Deep Active Learning for Joint Classification & Segmentation with Weak Annotator

Soufiane Belharbi[1], Ismail Ben Ayed[1], Luke McCaffrey[2], and Eric Granger[1]
[1] LIVIA, Dept. of Systems Engineering, École de technologie supérieure, Montreal, Canada
[2] Goodman Cancer Research Centre, Dept. of Oncology, McGill University, Montreal, Canada
soufiane.belharbi.1@ens.etsmtl.ca, luke.mccaffrey@mcgill.ca,
{ismail.benayed, eric.granger}@etsmtl.ca

Abstract

CNN visualization and interpretation methods, like class-activation maps (CAMs), are typically used to highlight the image regions linked to class predictions. These models allow to simultaneously classify images and extract class-dependent saliency maps, without the need for costly pixel-level annotations. However, they typically yield segmentations with high false-positive rates and, therefore, coarse visualisations, more so when processing challenging images, as encountered in histology. To mitigate this issue, we propose an active learning (AL) framework, which progressively integrates pixel-level annotations during training. Given training data with global image-level labels, our deep weakly-supervised learning model jointly performs supervised image-level classification and active learning for segmentation, integrating pixel annotations by an oracle. Unlike standard AL methods that focus on sample selection, we also leverage large numbers of unlabeled images via pseudo-segmentations (i.e., self-learning at the pixel level), and integrate them with the oracle-annotated samples during training. We report extensive experiments over two challenging benchmarks – high-resolution medical images (histology GlaS data for colon cancer) and natural images (CUB-200-2011 for bird species). Our results indicate that, by simply using random sample selection, the proposed approach can significantly outperform state-of the-art CAMs and AL methods, with an identical oracle-supervision budget. Our code is publicly available[1].

1. Introduction

Image classification and segmentation are fundamental tasks in many visual recognition applications involving natural and medical images. Given a large image dataset annotated with global image-level labels for classification or with pixel-level labels for segmentation, deep learning (DL) models achieve state-of-the-art performances for these tasks [12, 20, 32, 37, 38, 48]. However, the impressive accuracy of such fully-supervised learning models comes at the expense of a considerable cost for collecting and annotating large image data sets. While the acquisition of global image-level annotation can be relatively inexpensive, pixel-wise annotation involves a laborious process, a difficulty further accrued by the requirement of domain expertise, as in medical imaging, which increases the annotation costs.

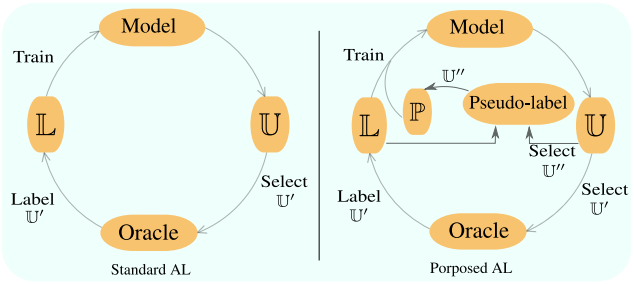


Figure 1: Proposed AL framework with weak annotator.

Weakly-supervised learning (WSL) has recently emerged as a paradigm that relaxes the need for dense pixel-wise annotations [49, 69]. WSL techniques depend on the type of application scenario and annotation, such as global image-level labels [4, 28, 45, 55, 61], scribbles [34, 54], points [2], bounding boxes [10, 27] or global image statistics such as the target-region size [1, 24, 25, 26]. This paper focuses on learning using only image-level labels, which enables to classify an image while yielding pixel-wise scores (i.e., segmentations), thereby localizing the regions of interest linked to the image-class predictions.

Several CNN visualization and interpretation methods have recently been proposed, based on either perturbation, propagation or activation approaches, and allowing

[1] https://github.com/sbelharbi/deep-active-learning-for-joint-classification-and-segmentation-with-weak-annotator

to localize the salient image regions responsible for a CNN's predictions [17]. In particular, WSL techniques [49] rely on activation-based methods like CAM and, more recently, Gradient-weighted Class Activation Mapping (Grad-CAM), Grad-CAM++, Ablation-CAM and Axiom-based Grad-CAM [17, 36, 46]. Trained with only global image-level annotations, these methods locate the regions of interest (ROIs) of the corresponding class in a relatively inexpensive and accurate way. However, while these WSL techniques can provide satisfying results in natural images, they typically yield poor segmentations in relatively more challenging scenarios, for instance, histology data in medical image analysis [49]. We note two limitations associated with CAMs: (1) they are obtained in an unsupervised way (*i.e*. without pixel-level supervision under an ill-posed learning problem [9]); and (2) they have low resolution. For instance, CAMs obtained from ResNet models [22] have a resolution of $1/32$ of the input image. Interpolation is required to restore full image resolution. Both of these limitation with CAM-based methods lead to high false-positive rates, which may render them impractical [49].

Enhancing deep WSL models with pixel-wise annotation, as supported by a recent study in weakly-supervised object localization [9], can improve localization and segmentation accuracy, which is the central goal of this paper. To do so, we introduce a deep WSL model that allows supervised learning for classification, and active learning for segmentation, with the latter providing more accurate and high-resolution masks. We assume that the images in the training set are globally annotated with image-class labels. Relevant images are *gradually* labeled at the pixel level through an oracle that respects a low annotation-budget constraint. Consequently, this leads us to an active learning (AL) paradigm [51], where an oracle is requested to annotate pixels in a subset of samples.

Different sample-acquisition techniques have been successfully applied to deep AL for classification based on, e.g., certainty [13, 18, 30] or representativeness [29, 52]. However, very few deep AL techniques were investigated in the context of segmentation [19, 21, 40]. Most AL techniques focus mainly on the sampling criterion (Fig.1, left) to populate the labeled pool using an oracle. During training, only the labeled pool is used, while the unlabeled pool is left dormant. Such an AL protocol may limit the accuracy of DL models under constrained oracle-supervision budget in real-world applications for multiple reasons:

(1) Standard AL protocols may be relevant to small/shallow models that can learn and provide reliable queries using a few training samples. Since training accurate DL models typically depends on large training sets, large numbers of queries may be needed to build reliable DL models, which may incur a high annotation cost.

(2) In most AL work, the experimental protocol starts with a large labeled pool, and acquires a large number of queries for sufficient supervision, neglecting the workload placed on the oracle. This typically reaches a plateau-performance of a DL quickly, hampering a reliable study of the impact of different AL selection techniques. Moreover, model-based sampling techniques are inconsistent [19] in the sense that the model is used to query samples while it is still in an early learning stage.

(3) Segmentation and classification problems are associated with different properties and challenges, such as decision boundaries and uncertainty, which provide additional challenges to AL. For instance, the class boundaries defined by different classification methods [14, 51, 56] are not defined in the context of segmentation, making such a branch of methods inadequate for segmentation.

(4) In critical fields such as medical imaging, acquiring a sample itself can be very expensive[2]. The time and cost associated with each sample makes them valuable items. Such considerations may be overlooked for large-scale data sets with almost-free samples, as in the case of natural images. Given this high cost, keeping the unlabeled pool dormant during learning may be ineffective.

Based on the aforementioned arguments, we advocate that focusing solely on the sample acquisition and supervision pool may not be an efficient way to build high-performing DL models in an AL framework for segmentation. Therefore, we consider augmenting the labeled pool using the model as a second source of annotation, in a self-learning fashion [42] (Fig.1, right). This additional annotation might be less accurate (*i.e*., weak[3]) compared to the oracle that provides strong but expensive annotations. However, it is expected to fast-improve the model's performance [42], while using a few oracle-annotated samples, reducing the annotation cost.

Our main contributions are the following. **(1) Architecture design**: As an alternative to CAMs, we propose using a segmentation mask trained with pixel-level annotations, which yields more accurate and high-resolution ROIs. This

[2]For instance, prior to a diagnosis of breast cancer from a histological sample, a patient undergoes a bilateral diagnostic mammogram and breast ultrasound that are interpreted by a radiologist, one to several needle biopsies (with low risks under 1% of hematoma and wound infection) to further assess areas of concern, surgical consultation and pre-operative blood work, and surgical excision of the positive tissue for breast cancer cells. The biopsy and surgical tissues are processed (fixation, embedding in parraffin, H&E staining) and interpreted by a pathologist. Depending on the cancer stage, the patient may undergo additional procedures. Therefore, accounting for all the steps required for breast cancer diagnosis from histological samples, a rough estimation of the final cost associated with obtaining a Whole Slide Image (WSI) is about $400 (Canadian dollars, by 1999) [62]. Moreover, some cancer types are rare [62], adding to the values of these samples. All these procedures are conducted by highly trained experts, with each procedure taking from a few minutes to an hour and requiring expensive specialized medical equipment.

[3]Not to be confused with the weak annotation of data in weakly supervised learning frameworks.

is achieved through a convolutional architecture capable of simultaneously classifying and segmenting images, with the segmentation task trained using annotations acquired using an AL framework. As illustrated in Fig.3, the architecture combines well-known DL models for classification (ResNet [22]) and segmentation (U-Net [48]), although other architectures could also be used. **(2) Active learning**: We augment the size of the labeled pool by weak-annotating a large number of unlabeled samples based on predictions of the DL model itself, providing a second source of annotation (Fig.1). This enables rapid improvements of the segmentation accuracy, with less oracle-based annotation. Moreover, our method can be integrated on top of any sample-acquisition method. **(3) Experimental study**: We conducted comprehensive experiments over two challenging benchmarks – high-resolution medical images (histology GlaS data for colon cancer) and natural images (CUB-200-2011 for bird species). Our results indicate that, by simply using random sample selection, the proposed approach can significantly outperform state-of the-art CAMs and AL methods, with an identical oracle-supervision budget.

2. Related work

Deep active learning: AL has been studied for a long time in machine learning, mainly for classification and regression, using linear models in particular [51]. Recently, there has been an effort to transfer such techniques to DL models for classification tasks by mimicking their intuition or by adapting them, taking into consideration model specificity and complexity. Such methods include, for instance, different mechanisms for uncertainty [5, 13, 14, 18, 29, 30, 33, 60, 65] and representativeness estimation [29, 50, 52]. However, most deep AL techniques are validated on synthetic, simple or tiny data, which does not explore their full potential in real applications.

While deep AL for classification is rapidly growing, deep AL models for segmentation are uncommon in the literature. In fact, the very few methods in the literature mostly focused on the direct application of deep AL classification methods. The limited research in this area may be explained by the fact that segmentation tasks bring challenges to AL, such as the additional spatial information and the fact that a segmentation mask lies in a much larger dimension than a classification prediction. In classification, AL often deals with one output that is used to drive queries [23]. The spatial information in segmentation does not naturally provide a direct scoring function that can indicate the overall quality or certainty of the output. Most of deep AL methods for segmentation consider pixels as classification instances, and apply standard AL techniques to each pixel.

For instance, the authors of [19] exploit a variant of entropy-based acquisition at the pixel level, combined with a distribution-based term that encodes diversity using a complex hierarchical clustering algorithm over sliding windows, with application to microscopic membrane segmentation. Similarly, [21, 40] apply Monte-Carlo dropout uncertainty [18] at the pixel level, with application to myelin segmentation using spinal cord and brain microscopic histology images. In [47], the authors experiment with five acquisition functions of classification for a segmentation task, including entropy-based, core-set [50], k-mean and Bayesian [18] sampling, with application to electron microscopy segmentation. Entropy-based methods seem to be dominant over multiple datasets. In [64], the authors combine two sampling terms for histology image segmentation. The first employs bootstrapping over fully convolutional networks (FCN) to estimate uncertainty, where a set of FCNs are trained on different subsets of samples. The second term is a representation-based term that selects the most representative samples. This is achieved by solving an optimization of a generalization version of the maximum cover set problem [16] using sample description extracted from an FCN. Despite the obtained promising results, this approach remains complex and impractical due to the use of bootstrapping over DL models and an optimization step. Moreover, the authors of [64] do not provide a comparison to other acquisition functions. The work in [8] considers a specific case of AL using reinforcement learning for *region-based* AL for segmentation, where only a selected region of the image is labeled. This method is adequate for data sets with large and unbalanced classes, such as street-view images. While the method in [8] outperforms random and Bayesian [18] selection, surprisingly, it performs close to entropy-based selection.

Weak annotators: The AL paradigm does not prohibit the use of unlabelled data for training [51], but it mainly constrains the oracle-labeling budget. The standard AL experimental protocol (Fig.1, left) was inherited from AL of simple/linear ML models, and adopted in subsequent works. Budget-constrained oracle annotation may not be sufficient to build effective DL models, due to the lack of labeled samples. Furthermore, several studies in AL for classification have successfully leveraged the unlabelled data to provide additional supervision [35, 39, 58, 60, 68, 71].

For instance, the authors of [35, 60] propose to pseudo-label a part of the unlabeled pool. The latter is selected using dynamic thresholding based on confidence, through the model itself, so as to learn a better embedding. Furthermore, a theoretical framework for AL using strong and weak annotators for classification task is introduced in [66]. Their results suggest that using multiple annotators can reduce the cost of oracle annotation, in comparison to one annotator. Multiple sources of annotations that include both strong and weak annotators were used in AL, crowdsourcing, self-paced learning and other interactive learning scenarios for classification to help reducing the number of

requests for the strong annotator [31, 41, 43, 44, 57, 63, 66]. Using the model itself for pseudo-annotation is motivated mainly by the success of deep self-supervised learning [42].

Label Propagation (LP): Our approach is also related to LP methods [6, 67, 70] for classification, which aim to label unlabeled samples using knowledge from the labeled ones (Fig.2). However, while LP propagates labels to unlabeled samples through an iterative process, our approach bypasses this using the model itself. In our case, the propagation is limited to the neighbors of labeled samples defined through k-nearest neighbors (k-nn) (Fig.2). Using k-nn has been also studied to combine AL and domain adaptation [7], where the goal is to query samples from the target domain. Such an approach is connected to the recently developed core-set method for deep AL [50]. Our method intersects with [7] only in the sense of predicting the labels to query samples using their labeled neighbors.

In contrast to state-of-the-art DL models for AL segmentation, we consider increasing the unlabeled pool through pseudo-annotated samples (Fig.1, right). To this end, the model is used for pseudo-labeling samples within the neighborhood of samples with strong supervision (Fig.2). From a self-learning perspective, the works in [35, 60] on face recognition are the closest to ours. While both rely on pseudo-labeling, they mainly differ in the sample selection for pseudo-annotation. In [35, 60], the authors considered model confidence, where samples with high confidence are pseudo-labeled, while low-confidence samples are queried. While this yields good results, it makes the overall method strongly dependent on model confidence. As we consider segmentation tasks, model-confidence is not well-defined. Moreover, using the expected pixel-wise values can be less representative for model confidence.

Our approach relies on the spatial assumption in Fig.2, where the samples to pseudo-label are selected to be near the labeled samples, and expected to have good pseudo-segmentations. This makes the oracle-querying technique independent from the pseudo-labeling method, giving more flexibility to the user. Our pseudo-labeled samples are added to the labeled pool, along with the samples annotated by the oracle. The underlying assumption is that, given a sample labeled by an oracle, the model is more likely to produce good segmentations of images located nearby that sample. Our assumption is verified empirically in the experimental section of this paper. This simple procedure enables to rapidly increase the number of pseudo-labeled samples, and helps improving segmentation performance under a limited amount of oracle-based supervision.

3. Proposed approach

We consider an AL framework for training deep WSL models, where all the training images have class-level annotations, but no pixel-level annotations. Due to their high cost, pixel annotations are gradually acquired for training through oracle queries. It propagate pixel-wise knowledge encoded in the model though the labeled images.

Active learning training consists of sequential training rounds. At each round r, the total training set $\mathbb{D}$ that contains n samples with unlabeled and labeled subsets (Fig.1). **(1) Unlabeled subset:** contains samples without pixel-wise annotation (unlabeled samples) $\mathbb{U} = \{\boldsymbol{x}_i, y_i, -\}_{i=1}^{u}$ where $\boldsymbol{x} \in \mathcal{X}$ is the input image, y is its global label; and the pixel label is missing. **(2) Labeled subset:** contains samples with full supervision $\mathbb{L} = \{\boldsymbol{x}_i, y_i, \boldsymbol{m}_i\}_{i=1}^{l}$ where $\boldsymbol{m}$ is the pixel-wise annotation of the sample. $\mathbb{L}$ is initially empty. It is gradually populated from $\mathbb{U}$ by querying the oracle using an acquisition function. Let $f(\cdot : \boldsymbol{\theta})$ denotes a DL model that is able to classify and segment an image $\boldsymbol{x}$ (Fig.3). For clarity, and since we focus on the segmentation task, we omit the notation for the classification task (to simplify the presentation). Therefore, $f(\boldsymbol{x})$ refers to the predicted segmentation mask. Let $\mathbb{U}' \subseteq \mathbb{U}$ and $\mathbb{U}'' \subseteq \mathbb{U}$ denote two subsets (Fig.1), with $\mathbb{U}' \cap \mathbb{U}'' = \varnothing$. In our method, we introduce $\mathbb{P}$ as a subset holder for pseudo-labeled samples, which is initially empty and gradually replenished (Fig.1, right). To express the dependency of each subset on round r, we introduce the following notations: $\mathbb{U}_r, \mathbb{L}_r, \mathbb{P}_r, \mathbb{U}'_r, \mathbb{U}''_r$. The samples in $\mathbb{P}_r$ are denoted as $\{\boldsymbol{x}_i, y_i, \hat{\boldsymbol{m}}_i\}$. The following holds: $\forall r : \mathbb{D} = \mathbb{L}_r \cup \mathbb{U}_r \cup \mathbb{P}_r$.

Alg.1 describes the overall AL process with our pseudo-annotation method. First, $\mathbb{U}'_r$ is queried, then labeled by the oracle, and added to $\mathbb{L}_r$. Using k-nn, $\mathbb{U}''_r$ is selected based on their proximity to $\mathbb{L}_r$ (Fig.2); and pseudo-labeled by the

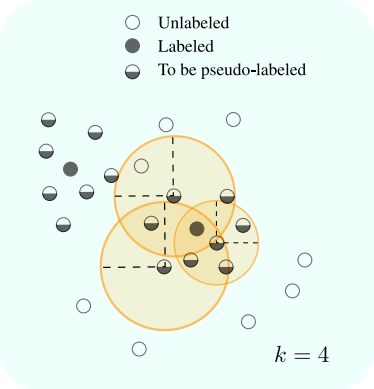


Figure 2: The k-nn method for selecting $\mathbb{U}''$ subset to be speudo-labeled. Assumption to select $\mathbb{U}''$: since $\mathbb{U}''$ lives nearby supervised samples, it is more likely to be assigned good segmentation by the model. We consider measuring k-nn for each **unlabeled** sample. In this example, using $k = 4$ allows $|\mathbb{U}''| = 14$. If k-nn is considered for each **labeled** sample: $|\mathbb{U}''| = 8$. $|\cdot|$ is the set cardinal. Note that k-nn is only considered between samples of the *same class*.

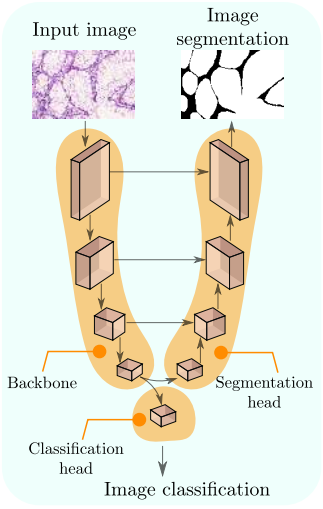


Figure 3: Out proposed DL architecture for classification and segmentation composed of: (1) a shared **backbone** for feature extraction; (2) a **classification head** for the classification task; (3) and a **segmentation head** for the segmentation task with a U-Net style [48]. The latter merges representations from the backbone, while gradually upscaling the feature maps to reach the full image resolution for the predicted mask, similarly to the U-Net model.

model, then added to $\mathbb{P}_r$. To fast-increase the size of $\mathbb{L}_r$, $\mathbb{P}_r$ is protected from being queried for the oracle until it is inevitable. In the *latter case*, queried samples from $\mathbb{P}_r$ are used to fill $\mathbb{U}'$; and they are no longer considered pseudo-labeled since they will be assigned the oracle annotation.

To measure image similarity for the k-nn method, we used the color distribution to describe image content. This can be a flexible descriptor for highly unstructured images such as histology images. Note that the k-nn method is considered *only* for pairs of samples of the *same class*. The underlying assumption is that samples of the same class, with similar color distributions, are likely to contain relatively similar objects. Consequently, labeling representative samples could be a proxy for supervised learning based on the underlying data distribution. This can increase the likelihood of the model to provide relatively good segmentations of the other samples. The proximity between two images $(\boldsymbol{x}_i, \boldsymbol{x}_j)$ is measured using the Jensen-Shannon divergence between their respective color distributions (measured as normalized histograms). For an image with multiple color planes, the similarity is formulated as the sum of similarities, one for each plane.

At round r, the queried and pseudo-labeled samples are both used in training by optimizing the following loss function:

$$\min_{\boldsymbol{\theta}} \sum_{\boldsymbol{x}_i \in \mathbb{L}_{r-1}} \mathcal{L}(f(\boldsymbol{x}_i), \boldsymbol{m}_i) + \lambda \sum_{\boldsymbol{x}_i \in \mathbb{P}_{r-1}} \mathcal{L}(f(\boldsymbol{x}_i), \hat{\boldsymbol{m}}_i), \tag{1}$$

where $\mathcal{L}(\cdot,\cdot)$ is a segmentation loss, and λ a positive scalar. Eq.(1) corresponds to training the model (Fig.3) solely for the segmentation task. Simultaneous training for classification and segmentation in this AL setup is avoided due to the unbalance between the number of samples that are labeled globally and at the pixel level. Therefore, we consider training the model for classification first. Then, we freeze the classifier parameters. Training for the segmentation tasks is resumed later. This yields the best classification performance, and allows a better study of the impact of the queried samples on the segmentation task.

Considering the relation of our method and label propagation algorithm [6, 67, 70], we refer to our proposal as Label_prop.

Algorithm 1: Standard AL procedure and our proposal. The extra instructions associated with our method are indicated with a blue background.

Input: $\mathbb{P}_0 = \mathbb{L}_0 = \varnothing$
$\boldsymbol{\theta}^0$: Initial parameters of f trained on the classification task.
`maxr` : Maximum number of AL rounds.

1 Select $\mathbb{U}'_0$ randomly and label them by an oracle.
2 $\mathbb{L}_0 \leftarrow \mathbb{U}'_0$.
3 **for** $r \in 1 \cdots$ `maxr` **do**
4 $\quad \boldsymbol{\theta} \leftarrow \boldsymbol{\theta}^0$.
5 $\quad$ Train f using $\mathbb{L}_{r-1} \cup \mathbb{P}_{r-1}$ and the loss in Eq. (1).
6 $\quad$ Select $\mathbb{U}'_r$ and label them by an oracle.
7 $\quad \mathbb{L}_r \leftarrow \mathbb{L}_{r-1} \cup \mathbb{U}'_r$.
8 $\quad$ Select $\mathbb{U}''_r$.
9 $\quad \mathbb{P}_r \leftarrow \mathbb{P}_{r-1} \cup \mathbb{U}''_r$.
10 $\quad$ Pseudo-label $\mathbb{P}_r$.

4. Results and discussion

4.1. Experimental methodology:

a) Datasets. For evaluation, datasets should have global and pixel-wise annotation. We consider two public datasets

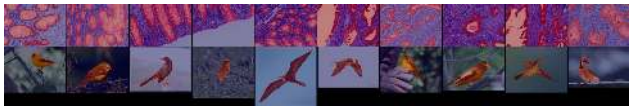


Figure 4: **Top row**: GlaS dataset [53]. **Bottom row**: CUB dataset [59]. (Best visualized in color.)

Table 1: Number of samples selected for the oracle per round.

Dataset	#selected samples *per-class* ($r = 1$)	#selected samples *per-class* ($r > 1$)	Max AL rounds (`maxr` in Alg.1)
GlaS	4	1	25
CUB	1	1	20

including both medical (histology) and natural images (Fig.4). (1) **GlaS dataset**: This dataset contains histology images for colon cancer diagnosis[4] [53]. It includes 165 images derived from 16 Hematoxylin and Eosin (H&E) histology sections of two grades (classes): benign and malignant. It is divided into 84 samples for training and 80 samples for testing. The ROIs to be segmented are the glandes. (2) **CUB-200-2011 dataset (CUB)**[5] [59] is a dataset for bird species with $11,788$ samples ($5,994$ for training and $5,794$ for testing) and 200 species. The ROIs to be segmented are the birds. In GlaS and CUB datasets, we randomly select 80% of the training samples for effective training, and 20% for validation (with full supervision) to perform early stopping. The splits are identical to the ones used in [3, 49] (split 0, fold 0), and are publicly available.

b) Active learning setup. To assess the performance of different AL acquisition methods, we consider a realistic scenario with respect to the number of samples to be labeled at each AL round, accounting for the load imposed on the oracle. Therefore, only a few samples are selected at each round for oracle annotation, and $\mathbb{L}$ is slowly replenished. This allows better comparison between AL selection techniques since we spend more time in a phase where $\mathbb{L}$ holds a few samples. Such a phase allows to better measure the impact of the selected samples. Filling $\mathbb{L}$ quickly brings the model's performance to a plateau that hides the impact of newly selected samples. The initial replenishment ($r = 1$) is achieved by randomly selecting a few samples. The same samples are used for all AL techniques at round $r = 1$ for a fair comparison. To avoid any bias from selecting unbalanced classes that can directly affect the segmentation performance, and hinder AL evaluation, the same number of samples is selected from each class (since the global annotation is known beforehand for all the samples). Note that the oracle is used only to provide pixel-wise annotation. Tab.1 provides the selection details.

c) Evaluation. We report the classification accuracy obtained by the classification head (Fig.3). Average Dice index is used to measure the segmentation quality at each AL round forming a Dice index curve over all the rounds. To better assess the *dominance* of each method [51], the Area Under the Dice index Curve is used (AUC). This provides a fair indicator of the dominant curve, but contrasts with standard AL works, where one or multiple specific operating points in the curve are selected (leading to a biased and less accurate protocol). The average and standard deviation of Dice index curve and AUC metric are reported based on 5 replications of a complete AL session, using a different seed for each session. An AL session across different methods uses the same seed.

[4]GlaS: warwick.ac.uk/fac/sci/dcs/research/tia/glascontest.

[5]CUB: www.vision.caltech.edu/visipedia/CUB-200-2011.html

Table 2: Classification accuracy over of the proposed deep WSL model on GlaS and CUB test datasets.

Dataset	GlaS	CUB
Classification accuracy (%)	99.50 ± 0.61	73.22 ± 0.19

Table 3: Readings of Dice index (mean $\pm$ standard deviation) from Fig.5 over test set for the **first 5 queries** formed by each method. We start from the second query since the first query is random but identical for all methods.

Queries	**q2**	**q3**	**q4**	**q5**	**q6**
GlaS					
WSL	66.44 ± 0.20				
Random	70.26 ± 3.02	71.58 ± 3.14	71.43 ± 1.83	74.05 ± 3.14	75.36 ± 3.45
Entropy	$\mathbf{72.75 \pm 2.96}$	70.93 ± 3.58	72.60 ± 1.44	73.44 ± 1.38	75.15 ± 1.63
MC_Dropout	68.44 ± 2.89	69.70 ± 1.96	69.97 ± 1.95	72.71 ± 2.21	73.00 ± 1.04
Label_prop (ours)	71.02 ± 4.19	$\mathbf{74.07 \pm 3.93}$	$\mathbf{76.52 \pm 3.49}$	$\mathbf{77.63 \pm 2.73}$	$\mathbf{78.41 \pm 1.23}$
Full_sup	86.53 ± 0.31				
CUB					
WSL	39.22 ± 0.18				
Random	56.86 ± 2.07	61.39 ± 1.85	62.97 ± 1.13	63.56 ± 4.02	66.56 ± 2.50
Entropy	53.37 ± 2.06	59.11 ± 2.50	60.48 ± 3.56	63.81 ± 2.75	63.59 ± 2.34
MC_Dropout	57.13 ± 0.83	59.98 ± 2.06	63.52 ± 2.26	63.02 ± 2.68	64.68 ± 1.41
Label_prop (ours)	$\mathbf{62.58 \pm 2.15}$	$\mathbf{66.32 \pm 2.34}$	$\mathbf{67.01 \pm 2.85}$	$\mathbf{69.40 \pm 3.40}$	$\mathbf{68.28 \pm 1.60}$
Full_sup	75.29 ± 1.50				

While our approach, referred to as (Label_prop), can operate on top of any AL selection criterion, we demonstrate its efficiency using simple random selection, which is often a baseline for AL experiments. Note that our pseudo-annotations are obtained from the segmentation head shown in Fig.3. Our method is compared to three different AL selection approaches for segmentation: **(1) random selection (Random)**: the samples are randomly selected; **(2) entropy-based selection (Entropy)**: the scoring function per sample is the average entropy at the pixel level [19]. Samples with high entropy are selected; and **(3) Monte-Carlo dropout uncertainty (MC_Dropout)**: we use Monte-Carlo dropout [21, 40] at the pixel level to compute the uncertainty score per sample. Samples are forwarded 50 times in the model, where dropout is set to 0.2 [21, 40]. Then, the pixel-wise variance is estimated. Samples with high mean variance are selected.

Lower bound performance (WSL): We consider the segmentation performance obtained by WSL method as a lower bound. It is trained using only global annotation. CAMs are used to extract the segmentation mask. WILDCAT method

is considered [15] (Fig.3) at the classification head to obtain the CAMs. For WSL method, a pre-trained model over ImageNet [11] is used to initialize the weights of the backbone, which is then fined-tuned. The model is trained over the entire dataset, where samples are labeled globally only. The obtained classifier using seed=0 is frozen and used as a backbone for *all* the other methods.

Upper bound performance (Full_sup): Fully supervised segmentation is considered as an upper bound on performance. The model in Fig.3 is trained for segmentation only using the entire dataset, where samples are labeled at the pixel level.

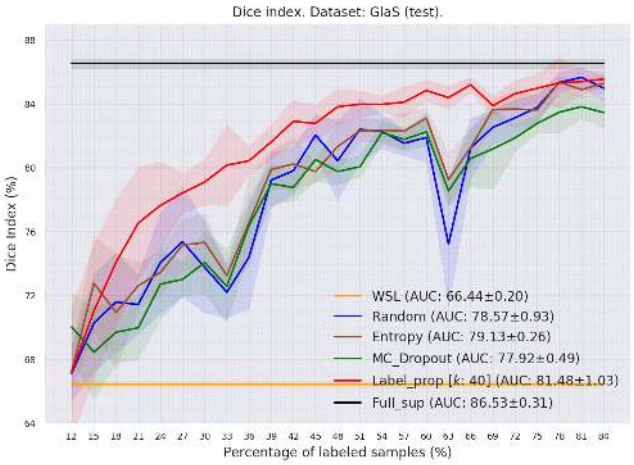
For a fair comparison, all the methods are trained using the same hyper-parameters over the same dataset. WSL and Full_sup methods have minor differences. Due to space limitation, all the hyper-parameters are presented in the supplementary material. In Alg.1, notice that for our method, $\mathbb{P}_r$ is not used at the current round r but until the next round $r+1$. To take advantage of $\mathbb{P}_r$ at round r, instructions from line-4 to line-10 are repeated twice in the provided results.

4.2. Results

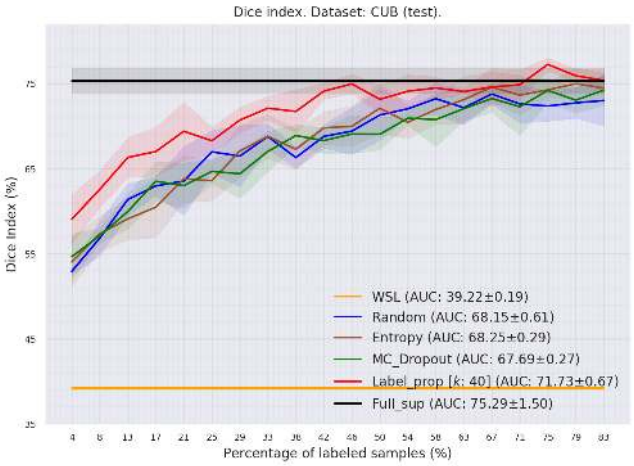
We report the classification and segmentation performances following the training the proposed deep WSL model in Fig.3. Tab.2 reports the Classification accuracy of the classification head using WSL, which is close to the results reported in [4, 49]. The results of GlaS suggest that it is an easy dataset for classification.

The segmentation results are reported in Tabs. 3 and 4, and in Fig 5.

Fig. 5a compares Dice accuracy on the **GlaS dataset**. On the latter, we observe that adding more labels increases Dice index for all AL methods, yielding, as expected, better performance than the WSL method. Reading from Tab.3, randomly labeling only 4 samples per class enables to easily outperform WSL. This means that using our approach in Fig.3, with limited supervision, can lead to more accurate masks compared to using CAMs in the WSL method. From Fig.5a, one can also observe that Random, Entropy, and MC_Dropout methods grow relatively in the same way, leading to the same overall performance, with the Entropy method slightly ahead. Considering the overall behavior of the curves, one may conclude that using advanced selection techniques such as MC_Dropout and Entropy provides an accuracy similar to simple random selection. On the one hand, since both methods have shown substantial improvements in AL for classification, and based on the results in Fig.5a, one may conclude that all samples are equivalently informative for the model. Therefore, there is no better order to acquire them. On the other hand, using simply random selection and pseudo-labeled samples allowed our method to substantially improve the overall performance, demonstrating the benefits of self-learning.



(a)



(b)

Figure 5: Average Dice index of the proposed and baseline methods over test sets. (a) GlaS. (b) CUB.

Fig.5b and Tab.3 compare Dice accuracy on the **CUB dataset**, where labeling only one sample per class yielded a large improvement in Dice index, in comparison to WSL. Adding more samples increases the performance of all the methods. One can observe similar pattern as for GlaS: Random, Entropy and MC_Dropout methods yield similar curves, while the AUC performances of Random and Entropy methods are similar, and slightly ahead of MC_Dropout. Similar to GlaS analysis, and based on the results of these three methods, one can conclude that none of the methods for ordering the samples is better than simple random selection. Using self-labeled samples in our method shows again its benefits. Simple random selection combined with self-annotation yields an overall best performance. Using two datasets, our empirical results suggest that self-learning, under limited oracle-annotation, has

the potential to provide a reliable second source of annotation, which can efficiently enhance model performance, while using simple sample acquisition techniques.

Pseudo-annotation performance. Furthermore, the proposed approach is assessed on the pseudo-labeled samples at each AL round. Fig.6 shows that the model provides good segmentations at the initial rounds. Then, the more supervision, the more accurate the pseudo-segmentation, as expected. This figure shows the interest and potential of self-learning in segmentation, and confirms our assumption that samples near the labeled ones are likely to achieve accurate pseudo-segmentation by the model.

Hyper-parameters. Our approach requires two main hyper-parameters: k and λ. We conducted an ablation study over k on GlaS dataset, and over λ on both datasets. Results, which are presented in the supplementary material, suggest that our method is less sensitive to k. λ plays an important role, and based on our study, we recommend using small values of this weighting parameter. In our experiments, we used $\lambda = 0.1$ for Glas and $\lambda = 0.001$ for CUB. We set $k = 40$. We note that hyper-parameter tuning in AL is challenging due to the change of the size of the data set, which in turn changes the training dynamics. In all the experiments, we used fixed hyper-parameters across the AL rounds. Fig.6 suggests that a dynamic $\lambda(r)$ that is increased through AL rounds could be more beneficial. However, this requires a principled update protocol for λ, which was not explored in this work. Nonetheless, using a fixed value seems to yield promising results overall.

Supplementary material. Due to space limitation, we deferred the hyper-parameters used in the experiments, results of the ablation study, visual results for the similarity measure and examples of predicted masks to the supplementary materials.

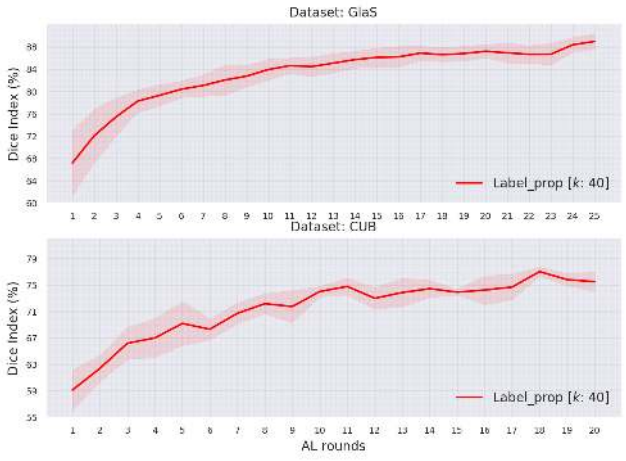


Figure 6: Average Dice index over the pseudo-labeled samples of our method in **each** AL round.

Table 4: Average AUC and standard deviation (Fig.5) for Dice index performance over GlaS and CUB test sets.

Dataset	GlaS	CUB
WSL	66.44 ± 0.20	39.22 ± 0.19
Random	78.57 ± 0.93	68.15 ± 0.61
Entropy	79.13 ± 0.26	68.25 ± 0.29
MC_Dropout	77.92 ± 0.49	67.69 ± 0.27
Label_prop (ours)	$\mathbf{81.48 \pm 1.03}$	$\mathbf{71.73 \pm 0.67}$
Full_sup	86.53 ± 0.31	75.29 ± 1.50

5. Conclusion

Deep WSL models trained with global image-level annotations can play an important role in CNN visualization and interpretability. However, they are prone to high false-positive rates, especially for challenging images, leading to poor segmentations. To alleviate this issue, we considered using pixel-wise supervision provided gradually through an AL framework. This annotation is integrated into training using an adequate deep convolutional model that allows supervised learning for both tasks: classification and segmentation. Through a few pixel-supervised samples, such a design is intended to provide full-resolution and more accurate masks compared to standard CAMs, which are trained without pixel supervision and often provide coarse resolution. Therefore, it enables a better CNN visualization and interpretation of CNN predictions. Furthermore, and unlike standard deep AL methods that focus solely on the acquisition function, we considered using self-learning as a second source of supervision to fast-improve the model segmentation. Evaluating our method using a realistic AL protocol over two challenging benchmarks, our results indicate that: (1) using a *few* supervised samples, the proposed architecture yielded more accurate segmentations compared to CAMs, with a large margin using different AL methods. Thus, it provides a solution to enhance pixel-wise predictions in real-world visual recognition applications. (2) using self-learning with random selection yielded substantial improvements. Self-learning under a limited oracle-budget can, therefore, provide a cost-effective alternative to standard AL protocols, where most of the effort is spent on the acquisition function.

Acknowledgment

This research was supported in part by the Canadian Institutes of Health Research, the Natural Sciences and Engineering Research Council of Canada, Compute Canada, MITACS, and the Ericsson Global AI Accelerator Montreal.

References

[1] M. Bateson, H. Kervadec, J. Dolz, H. Lombaert, and I. Ben Ayed. Constrained domain adaptation for segmentation. In *MICCAI*, 2019.

[2] A. Bearman, O. Russakovsky, V. Ferrari, and F.-F. Li. What's the point: Semantic segmentation with point supervision. In *ECCV*, 2016.

[3] S. Belharbi, I. Ben Ayed, L. McCaffrey, and E. Granger. Deep ordinal classification with inequality constraints. *CoRR*, abs/1911.10720, 2019.

[4] S. Belharbi, J. Rony, J. Dolz, I. Ben Ayed, L. McCaffrey, and E. Granger. Min-max entropy for weakly supervised pointwise localization. *CoRR*, abs/1907.12934, 2019.

[5] W. H. Beluch, T. Genewein, A. Nürnberger, and J. M. Köhler. The power of ensembles for active learning in image classification. In *CVPR*, 2018.

[6] Y. Bengio, O. Delalleau, and N. Le Roux. Label propagation and quadratic criterion. In O. Chapelle, B. Scholkopf, and A. Zien, editors, *Semi-supervised learning*, chapter 11. The MIT Press, 2010.

[7] C. Berlind and R. Urner. Active nearest neighbors in changing environments. In *ICML*, 2015.

[8] A. Casanova, P.O. Pinheiro, N. Rostamzadeh, and C. J. Pal. Reinforced active learning for image segmentation. In *ICLR*, 2020.

[9] J. Choe, S. J. Oh, S. Lee, S. Chun, Z. Akata, and H. Shim. Evaluating weakly supervised object localization methods right. In *CVPR*, 2020.

[10] Jifeng Dai, Kaiming He, and Jian Sun. Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In *ICCV*, 2015.

[11] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR*, 2009.

[12] J. Dolz, C. Desrosiers, and I. Ben Ayed. 3d fully convolutional networks for subcortical segmentation in mri: A large-scale study. *NeuroImage*, 170, 2018.

[13] M. Ducoffe and F. Precioso. Qbdc: query by dropout committee for training deep supervised architecture. *CoRR*, abs/1511.06412, 2015.

[14] Melanie Ducoffe and Frederic Precioso. Adversarial active learning for deep networks: a margin based approach. *CoRR*, abs/1802.09841, 2018.

[15] T. Durand, T. Mordan, N. Thome, and M. Cord. Wildcat: Weakly supervised learning of deep convnets for image classification, pointwise localization and segmentation. In *CVPR*, volume 2, 2017.

[16] U. Feige. A threshold of ln n for approximating set cover. *Journal of the ACM*, 45(4), 1998.

[17] R. Fu, Q. Hu, X. Dong, Y. Guo, Y. Gao, and B. Li. Axiom-based grad-cam: Towards accurate visualization and explanation of cnns. *BMVC*, 2020.

[18] Y. Gal, R. Islam, and Z. Ghahramani. Deep bayesian active learning with image data. In *ICML*, 2017.

[19] U. Gaur, M. Kourakis, E. Newman-Smith, W. Smith, and B.S. Manjunath. Membrane segmentation via active learning with deep networks. In *ICIP*, 2016.

[20] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. `http://www.deeplearningbook.org`.

[21] M. Górriz Blanch. Active deep learning for medical imaging segmentation. B.S. thesis, Universitat Politècnica de Catalunya, 2017.

[22] K. He, X. Zhang, S.g Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[23] S.-J. Huang, R. Jin, and Z.-H. Zhou. Active learning by querying informative and representative examples. In *NIPS*, 2010.

[24] Z. Jia, X. Huang, E. I-C. Chang, and Y. Xu. Constrained deep weak supervision for histopathology image segmentation. *Transactions on medical imaging*, 36(11), 2017.

[25] Hoel Kervadec, Jose Dolz, Eric Granger, and Ismail Ben Ayed. Curriculum semi-supervised segmentation. In *MICCAI*, 2019.

[26] H. Kervadec, J. Dolz, M. Tang, E. Granger, Y. Boykov, and I. Ben Ayed. Constrained-cnn losses for weakly supervised segmentation. *Medical image analysis*, 54, 2019.

[27] A. Khoreva, R. Benenson, J.H. Hosang, M. Hein, and B. Schiele. Simple does it: Weakly supervised instance and semantic segmentation. In *CVPR*, 2017.

[28] D. Kim, D. Cho, D. Yoo, and I. So Kweon. Two-phase learning for weakly supervised object localization. In *ICCV*, 2017.

[29] K. Kim, D. Park, K. I. Kim, and S. Y. Chun. Task-aware variational adversarial active learning. *CoRR*, abs/2002.04709, 2020.

[30] A. Kirsch, J. van Amersfoort, and Y. Gal. Batchbald: Efficient and diverse batch acquisition for deep bayesian active learning. In *NIPS*, 2019.

[31] J. Kremer, F. Sha, and C. Igel. Robust active label correction. In *International Conference on Artificial Intelligence and Statistics*, 2018.

[32] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*. 2012.

[33] B. Lakshminarayanan, A. Pritzel, and C. Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *NIPS*, 2017.

[34] D. Lin, J. Dai, J. Jia, K. He, and J. Sun. Scribblesup: Scribble-supervised convolutional networks for semantic segmentation. In *CVPR*, 2016.

[35] L. Lin, K. Wang, D. Meng, W. Zuo, and L. Zhang. Active self-paced learning for cost-effective and progressive face identification. *TPAMI*, 40(1), 2017.

[36] M. Lin, Q. Chen, and S. Yan. Network in network. *coRR*, abs/1312.4400, 2013.

[37] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, and all. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42, 2017.

[38] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015.

[39] J. Long, J. Yin, W. Zhao, and E. Zhu. Graph-based active learning based on label propagation. In *International Conference on Modeling Decisions for Artificial Intelligence*, 2008.

[40] M. Lubrano di Scandalea, C. S Perone, M. Boudreau, and J. Cohen-Adad. Deep active learning for axon-myelin segmentation on histology data. *CoRR*, abs/1907.05143, 2019.

[41] L. Malago, N. Cesa-Bianchi, and J. Renders. Online active learning with strong and weak annotators. In *NIPS Workshop on Learning from the Wisdom of Crowds*, 2014.

[42] H. H. Mao. A survey on self-supervised pre-training for sequential transfer learning in neural networks. *CoRR*, abs/2007.00800, 2020.

[43] B. Mattsson. Active learning of neural network from weak and strong oracles. Master's thesis, 2017.

[44] K. Murugesan and J. Carbonell. Active learning from peers. In *NIPS*, 2017.

[45] D. Pathak, P. Krahenbuhl, and T. Darrell. Constrained convolutional neural networks for weakly supervised segmentation. In *ICCV*, 2015.

[46] P. H. O. Pinheiro and R. Collobert. From image-level to pixel-level labeling with convolutional networks. In *CVPR*, 2015.

[47] J. Roels and Y. Saeys. Cost-efficient segmentation of electron microscopy images using active learning. *CoRR*, abs/1911.05548, 2019.

[48] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015.

[49] J. Rony, S. Belharbi, J. Dolz, I. Ben Ayed, L. McCaffrey, and E. Granger. Deep weakly-supervised learning methods for classification and localization in histology images: a survey. *CoRR*, abs/1909.03354, 2019.

[50] O. Sener and S. Savarese. Active learning for convolutional neural networks: A core-set approach. In *ICLR*, 2018.

[51] B. Settles. Active learning literature survey. Technical report, University of Wisconsin-Madison Department of Computer Sciences, 2009.

[52] S. Sinha, S. Ebrahimi, and T. Darrell. Variational adversarial active learning. In *ICCV*, 2019.

[53] K. Sirinukunwattana, J. P.W. Pluim, H. Chen, et al. Gland segmentation in colon histology images: The glas challenge contest. *Medical image analysis*, 35, 2017.

[54] M. Tang, A. Djelouah, F. Perazzi, Y. Boykov, and C. Schroers. Normalized Cut Loss for Weakly-supervised CNN Segmentation. In *CVPR*, 2018.

[55] E. W. Teh, M. Rochan, and Y. Wang. Attention networks for weakly supervised object localization. In *BMVC*, 2016.

[56] S. Tong and D. Koller. Support vector machine active learning with applications to text classification. *JMLR*, 2, 2001.

[57] R. Urner, S. B. David, and O. Shamir. Learning from weak teachers. In *Artificial intelligence and statistics*, 2012.

[58] S. Vijayanarasimhan and K. Grauman. Active frame selection for label propagation in videos. In *ECCV*, 2012.

[59] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical report, California Institute of Technology, 2011.

[60] K. Wang, D. Zhang, Y. Li, R. Zhang, and L. Lin. Cost-effective active learning for deep image classification. *IEEE Transactions on Circuits and Systems for Video Technology*, 27(12), 2016.

[61] Y. Wei, J. Feng, X. Liang, M.-M. Cheng, Y. Zhao, and S. Yan. Object region mining with adversarial erasing: A simple classification to semantic segmentation approach. In *CVPR*, 2017.

[62] BP Will, C Le Petit, JM Berthelot, EM Tomiak, S Verma, and WK Evans. Diagnostic and therapeutic approaches for nonmetastatic breast cancer in canada, and their associated costs. *British journal of cancer*, 79(9), 1999.

[63] S. Yan, K. Chaudhuri, and T. Javidi. Active learning from imperfect labelers. In *NIPS*, 2016.

[64] L. Yang, Y. Zhang, J. Chen, S. Zhang, and D. Z. Chen. Suggestive annotation: A deep active learning framework for biomedical image segmentation. In *MICCAI*, 2017.

[65] D. Yoo and I. S. Kweon. Learning loss for active learning. In *CVPR*, 2019.

[66] C. Zhang and K. Chaudhuri. Active learning from weak and strong labelers. In *NIPS*, 2015.

[67] D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Schölkopf. Learning with local and global consistency. In *NIPS*, 2004.

[68] S. Zhou, Q. Chen, and X. Wang. Active deep networks for semi-supervised sentiment classification. In *Proceedings of International Conference on Computational Linguistics: Posters*, 2010.

[69] Z.-H. Zhou. A brief introduction to weakly supervised learning. *National Science Review*, 5, 2017.

[70] X. Zhu and Z. Ghahramani. Learning from labeled and unlabeled data with label propagation. 2002.

[71] X. Zhu, Z. Ghahramani, and J. Lafferty. Semi-supervised learning using gaussian fields and harmonic functions. In *ICML*, 2003.