# Computational Evaluation of the Combination of Semi-Supervised and Active Learning for Histopathology Image Segmentation with Missing Annotations

Laura Gálvez Jiménez*
Université Libre de Bruxelles
Brussels, Belgium
laura.galvez.jimenez@ulb.be

Lucile Dierckx*
Université Catholique de Louvain
Louvain-la-Neuve, Belgium
lucile.dierckx@uclouvain.be

Maxime Amodei*
Université de Liège
Liège, Belgium
maxime.amodei@uliege.be

Hamed Razavi Khosroshahi
Université Libre de Bruxelles
Brussels, Belgium
hamed.razavi.khosroshahi@ulb.be

Natarajan Chidambaran
Université de Mons
Mons, Belgium
natarajan.chidambaram@umons.ac.be

Anh-Thu Phan Ho
Multitel
Mons, Belgium
phanho@multitel.be

Alberto Franzin
Université Libre de Bruxelles
Brussels, Belgium
alberto.franzin@ulb.be

## Abstract

*Real-world segmentation tasks in digital pathology require a great effort from human experts to accurately annotate a sufficiently high number of images. Hence, there is a huge interest in methods that can make use of non-annotated samples, to alleviate the burden on the annotators. In this work, we evaluate two classes of such methods, semi-supervised and active learning, and their combination on a version of the GlaS dataset for gland segmentation in colorectal cancer tissue with missing annotations. Our results show that semi-supervised learning benefits from the combination with active learning and outperforms fully supervised learning on a dataset with missing annotations. However, an active learning procedure alone with a simple selection strategy obtains results of comparable quality.*

## 1. Introduction

Deep learning methods have demonstrated their ability to obtain state-of-the-art performance for image segmentation in different domains, from natural images to more complex problems such as biomedical images [3, 35]. However, training a deep learning model requires extensive and well-annotated datasets, which are not always available.

In critical domains such as digital pathology, annotations need to be provided by highly trained experts. Providing segmentation masks is a particularly costly operation, for which the annotators have to invest a relatively high amount of time and, in real-world medical applications, the scarcity of annotations is very common [56]. Machine learning models that can operate with partially annotated datasets would therefore alleviate the annotation effort.

Several techniques have been applied to segmentation tasks on partially annotated datasets in medical applications [53]. Notably, semi-supervised learning (SSL) methods make use of both annotated and non-annotated samples in a fully automated fashion [60]. Select the labeled data to use is however still a challenge. Its results, therefore, depend heavily on the quality of the labeled data initially available. Conversely, active learning (AL) involves a human annotator in the process, proposing a set of carefully selected unlabeled samples to an expert who will provide the required annotations, in order to minimize the effort required by the expert without sacrificing the expected quality of the results. However, the performance of AL strongly depends on its initial labeled pool [23, 28]. To combine the strength of each method while alleviating their issues, efforts have been made to combine AL and SSL for image segmentation [51, 64], but their application to digital pathology is still very limited [28].

In this work, we explore the combination of SSL and AL in the context of histopathology semantic image segmentation, using a state-of-the-art model for this task [48].

*These authors contributed equally to the work.

More precisely, we consider four AL selection criteria and combine them with a family of novel SSL methods that obtained state-of-the-art results for classification [50, 65, 73], to evaluate their potential on a segmentation task. We use the GlaS dataset for the assessment of gland segmentation in colorectal cancer tissue [48, 49], from which we remove annotations to simulate a realistic scenario in a controlled manner. We assess (i) the efficiency of the combination of SSL and AL in the case of histopathology image segmentation, (ii) the robustness of this framework against a dataset with missing annotations, and (iii) whether one AL selection criterion, SSL strategy, or their particular combination is especially recommended for medical image segmentation. Our results show that the combination of SSL and AL improves the performance of SSL, but not the performance of AL, regardless the strategy. We observe that a simple Only Positive strategy with AL is sufficient to obtain good quality results, without requiring SSL.

This paper is organized as follows. In Section 2 we give an overview of the related works. In Section 3 we present the dataset and the SSL+AL framework. The experimental protocol and results are presented and discussed in Section 4, before concluding in Section 5.

## 2. Related works

### 2.1. Semantic segmentation with missing annotations in digital pathology

Semantic segmentation is the task of partitioning an image into objects or regions. It is a pixel-level classification task that assigns the same label to pixels belonging to objects of the same class. Semantic segmentation is a fundamental problem in digital pathology, where objects such as nuclei or glands have to be precisely identified in an image [5, 15, 32, 63]. In recent years, deep learning models have consistently outperformed traditional computer vision approaches [48, 10, 22, 45, 52]. The task is formulated as a supervised learning one, where images are matched to a *segmentation mask*, an equally-sized image whose pixels are the labels of the corresponding pixels in the sample.

In real-world applications, the annotations represented in the segmentation mask are manually provided by experts. This is a time-consuming operation, and even experts are likely to omit annotations, provide inaccurate annotations, or make mistakes [16]. There is therefore a great interest in methods that can deal with missing and/or imperfect annotations [9, 17, 53, 59, 61]. Focusing specifically on medical image segmentation problems, authors have evaluated the impact of data augmentation techniques in different models, [12, 75] and data augmentation via generative models [8, 54]. Transfer learning is a popular approach that has been applied for colorectal cancer image segmentation [26, 43], breast cancer [62] and several other medical domains, see e.g. [11, 61] for more in-depth surveys on this topic. Model-specific techniques have also been proposed, such as in [39] that uses ensembles of independent models that do not consider pixels of uncertain status.

In this work, we consider semi-supervised learning and active learning, two families of machine learning methods that make use of partially annotated data during the training phase. We discuss them in greater detail in the rest of this section, with particular attention to their application to medical image segmentation.

### 2.2. Semi-supervised learning for image segmentation

Semi-supervised learning considers both labeled and unlabeled samples in the training set. One family of semi-supervised methods generates pseudo-annotations for the unlabeled images, in order to obtain a completely annotated set of images to use for subsequent trainings, until some convergence criterion is met. This pseudo-labeling procedure, also called self-training by some authors, is an application of the classic machine learning Expectation-Maximization (EM) procedure [14, 30, 38, 40, 74]. Similar procedures have been applied to cardiac MR segmentation [2] and multi-organ 3D segmentation [77]. Multi-model settings have also been explored, where predictions from an ensemble of models are aggregated into a single pseudo-label [34, 41], or where each model estimates the confidence on its prediction [67].

Consistency-based methods iteratively augment the set of labeled samples by including into it samples for which it is possible to generate pseudo-labels that are *consistent*, that is, robust to perturbations [4, 29, 46, 72]. The teacher-student approach uses two networks, a "student" one that generates predictions that are evaluated against the ones generated by the "teacher" network, whose weights are in turn updated as the moving average of the weights of the student network [33, 55].

We consider in particular a novel class of consistency-based methods based on exploiting both weak and strong sample perturbations, The first of these methods is FixMatch [50], which generates pseudo-labels for unlabeled samples by generating both a candidate label and a target one. The target label is obtained with a weak perturbation of the original sample and it is converted into a hot-one value if its probability exceeds a fixed threshold. The candidate label is, instead, obtained with a strong perturbation, and compared against the target. FixMatch keeps only pseudo-labels having sufficiently high predicted confidence. Originally proposed for image classification, FixMatch has also been applied to medical image segmentation [27, 58]. Similar works tackle instead semantic segmentation in different domains [18, 24, 78]. The choice of the threshold is however crucial, as it controls the trade-off between the quality of the results and the speed of convergence. The performance of FixMatch can also suffer in the case of scarcely annotated classes [37].

Several methods have been proposed to improve over FixMatch. Dash and FlexMatch use a dynamic threshold [69, 73]. FreeMatch uses class-specific adaptive thresholds [65]. Unimatch introduces feature perturbations on the weakly-augmented samples and applies strong perturbations in a contrastive fashion, selecting two augmentations at random from a pool of available ones [70]. AlphaMatch generates candidate predictions by aggregating multiple predictions obtained using an ensemble of strong and weak perturbations [21]. ComWin uses multiple networks to obtain pseudo-labels for the unlabeled samples and chooses the pseudo-labels for which the generating model displays the highest confidence [66]. UDA introduces unsupervised data augmentation and minimizes a divergence metric between the prediction for a sample and the one obtained when introducing some noise on the sample [68].

### 2.3. Active learning for image segmentation

Active learning iteratively proposes a subset of unlabeled samples to a human annotator, to be added to the labeled part of the training set. The objective is to minimize the cost of labeling by prioritizing the selection of highly-informative data to increase the performance of the model as much and as soon as possible.

Particularly, in active learning the model is first trained on a small initial set of labeled data referred to as the initial label pool. An acquisition function is then used to identify the samples that require annotation by an external oracle. Afterward, the newly labeled samples are added to the labeled pool, and the model is retrained using the updated training set. This process is repeated until the labeling budget is exhausted. Currently, research studying AL strategies mainly concentrates on low-dimensional annotation tasks such as image classification [20]. Some authors have used AL for segmentation of natural images [6], [7]. The general approach is to train a neural network model on the labeled data to infer annotations for the unlabeled samples. In the medical domain, a common selection criterion to choose the samples to be labeled is based on uncertainty and relevance, to include samples that are difficult to segment and include the most information possible. One possibility is to use bootstrap to compute the variability on the predictions [71]. Other authors observe the difference in prediction when a sample is altered with different augmentations [19]. The discrepancy is instead computed considering Class Activation Maps, which indicate what parts of the images are more relevant for the final decision, in [1]. In [47], the authors rank the samples by a score that combines the uncertainty of the prediction on the unlabeled samples and a similarity metric between the labeled and the unlabeled samples at the feature level. The authors of [31] use an ensemble of $k$ DL models to compute $k$ probability maps to estimate the uncertainty of the prediction. The samples highest ranked by uncertainty are subsequently ranked by discrepancy with respect to the average of the features computed by the ensemble of DL models, to select the smallest informative subset to be sent to a human annotator.

### 2.4. Combining semi-supervised and active learning

Active learning and semi-supervised learning can be combined in different ways to increase the efficiency of the labeling task. Authors in different fields introduce SSL in an AL framework for classification tasks, generating pseudo labels for the unlabeled samples and selecting for manual labeling those that have high inconsistency between the original sample and the corresponding augmentations [20, 57].

A similar procedure is applied to object detection, where the samples in the AL step are selected using criteria of uncertainty and diversity computed during the SSL step [44].

Other works have combined SSL and AL for image segmentation. In [76] the inferred pseudo-labels are ranked by uncertainty; the most uncertain ones are sent to a human annotator, while the least uncertain ones are reannotated by an ensemble of DL models. The combined dataset is used for fine-tuning before the subsequent pseudo-label generation. In [64] pseudo-labels are generated aggregating information from three projection heads for segmentation, detection, and classification, and ranked according to an uncertainty score. The highest-ranked samples, deemed the most informative ones, are sent to the human annotators for the AL step; the remaining pseudo-labels are kept for the next training. The mean teacher approach of [42] uses SSL in the teacher network to generate pseudo labels used to train the student network; the samples with the best performance are labeled and inserted in the labeled pool for the subsequent iteration, and the weights of the teacher network are updated as moving average of the weights of the student network.

Gigapixel histopathology images are considered in [28]. The images are divided into patches, and a FixMatch procedure is augmented with AL with the introduction, at each iteration, of patches whose segmented regions have the highest uncertainty.

## 3. Materials and methods

### 3.1. Dataset

We use the GlaS dataset for gland segmentation in colorectal cancer tissue [48, 49]. The dataset is composed of 165 images of Hematoxylin and Eosin (H&E) stained slides of different stages of colorectal adenocarcinoma annotated by expert pathologists based on the shape of the glands. 85 images are used for training (37 benign, 48 malignant). Originally, GlaS does not have a validation set, so we use

20% of the test set (16 images, 8 benign, 8 malignant) for validation and the remaining 64 (29 benign, 35 malignant) for testing. Each image is annotated pixel-wise, indicating whether each pixel belongs to the background (labeled as 0) or to a gland (label 1). We generate non-overlapping patches of size 256×256 pixels. Smaller patches are padded by reflecting the image.

**Dataset corruption** In order to test the AL and SSL strategies and their combination, we introduce corrupting the original dataset by removing annotations. We do not add imperfections in the annotations since it has been demonstrated that an imprecise segmentation in histopathology images does not have an impact as strong as missing objects of interest [16]. We remove 80% gland labels by assigning them a value of 0, thus confounding them with the background. Such a high level of noise ensures that the model's performance is heavily impacted and the possible effects of using SSL and AL techniques are noticeable enough. This label removal is performed on the whole image before splitting it into patches, so that we ensure that a gland is consistently removed across the different patches.

Regarding the initial dataset for SSL and AL strategies, we divide the corrupted training set into labeled and unlabeled subsets following the Only Positive approach proposed by Foucart *et al.* [17], keeping as labeled patches those that contain at least a part of an annotated gland. For this reason, some of the patches considered as labeled are actually partially annotated as they can have multiple (part of) glands present but not annotated as class 1 (gland) as a result of the corruption introduced. With the level of noise we consider, once the images are split, we remain with around 43% annotated patches, that is, 426 from 983 patches, with 36% of their original annotations.

### 3.2. Problem statement

We focus on a binary gland segmentation problem for histopathology images with missing annotations, where for some of the glands that are present in the original samples there is no corresponding annotation in the segmentation mask. Formally, the goal is to develop an algorithm that takes as input a set of patches of size 256 × 256 extracted from a Whole Sliding Image (WSI) and to compute a semantic segmentation mask by classifying each pixel as either background or belonging to a gland. The training data $\mathcal{D} = \mathcal{D}_l \cup \mathcal{D}_u$ contains a labeled set $\mathcal{D}_l = \{(\boldsymbol{x}^{l,1}, \boldsymbol{y}^1), \ldots, (\boldsymbol{x}^{l,N}, \boldsymbol{y}^N)\}$ and an unlabeled set by $\mathcal{D}_u = \{\boldsymbol{x}^{u,1}, \ldots, \boldsymbol{x}^{u,M}\}$, with $N$ potentially $\ll M$, where $\boldsymbol{x}^{l,i}$ and $\boldsymbol{y}^i$ are respectively the $i$-th labeled image patch and its corresponding mask (each pixel $x_{ij}$ has an associated value in the mask $y_{ij}$ of 0 for background and 1 for glands) of size $H_p \times W_p = 256 \times 256$, and $\boldsymbol{x}^{u,j}$ is the $j$-th unlabeled patch.

### 3.3. Combination of semi-supervised and active learning

We use the SSL+AL framework from [20] originally proposed for classification, and adapt it for segmentation. In a nutshell, this method embeds an SSL procedure in each AL iteration. Thanks to its generality we can use it to implement different SSL and AL strategies for respectively pseudo-label generation and relabeling, which can be provided as hyperparameters at runtime. It also allows for easy customization of weak and strong augmentations in the SSL step that are consistent with our dataset. The algorithmic framework is reported in Algorithm 1.

**Algorithm 1** Semi-supervised learning-based AL framework, adapted from [20].

```
Require: Dataset $\mathcal{D} = \mathcal{D}_l \cup \mathcal{D}_u$, AL strategy, SSL strategy,
  target cardinality of unlabeled data $\mathcal{K}_u$, selected sample
  batch set $B$, AL batch size $K$, segmentation model $M_0$
  randomly initialized at start.
  $B_0 \leftarrow \mathcal{D}_l$
  $U_0 \leftarrow \mathcal{D}_u$
  $L_0 \leftarrow \{(x, y) : x \in B_0\}$
  step $t = 0$, accuracy $A_0 = 0$
  while $|U_t| > \mathcal{K}_u$ do
      train $M_t$ using SSL strategy
      $A_t \leftarrow accuracy(M_t)$
      select $B_{t+1}, |B| = K$, from $U_t$ using AL strategy
      labeling $L_{t+1} \leftarrow L_t \cup \{(x, y) : x \in B_{t+1}\}$
      unlabeled pool update $U_{t+1} \leftarrow U_t \setminus B_{t+1}$
      $t = t + 1$
  end while
      return $M_{t-1}$
```

**Semi-supervised learning methods.** We consider three SSL methods, *FixMatch* and the derived *FlexMatch* and *FreeMatch*, whose general principles have been described in Section 2.2. These are recent state-of-the-art methods for classification, whose application to semantic segmentation in digital pathology has not been explored in depth. To the best of our knowledge, FlexMatch and FreeMatch in particular have not yet been applied to image segmentation.

Our FixMatch implementation for segmentation follows the approach of [58]. At each iteration $t$, the target model $M_t$ is learned by minimizing the loss function of the form $\mathcal{L}_l + \mathcal{L}_u$, where, at each batch, $\mathcal{L}_l$ and $\mathcal{L}_u$ are respectively supervised and unsupervised losses [25]. In the three cases, the supervised loss is computed as $\mathcal{L}_l = \mathcal{L}_{dice} + \mathcal{L}_{CE}$. The dice loss $\mathcal{L}_{dice}$ is

$$\mathcal{L}_{dice} = 1 - \frac{1}{N_t} \sum_{n=1}^{N_t} \frac{2 \sum_{ij} p_{ij,n} y_{ij,n} + 1}{\sum_{ij} p_{ij,n} + \sum_{ij} y_{ij,n} + 1} \tag{1}$$

where $\boldsymbol{p} = P\left(\boldsymbol{y} \mid \boldsymbol{x}^l\right) = \left(p_{ij}\right)_{i=\overline{1,H_p}, j=\overline{1,W_p}}$ is the softmax probability matrix predicted by the model. The cross-entropy loss $\mathcal{L}_{CE}$ is

$$\mathcal{L}_{CE} = \frac{1}{N_t} \sum_{n=1}^{N_t} \frac{1}{H_p \times W_p} \sum_{i,j} l\left(p_{ij,n}, y_{ij,n}\right), \tag{2}$$

where $l\left(\hat{y}, y\right) = y \log \hat{y} + (1 - y) \log\left(1 - \hat{y}\right)$, $\boldsymbol{p} = P\left(\boldsymbol{y} \mid \boldsymbol{x}^l\right) = \left(p_{ij}\right)_{ij=\overline{1,H_p},\overline{1,W_p}}$ is the softmax probability matrix predicted by the model for a sample $x$.

The unsupervised loss $\mathcal{L}_u$ is the cross-entropy loss computed according to the model's predicted class distribution given a weakly augmented version of a given unlabeled image $\boldsymbol{q}^\alpha = P\left(\boldsymbol{y} \mid \alpha\left(\boldsymbol{x}^u\right)\right)$. We use $\hat{\boldsymbol{q}}^\alpha = \arg\max\left(\boldsymbol{q}^\alpha\right)$ to denote a pseudo label. The $\mathcal{L}_u$ is computed as

$$\mathcal{L}_u = \frac{1}{\mu B} \frac{1}{H_p \times W_p} \sum_{m=1}^{\mu B} \sum_{ij} \mathbb{1}_{\left(max\left(q^\alpha_{ij,m}\right) \geq \tau\right)} l\left(\hat{q}^\alpha_{ij,m}, q^A_{ij,m}\right), \tag{3}$$

where $\mu$ the ratio between labeled and unlabeled samples in the batch, $\boldsymbol{q}^A = P\left(\boldsymbol{y} \mid A\left(\boldsymbol{x}^u\right)\right)$, $A\left(.\right)$ is the strong augmentation operation and $\tau$ denotes the threshold above which we take a pseudo-label.

The difference between the three SSL methods is in the value of $\tau$. In FixMatch $\tau$ is a given hyperparameter [50].

At each iteration $t_{\text{SSL}}$, FlexMatch uses a class-specific threshold $\tau_{t_{\text{SSL}}}(c)$ that makes the learning more lenient towards classes that are difficult to learn [73]. A coefficient $\sigma_{t_{\text{SSL}}}(c)$ for a class $c$ at SSL iteration $t_{\text{SSL}}$ is computed as

$$\begin{aligned} \sigma_{t_{\text{SSL}}}(c) = \quad & \textstyle\sum_{m=1}^{B} \sum_{ij} \mathbb{1}(\max(q^\alpha_{ij,m}) > \tau) \cdot \\ & \mathbb{1}(\hat{q}^\alpha_{ij,m} = c), \end{aligned} \tag{4}$$

This coefficient is normalized to $[0, 1]$ and used to rescale $\tau$. FlexMatch also includes a warm-up threshold when the number of unlabeled data that is unused is too high. The global flexible threshold is then expressed as

$$\tau_{t_{\text{SSL}}}(c) = \frac{\sigma_{t_{\text{SSL}}}(c)}{\max\left\{\max\limits_c \sigma_{t_{\text{SSL}}}, M - \sum_c \sigma_{t_{\text{SSL}}}\right\}} \cdot \tau. \tag{5}$$

FreeMatch can be considered a variant of FlexMatch whose threshold values reflect the stage of the training process [65]. The thresholds are set at low initial values to favor a higher acceptance of pseudo-labeled samples, and the values are progressively increased during the training to ensure that only high-quality pseudo-labels are kept. First, a global threshold $\tau_{t_{\text{SSL}}}$ is defined as the Exponential Moving Average (EMA) of the confidence at each training step, namely $1/C$ at the first iteration, with $C$ being the number of classes, and as $\lambda\tau_{t_{\text{SSL}}-1} + (1 - \lambda)\frac{1}{B}\frac{1}{H_p \times W_p}\sum_{b=1}^{B}\sum_{ij}(\hat{q}^\alpha_{ij,b})$ for the subsequent iterations,

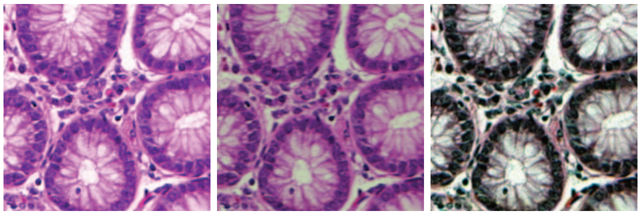

Figure 1: An example of, left to right, a basic, a weak, and a strong augmentation of a random patch.

where $\lambda \in (0, 1)$ is the momentum decay of EMA, and $B$ the batch size, $\hat{q}^\alpha_b$ the argmax pseudo-labels generated with a weak augmentation for the $b$-th sample in the batch.The class-specific thresholds are computed in a similar way. The expectation of the model prediction $\tilde{p}_{t_{\text{SSL}}}(c)$ for each class $c$ is computed as $1/C$ at the first iteration and $\lambda\tilde{p}_{t_{\text{SSL}}-1}(c) + (1 - \lambda)\frac{1}{B}\frac{1}{H_p \times W_p}\sum_{b=1}^{B}\sum_{ij} q^\alpha_{ij,b}(c)$ for the subsequent ones. This value is then normalized and rescaled by the global threshold as

$$\tau_{t_{SSL}}(c) = \frac{\tilde{p}_{t_{\text{SSL}}}(c)}{\max\limits_c \tilde{p}_{t_{\text{SSL}}}} \cdot \tau_{t_{\text{SSL}}}. \tag{6}$$

FreeMatch also introduces an additional term in the loss function called Self-Adaptive Fairness to take into account a possible class imbalance, which we implement as in [65] and do not describe in further detail for the sake of brevity.

As a baseline, we consider the case of not including a SSL strategy in a the AL iteration. In this case, the AL selection criterion operates on the pseudo-labels predicted by the model trained on the samples available with the Only Positive strategy.

**Augmentations.** The three SSL methods we use in our experiments are based on weak and strong augmentations. First, a base augmentation, composed of random horizontal and vertical flips, as well as random rotation by 90 and 180 degrees is applied to the image as well as the segmentation mask. The weak augmentation applies a color perturbation in the HED color space [36] and a random Gaussian blur to the base augmentation. The strong augmentation leverages the work from RandAugment[1] [13] to produce a random augmentation of 3 layers. We only consider the annotation that preserve geometry of the image, in order to have the same segmentation mask for the basic, weak, and strong augmentations. An example is given in Figure 1.

**Active learning methods.** The SSL methods leverage the unlabeled data by ensuring a consistent prediction between a weakly-transformed sample and the corresponding

[1] https://github.com/DIAGNijmegen/pathology-he-auto-augment

strongly-transformed one. Hence, the AL selection methods make use of the predicted pseudo-labels to select the samples to be shown to the annotator.

The first AL selection strategy, the *Consistency*-based method, is to choose samples that demonstrate highly inconsistent prediction with different distorted versions. Different from [20], we propose a metric to measure the inconsistency of predictions over a random set of distortion versions of a given image $\boldsymbol{x}$. Here, consistency is defined as the variance of the probability across different augmentations of the input, where all augmentations should have the same output.

$$\mathcal{E}\left(\boldsymbol{x}, M\right) = \frac{1}{H \times W} \sum_{i,j=1}^{H,W} \sum_{c=1}^{C} \mathrm{Var}\left[P_{ijc}^{\alpha}, P_{ijc}^{A_1}, \ldots, P_{ijc}^{A_{N_A}}\right] \tag{7}$$

where

$$P_{ijc}^{\alpha} = P\left(q_{ij}^{\alpha} = c \mid \alpha\left(x_{ij}\right), M\right) \tag{8}$$

$$P_{ijc}^{A} = P\left(q_{ij}^{\alpha} = c \mid A\left(x_{ij}\right), M\right) \tag{9}$$

and $C = 2$ is the number of classes, $x_{ij}$ is the pixel value at the position $i, j$ of the image $\boldsymbol{x}$ and $q_{ij}$ is its corresponding predicted label. $N_A$ is the number of augmentation on the original input image $\boldsymbol{x}$.

For batch selection, we aim to choose a batch $B$ such that the following consistency metric is maximized

$$\mathcal{C}\left(B, M\right) = \sum_{\boldsymbol{x} \in B} \mathcal{E}\left(\boldsymbol{x}, M\right) \tag{10}$$

The *Least Confidence* method computes the average probability for the predicted class of each pixel and selects the $K$ patches with the lowest confidence score. Given $\boldsymbol{q} = P(y_{ij} = q|\boldsymbol{x}^u)$ a probability distribution for the pseudo-labels of all pixels of a patch $\boldsymbol{x}^u$, the metric is defined as

$$S_{\text{conf}}(\boldsymbol{x}) = \frac{1}{H_p \cdot W_p} \sum_{ij} \max_{c} q_{ijc}^{\alpha} \tag{11}$$

The $K$ patches with the lowest $S_{\text{conf}}$ are selected for the labeling step.

The *Entropy*-based method computes the entropy of the class distribution for each pixel and selects the $K$ patches with the lowest entropy. It is defined as

$$S_{\text{entr}}(\boldsymbol{x}) = \frac{1}{H_p \cdot W_p} \sum_{ij} \sum_{c} q_{ijc}^{\alpha} \cdot \log q_{ijc}^{\alpha}. \tag{12}$$

The $K$ patches with the lowest $S_{\text{entr}}$ are selected for the labeling step.

The baseline AL method is the *random selection* without replacement of a set of $K$ patches to be annotated at each iteration.

# 4. Experiments

## 4.1. Training setup

The model that we use as the base segmentation model during all our experiments is the U-Net, proposed by Ronneberger *et al.* [45], a popular model for segmentation.

The training is performed using *Stochastic Gradient Descent* (SGD) with a learning rate of $10^{-1}$ and a momentum coefficient of $0.9$. The learning rate is decreased to $10^{-2}$ at epoch $80$, and then to $10^{-3}$ at epoch $120$. Each iteration has a maximum of $200$ epochs. To avoid overfitting we stop the training using an early stopping strategy[2]. The early stopping counter starts at epoch $120$ and has a patience parameter of $30$ epochs with a delta of $0.001$. The early stopping is combined with callback based on the validation loss, which allows the retrieval of the last weights that generalize well on the validation set.

For the SSL approaches, in every batch we use one set of $4$ supervised samples, one set of $4$ weakly augmented samples, and one set of $4$ strongly augmented samples. For fully supervised learning, we use a batch of $12$.

The initial training set (before any active learning query) contains all the non-empty corrupted patches as described in Section 3.1. A first model is trained on this dataset using the settings described above. Then, the *active learning loop* goes as follows: using a *query strategy*, query $K = 0.05 \cdot |\mathcal{D}| = 49$ unlabeled samples to be labeled, add them to the labeled dataset, and train a new model (using the settings described above) on the updated dataset for the next iteration of the loop. As we want to evaluate the impact of the noise, we stop when all the samples have been added to $\mathcal{D}_l$, for a total of 13 iterations. We simulate the annotation by an expert by using the original ground truth mask. Therefore, the patches that we introduce after the labeling phase are fully annotated. The patches in $\mathcal{D}_l$ after the corruption remain unchanged throughout the whole process.

The different setups we want to try for our experiments are: to train the segmentation model in a fully supervised way using the corrupted dataset, train the model using only the different SSL methods, train the model using only the different AL strategies, and then finally train the segmentation model with the combination of the AL loop enhanced with an SSL training.

We evaluate the 16 combinations of AL and SSL methods (including the baselines). We compare the results also with fully supervised trainings on the original and corrupted datasets. The code and the material to reproduce our experiments are available online.[3]

[2] Implementation based on the code available in https://github.com/Bjarten/early-stopping-pytorch

[3] https://github.com/luciledierckx/Histopathology_Seg_SSL_AL.

### 4.2. Results and discussion

In our experiments, we evaluate the effect of AL, SSL and their combination on the corrupted dataset. In particular, we want to observe whether targeted strategies outperform the respective baselines (random selection, no SSL strategy, supervised learning). Starting from a high noise level, we observe also how much data needs to be annotated before the model performance becomes sufficiently good.

The metrics we consider for evaluation are the *Dice Score* (DSC)

$$\text{DSC} = \frac{1}{N}\sum_{n=1}^{N}\frac{2\sum\limits_{ij} p_{ij,n} y_{ij,n} + 1}{\sum\limits_{ij} p_{ij,n} + \sum\limits_{ij} y_{ij,n} + 1} \tag{13}$$

and the *Matthews Correlation Coefficient* (MCC)

$$\text{MCC} = \frac{TN \times TP + FN \times FP}{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}. \tag{14}$$

The results in terms of Dice Score on the test set are reported in Figure 2, divided by AL method. We include the results obtained using supervised learning on the original and corrupted dataset as a reference for the quality of the results. The MCC scores correlate perfectly with the DSC ones, so we do not report them here for the sake of brevity. The plots with the MCC scores and the results on the test set aggregated by SSL method are reported in the Supplementary Material.

In each plot of Figure 2 we therefore compare the performance of the different SSL methods. The results of the SSL methods alone on the corrupted dataset correspond to the first point in the respective curves in the plot.

Overall, the AL methods obtain results very close to each other, and even the random selection performs as well as its alternatives. A likely explanation for this phenomenon is the “cold start” problem of AL, when an initial dataset that is too small or too unbalanced causes the sampling strategy to be highly biased, and thus to select suboptimal samples for annotation [20, 23]. While no selection strategy drastically underperforms, there is no definitive best one either. The supervised approach with 80% of noise perform similarly to the only non-trainable method in [48], and all other studies approaches perform significantly better.

Concerning the SSL strategies, from our results in Figure 2 it appears evident that using an SSL strategy improves the performance over the fully supervised training, even though the results of FixMatch and FreeMatch are very close to those obtained when using the simple Only Positive strategy. We note how our results for FixMatch at its first iteration are consistent with those reported by [28] on another dataset. These results could be due to the fact that models developed in a domain often do not translate well into other domains, as discussed in [9]. In particular, while these SSL methods are state-of-the-art in classification, more efforts are needed to improve their performance on segmentation tasks, in particular in the case of histopathology images.

Surprisingly, FlexMatch obtains much poorer results than its alternatives even when very high portions of the dataset have been annotated. This fact suggests that its threshold updating rule is neither robust enough (as the fixed FixMatch value) nor flexible enough (as the FreeMatch one) to perform well on this task and kind of data. In particular, the method suffers the most from the small portion of initial annotations and fails to recover. In the Supplementary Material we include the metrics recorded for training and validation, from which we observe how FlexMatch fails to converge, resulting in a variability of the results on the test set. A longer runtime did not prove beneficial either. The threshold update rule of FreeMatch, which adapts the threshold values with respect to the previous iteration, rather than starting from the initial value, is clearly superior in terms of robustness, at least in the context considered in this work.

Finally, and consequently, no combination of SSL+AL methods stands out from the rest of the pool. No AL selection method benefits from the pseudo-labels generated by an SSL method. On the contrary, SSL methods benefit from the additional labeled samples provided by the AL selection, getting similar results with all the AL strategies.

The relatively good performance obtained by the Only Positive strategy with respect to its more complex alternatives confirms the observations of [16], where this strategy outperformed more sophisticated techniques for pseudo-label generation as SSL. Nevertheless, adding AL to the Only Positive strategy improves the model’s performance after the second or the third iteration. Considering the computational cost of the SSL procedures (in particular for FreeMatch), the Only Positive strategy, possibly combined with Active Learning, appears to be a method of choice for partially annotated histopathology datasets.

The small difference in performance between the SSL+AL methods and the fully supervised learning on the complete dataset can be explained by the absence of a significant amount of annotations in the initial labeled pool, as explained in section 3.1.

## 5. Conclusion

Image segmentation is one of the fundamental tasks in digital pathology. Modern deep learning approaches require a large amount of annotated data, which are difficult to obtain in many real-life situations due to the time-consuming effort demanded to experts to provide them. In this work, we have provided a computational evaluation of two techniques aimed at alleviating this issue by making use of non-annotated samples, active and semi-supervised learning, and their combination. In particular, as SSL methods

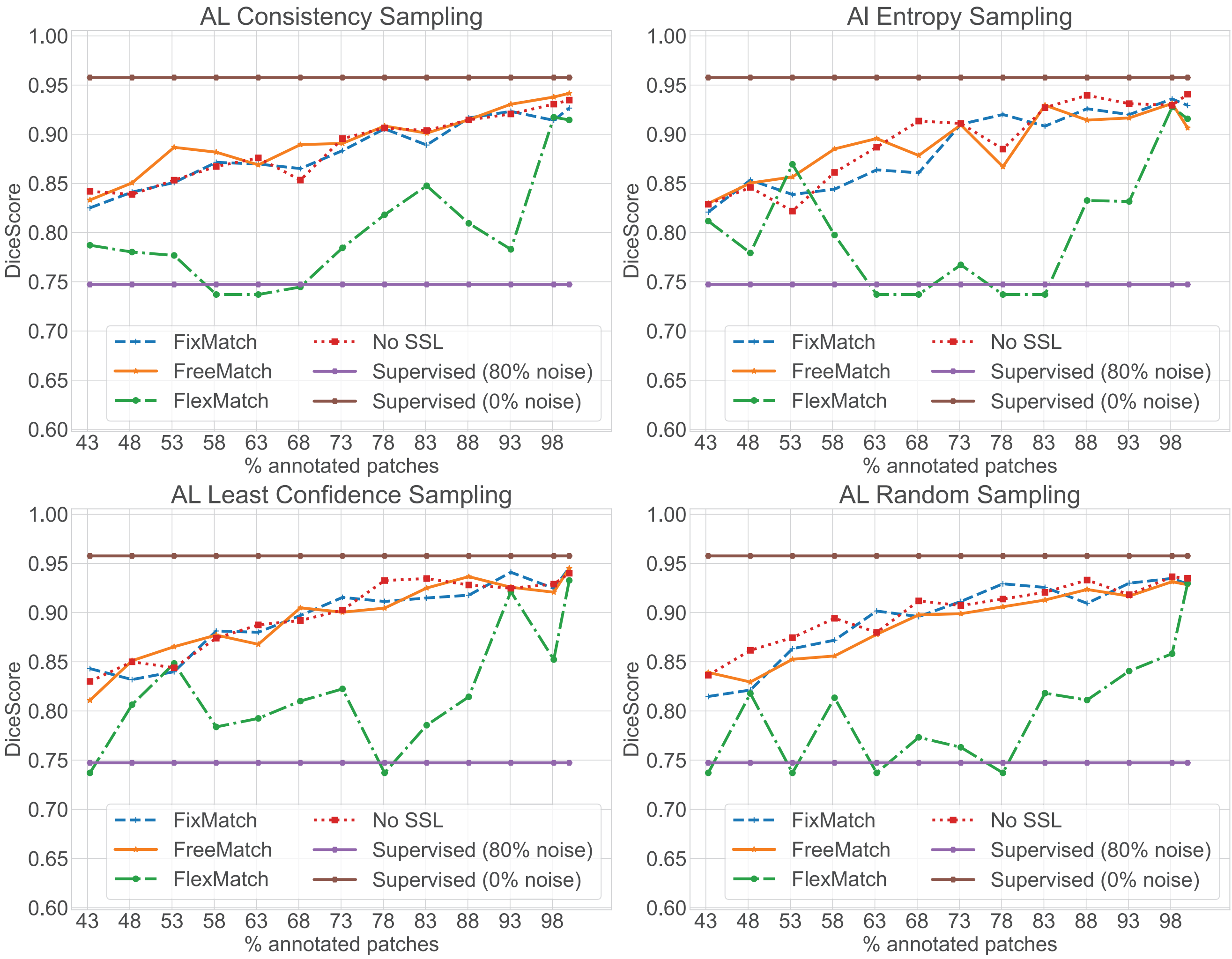

Figure 2: Dice Score on the test set for the four AL selection methods. In each plot, we include the performance obtained by combining each AL method with our pool of SSL methods. We include the results obtained by a fully supervised training on the noisy dataset as a lower bound and by a fully supervised training on the complete set of annotations as an upper bound.

we use the FixMatch for segmentation and we adapt FlexMatch and FreeMatch for this task, and four different AL strategies, namely consistency-based, entropy-based, random, and least confidence. We studied the impact of these methods on the GlaS dataset, from which we removed annotations in a controlled manner to simulate the errors made during the annotation process. Our results indicate that a simple Only Positive strategy performs as well as a recent family of consistency-based SSL methods, and can therefore be considered as a valid, cheaper alternative. Using AL with the Only Positive strategy improves the model's performance after a few iterations over using just the Only Positive strategy. However, an AL random strategy obtains the same results as more elaborated selection methodologies, such as entropy-based or consistency-based strategies.

In addition to the SSL and AL strategies we considered, more different strategies could be tested, as well as other types of combinations, such as the embedding of an AL selection strategy in an SSL procedure. We can also perform our evaluations on different segmentation tasks, such as multiclass nuclei segmentation.

## Acknowledgements

This work was supported by Service Public de Wallonie Recherche under grant n°2010235 - ARIAC by DIGITALWALLONIA4.AI. Computational resources have been provided by the Consortium des Équipements de Calcul Intensif (CÉCI), funded by the Fonds de la Recherche Scientifique de Belgique (F.R.S.-FNRS) under Grant No. 2.5020.11 and by the Walloon Region. We thank Prof. Christine Decaestecker and Charlotte Nachtegel for the useful discussions and comments.

## References

[1] Fan Bai, Xiaohan Xing, Yutian Shen, Han Ma, and Max Q-H Meng. Discrepancy-based active learning for weakly supervised bleeding segmentation in wireless capsule endoscopy images. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part VIII*, pages 24–34. Springer, 2022. 3

[2] Wenjia Bai, Ozan Oktay, Matthew Sinclair, Hideaki Suzuki, Martin Rajchl, Giacomo Tarroni, Ben Glocker, Andrew King, Paul M Matthews, and Daniel Rueckert. Semi-supervised learning for network-based cardiac MR image segmentation. In *Medical Image Computing and Computer-Assisted Intervention- MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part II 20*, pages 253–260. Springer, 2017. 2

[3] 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 *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*, pages 591–595. IEEE, 2017. 1

[4] Gerda Bortsova, Florian Dubost, Laurens Hogeweg, Ioannis Katramados, and Marleen De Bruijne. Semi-supervised medical image segmentation via learning consistency under transformations. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part VI 22*, pages 810–818. Springer, 2019. 2

[5] Gloria Bueno, M Milagro Fernandez-Carrobles, Lucia Gonzalez-Lopez, and Oscar Deniz. Glomerulosclerosis identification in whole slide images using semantic segmentation. *Computer methods and programs in biomedicine*, 184:105273, 2020. 2

[6] Lile Cai, Xun Xu, Jun Hao Liew, and Chuan Sheng Foo. Revisiting superpixels for active learning in semantic segmentation with realistic annotation costs. In *Conference on Computer Vision and Pattern Recognition*, pages 10988–10997. IEEE, 2021. 3

[7] Arantxa Casanova, Pedro O Pinheiro, Negar Rostamzadeh, and Christopher J Pal. Reinforced active learning for image segmentation. *arXiv preprint arXiv:2002.06583*, 2020. 3

[8] Krishna Chaitanya, Neerav Karani, Christian F Baumgartner, Anton Becker, Olivio Donati, and Ender Konukoglu. Semi-supervised and task-driven data augmentation. In *Information Processing in Medical Imaging: 26th International Conference, IPMI 2019, Hong Kong, China, June 2–7, 2019, Proceedings 26*, pages 29–41. Springer, 2019. 2

[9] 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*, 129:361–384, 2021. 2, 7

[10] Lyndon Chan, Mahdi S Hosseini, Corwyn Rowsell, Konstantinos N Plataniotis, and Savvas Damaskinos. Histosegnet: Semantic segmentation of histological tissue type in whole slide images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10662–10671, 2019. 2

[11] Veronika Cheplygina, Marleen de Bruijne, and Josien PW Pluim. Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. *Medical image analysis*, 54:280–296, 2019. 2

[12] Patrick Ferdinand Christ, Mohamed Ezzeldin A Elshaer, Florian Ettlinger, Sunil Tatavarty, Marc Bickel, Patrick Bilic, Markus Rempfler, Marco Armbruster, Felix Hofmann, Melvin D'Anastasi, et al. Automatic liver and lesion segmentation in ct using cascaded fully convolutional neural networks and 3d conditional random fields. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19*, pages 415–423. Springer, 2016. 2

[13] Ekin Dogus Cubuk, Barret Zoph, Jon Shlens, and Quoc Le. Randaugment: Practical automated data augmentation with a reduced search space. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 18613–18624. Curran Associates, Inc., 2020. 5

[14] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977. 2

[15] Shujian Deng, Xin Zhang, Wen Yan, Eric I-Chao Chang, Yubo Fan, Maode Lai, and Yan Xu. Deep learning in digital pathology image analysis: a survey. *Frontiers of medicine*, 14:470–487, 2020. 2

[16] Adrien Foucart, Olivier Debeir, and Christine Decaestecker. SNOW: semi-supervised, noisy and/or weak data for deep learning in digital pathology. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pages 1869–1872. IEEE, 2019. 2, 4, 7

[17] Adrien Foucart, Olivier Debeir, and Christine Decaestecker. SNOW supervision in digital pathology: managing imperfect annotations for segmentation in deep learning. Submitted, 2020. 2, 4

[18] Geoffrey French, Samuli Laine, Timo Aila, Michal Mackiewicz, and Graham Finlayson. Semi-supervised semantic segmentation needs strong, varied perturbations. In *British Machine Vision Conference*, 2020. 2

[19] Mélanie Gaillochet, Christian Desrosiers, and Hervé Lombaert. Taal: Test-time augmentation for active learning in medical image segmentation. In *Data Augmentation, Labelling, and Imperfections: Second MICCAI Workshop, DALI 2022, Held in Conjunction with MICCAI 2022, Singapore, September 22, 2022, Proceedings*, pages 43–53. Springer, 2022. 3

[20] 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. 3, 4, 6, 7

[21] Chengyue Gong, Dilin Wang, and Qiang Liu. Alphamatch: Improving consistency for semi-supervised learning with alpha-divergence. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13683–13692, 2021. 3

[22] Yanming Guo, Yu Liu, Theodoros Georgiou, and Michael S Lew. A review of semantic segmentation using deep neural networks. *International journal of multimedia information retrieval*, 7:87–93, 2018. 2

[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. 1, 7

[24] Hanzhe Hu, Fangyun Wei, Han Hu, Qiwei Ye, Jinshi Cui, and Liwei Wang. Semi-supervised semantic segmentation via adaptive equalization learning. *Advances in Neural Information Processing Systems*, 34:22106–22118, 2021. 2

[25] Fabian Isensee, Jens Petersen, Andre Klein, David Zimmerer, Paul F Jaeger, Simon Kohl, Jakob Wasserthal, Gregor Koehler, Tobias Norajitra, Sebastian Wirkert, et al. nnu-net: Self-adapting framework for u-net-based medical image segmentation. *arXiv preprint arXiv:1809.10486*, 2018. 4

[26] Sara Hosseinzadeh Kassani, Peyman Hosseinzadeh Kassani, Michal J Wesolowski, Kevin A Schneider, and Ralph Deters. Deep transfer learning based model for colorectal cancer histopathology segmentation: A comparative study of deep pre-trained models. *International Journal of Medical Informatics*, 159:104669, 2022. 2

[27] 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 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 1920–1923. IEEE, 2021. 2

[28] Zhengfeng Lai, Chao Wang, Luca Cerny Oliveira, Brittany N Dugger, Sen-Ching Cheung, and Chen-Nee Chuah. Joint semi-supervised and active learning for segmentation of gigapixel pathology images with cost-effective labeling. In *International Conference on Computer Vision*, pages 591–600. IEEE, 2021. 1, 3, 7

[29] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. In *International Conference on Learning Representations*, 2017. 2

[30] Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, page 896, 2013. 2

[31] Haohan Li and Zhaozheng Yin. Attention, suggestion and annotation: a deep active learning framework for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer Assisted Intervention–MICCAI*, pages 3–13. Springer, 2020. 3

[32] Jiayun Li, Karthik V Sarma, King Chung Ho, Arkadiusz Gertych, Beatrice S Knudsen, and Corey W Arnold. A multi-scale u-net for semantic segmentation of histological images from radical prostatectomies. In *AMIA Annual Symposium Proceedings*, volume 2017, page 1140. American Medical Informatics Association, 2017. 2

[33] Xiaomeng Li, Lequan Yu, Hao Chen, Chi-Wing Fu, Lei Xing, and Pheng-Ann Heng. Transformation-consistent self-ensembling model for semisupervised medical image segmentation. *IEEE Transactions on Neural Networks and Learning Systems*, 32(2):523–534, 2020. 2

[34] Shaobo Min, Xuejin Chen, Zheng-Jun Zha, Feng Wu, and Yongdong Zhang. A two-stream mutual attention network for semi-supervised biomedical segmentation with noisy labels. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence*, pages 4578–4585, 2019. 2

[35] Shervin Minaee, Yuri Boykov, Fatih Porikli, Antonio Plaza, Nasser Kehtarnavaz, and Demetri Terzopoulos. Image segmentation using deep learning: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(7):3523–3542, 2021. 1

[36] Romain Mormont, Mehdi Testouri, Raphaël Marée, and Pierre Geurts. Relieving pixel-wise labeling effort for pathology image segmentation with self-training. In *European Conference in Compter Vision (ECCV2022)*. Springer Cham, Genève, Switzerland, 2022. 5

[37] 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, 2018. 3

[38] George Papandreou, Liang-Chieh Chen, Kevin P Murphy, and Alan L Yuille. Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 1742–1750, 2015. 2

[39] Olivier Petit, Nicolas Thome, Arnaud Charnoz, Alexandre Hostettler, and Luc Soler. Handling missing annotations for semantic segmentation with deep convnets. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4*, pages 20–28. Springer, 2018. 2

[40] Emmanouil Antonios Platanios, Maruan Al-Shedivat, Eric Xing, and Tom Mitchell. Learning from imperfect annotations. *arXiv preprint arXiv:2004.03473*, 2020. 2

[41] Ilija Radosavovic, Piotr Dollár, Ross Girshick, Georgia Gkioxari, and Kaiming He. Data distillation: Towards omni-supervised learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4119–4128, 2018. 2

[42] Aneesh Rangnekar, Christopher Kanan, and Matthew Hoffman. Semantic segmentation with active semi-supervised learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5966–5977, 2023. 3

[43] Hatice Catal Reis and Veysel Turk. Transfer learning approach and nucleus segmentation with medclnet colon cancer database. *Journal of Digital Imaging*, 36(1):306–325, 2023. 2

[44] 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. 3

[45] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015. 2, 6

[46] Mehdi Sajjadi, Mehran Javanmardi, and Tolga Tasdizen. Regularization with stochastic transformations and perturbations for deep semi-supervised learning. *Advances in neural information processing systems*, 29, 2016. 2

[47] Dhruv Sharma, Zahil Shanis, Chandan K Reddy, Samuel Gerber, and Andinet Enquobahrie. Active learning technique for multimodal brain tumor segmentation using limited labeled images. In *Domain Adaptation and Representation Transfer and Medical Image Learning with Less Labels and Imperfect Data: First MICCAI Workshop, DART 2019, and First International Workshop, MIL3ID 2019, Held in Conjunction with MICCAI*, pages 148–156. Springer, 2019. 3

[48] Korsuk Sirinukunwattana, Josien PW Pluim, Hao Chen, Xiaojuan Qi, Pheng-Ann Heng, Yun Bo Guo, Li Yang Wang, Bogdan J Matuszewski, Elia Bruni, Urko Sanchez, et al. Gland segmentation in colon histology images: The glas challenge contest. *Medical image analysis*, 35:489–502, 2017. 1, 2, 3, 7

[49] Korsuk Sirinukunwattana, David RJ Snead, and Nasir M Rajpoot. A stochastic polygons model for glandular structures in colon histology images. *IEEE transactions on medical imaging*, 34(11):2366–2378, 2015. 2, 3

[50] 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. *Advances in neural information processing systems*, 33:596–608, 2020. 2, 5

[51] Shuang Song, David Berthelot, and Afshin Rostamizadeh. Combining mixmatch and active learning for better accuracy with fewer labels. *arXiv preprint arXiv:1912.00594*, 2019. 1

[52] Chetan L Srinidhi, Ozan Ciga, and Anne L Martel. Deep neural network models for computational histopathology: A survey. *Medical Image Analysis*, 67:101813, 2021. 2

[53] Nima Tajbakhsh, Laura Jeyaseelan, Qian Li, Jeffrey N Chiang, Zhihao Wu, and Xiaowei Ding. Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation. *Medical Image Analysis*, 63:101693, 2020. 1, 2

[54] You-Bao Tang, Sooyoun Oh, Yu-Xing Tang, Jing Xiao, and Ronald M Summers. Ct-realistic data augmentation using generative adversarial network for robust lymph node segmentation. In *Medical Imaging 2019: Computer-Aided Diagnosis*, volume 10950, pages 976–981. SPIE, 2019. 2

[55] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017. 2

[56] Hamid Reza Tizhoosh and Liron Pantanowitz. Artificial intelligence and digital pathology: challenges and opportunities. *Journal of pathology informatics*, 9(1):38, 2018. 1

[57] Katrin Tomanek and Udo Hahn. Semi-supervised active learning for sequence labeling. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2*, page 1039–1047, USA, 2009. Association for Computational Linguistics. 3

[58] Pratima Upretee and Bishesh Khanal. Fixmatchseg: Fixing fixmatch for semi-supervised semantic segmentation. *arXiv preprint arXiv:2208.00400*, 2022. 2, 4

[59] Jeroen Van der Laak, Geert Litjens, and Francesco Ciompi. Deep learning in histopathology: the path to the clinic. *Nature medicine*, 27(5):775–784, 2021. 2

[60] Jesper E. van Engelen and Holger H. Hoos. A survey on semi-supervised learning. *Machine Learning*, 109:373 – 440, 2019. 1

[61] Yves-Rémi Van Eycke, Adrien Foucart, and Christine Decaestecker. Strategies to reduce the expert supervision required for deep learning-based segmentation of histopathological images. *Frontiers in medicine*, page 222, 2019. 2

[62] Noorul Wahab, Asifullah Khan, and Yeon Soo Lee. Transfer learning based deep cnn for segmentation and detection of mitoses in breast cancer histopathological images. *Microscopy*, 68(3):216–233, 2019. 2

[63] Jiazhuo Wang, John D MacKenzie, Rageshree Ramachandran, and Danny Z Chen. A deep learning approach for semantic segmentation in histology tissue images. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19*, pages 176–184. Springer, 2016. 2

[64] Jun Wang, Shaoguo Wen, Kaixing Chen, Jianghua Yu, Xin Zhou, Peng Gao, Changsheng Li, and Guotong Xie. Semi-supervised active learning for instance segmentation via scoring predictions. In *British Machine Vision Conference (BMVC2020)*, 2020. 1, 3

[65] Yidong Wang, Hao Chen, Qiang Heng, Wenxin Hou, Marios Savvides, Takahiro Shinozaki, Bhiksha Raj, Zhen Wu, and Jindong Wang. Freematch: Self-adaptive thresholding for semi-supervised learning. *arXiv preprint arXiv:2205.07246*, 2022. 2, 3, 5

[66] Huimin Wu, Xiaomeng Li, Yiqun Lin, and Kwang-Ting Cheng. Compete to win: Enhancing pseudo labels for barely-supervised medical image segmentation. *IEEE Transactions on Medical Imaging*, 2023. 3

[67] Yingda Xia, Fengze Liu, Dong Yang, Jinzheng Cai, Lequan Yu, Zhuotun Zhu, Daguang Xu, Alan Yuille, and Holger Roth. 3d semi-supervised learning with uncertainty-aware multi-view co-training. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3646–3655, 2020. 2

[68] Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. Unsupervised data augmentation for consistency training. *Advances in neural information processing systems*, 33:6256–6268, 2020. 3

[69] Yi Xu, Lei Shang, Jinxing Ye, Qi Qian, Yu-Feng Li, Baigui Sun, Hao Li, and Rong Jin. Dash: Semi-supervised learning with dynamic thresholding. In *International Conference on Machine Learning*, pages 11525–11536. PMLR, 2021. 3

[70] Lihe Yang, Lei Qi, Litong Feng, Wayne Zhang, and Yinghuan Shi. Revisiting weak-to-strong consistency in semi-supervised semantic segmentation. In *CVPR*, 2023. 3

[71] Lin Yang, Yizhe Zhang, Jianxu Chen, Siyuan Zhang, and Danny Z Chen. Suggestive annotation: A deep active learning framework for biomedical image segmentation. In *Medical Image Computing and Computer Assisted Intervention-MICCAI*, pages 399–407. Springer, 2017. 3

[72] Lequan Yu, Shujun Wang, Xiaomeng Li, Chi-Wing Fu, and Pheng-Ann Heng. Uncertainty-aware self-ensembling model for semi-supervised 3d left atrium segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part II 22*, pages 605–613. Springer, 2019. 2

[73] Bowen Zhang, Yidong Wang, Wenxin Hou, Hao Wu, Jindong Wang, Manabu Okumura, and Takahiro Shinozaki. Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling. *Advances in Neural Information Processing Systems*, 34:18408–18419, 2021. 2, 3, 5

[74] Ling Zhang, Vissagan Gopalakrishnan, Le Lu, Ronald M Summers, Joel Moss, and Jianhua Yao. Self-learning to detect and segment cysts in lung ct images without manual annotation. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 1100–1103. IEEE, 2018. 2

[75] Ling Zhang, Xiaosong Wang, Dong Yang, Thomas Sanford, Stephanie Harmon, Baris Turkbey, Holger Roth, Andriy Myronenko, Daguang Xu, and Ziyue Xu. When unseen domain generalization is unnecessary? rethinking data augmentation. *arXiv preprint arXiv:1906.03347*, 2019. 2

[76] Ziyuan Zhao, Zeng Zeng, Kaixin Xu, Cen Chen, and Cuntai Guan. Dsal: Deeply supervised active learning from strong and weak labelers for biomedical image segmentation. *IEEE journal of biomedical and health informatics*, 25(10):3744–3751, 2021. 3

[77] Yuyin Zhou, Yan Wang, Peng Tang, Song Bai, Wei Shen, Elliot Fishman, and Alan Yuille. Semi-supervised 3d abdominal multi-organ segmentation via deep multi-planar co-training. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 121–140. IEEE, 2019. 2

[78] Yuliang Zou, Zizhao Zhang, Han Zhang, Chun-Liang Li, Xiao Bian, Jia-Bin Huang, and Tomas Pfister. Pseudoseg: Designing pseudo labels for semantic segmentation. *arXiv preprint arXiv:2010.09713*, 2020. 2