Multiple Instance Learning via Iterative Self-Paced Supervised Contrastive Learning

Kangning Liu*[1], Weicheng Zhu*[1], Yiqiu Shen[1], Sheng Liu[1], Narges Razavian[2]
Krzysztof J. Geras†[2,1], Carlos Fernandez-Granda†[1,3]
[1]NYU Center for Data Science [2]NYU Grossman School of Medicine
[3] Courant Institute of Mathematical Sciences

Abstract

Learning representations for individual instances when only bag-level labels are available is a fundamental challenge in multiple instance learning (MIL). Recent works have shown promising results using contrastive self-supervised learning (CSSL), which learns to push apart representations corresponding to two different randomly-selected instances. Unfortunately, in real-world applications such as medical image classification, there is often class imbalance, so randomly-selected instances mostly belong to the same majority class, which precludes CSSL from learning inter-class differences. To address this issue, we propose a novel framework, Iterative Self-paced Supervised Contrastive Learning for MIL Representations (ItS2CLR), which improves the learned representation by exploiting instance-level pseudo labels derived from the bag-level labels. The framework employs a novel self-paced sampling strategy to ensure the accuracy of pseudo labels. We evaluate ItS2CLR on three medical datasets, showing that it improves the quality of instance-level pseudo labels and representations, and outperforms existing MIL methods in terms of both bag and instance level accuracy. [1]

1. Introduction

The goal of multiple instance learning (MIL) is to perform classification on data that is arranged in *bags* of instances. Each instance is either positive or negative, but these instance-level labels are not available during training; only bag-level labels are available. A bag is labeled as positive if *any* of the instances in it are positive, and negative otherwise. An important application of MIL is cancer diagnosis from histopathology slides. Each slide is divided into hundreds or thousands of tiles but typically only slide-level labels are available [6, 9, 17, 35, 39, 53].

Histopathology slides are typically very large, in the order of gigapixels (the resolution of a typical slide can be as high as $10^5 \times 10^5$), so end-to-end training of deep neural networks is typically infeasible due to memory limitations of GPU hardware. Consequently, state-of-the-art approaches [6, 35, 39, 44, 53] utilize a two-stage learning pipeline: (1) a feature-extraction stage where each instance is mapped to a representation which summarizes its content, and (2) an aggregation stage where the representations extracted from all instances in a bag are combined to produce a bag-level prediction (Figure 1). Notably, our results indicate that even in the rare settings where end-to-end training is possible, this pipeline still tends to be superior (see Section 4.3).

In this work, we focus on a fundamental challenge in MIL: how to train the feature extractor. Currently, there are three main strategies to perform feature-extraction, which have significant shortcomings. (1) Pretraining on a large natural image dataset such as ImageNet [39, 44] is problematic for medical applications because features learned from natural images may generalize poorly to other domains [38]. (2) Supervised training using bag-level labels as instance-level labels is effective if positive bags contain mostly positive instances [11, 34, 50], but in many medical datasets this is not the case [5, 35]. (3) Contrastive self-supervised learning (CSSL) outperforms prior methods [14, 35], but is not as effective in settings with heavy class imbalance, which are of crucial importance in medicine. CSSL operates by pushing apart the representations of different randomly selected instances. When positive bags contain mostly negative instances, CSSL training ends up pushing apart negative instances from each other, which precludes it from learning features that distinguish positive samples from the negative ones (Figure 2). We discuss this finding in Section 2.

Our goal is to address the shortcomings of current feature-extraction methods. We build upon several key insights. First, it is possible to extract instance-level pseudo

---

[*] Equal Contribution

[†] Joint Last Author

[1] Code is available at https://github.com/Kangningthu/ItS2CLR

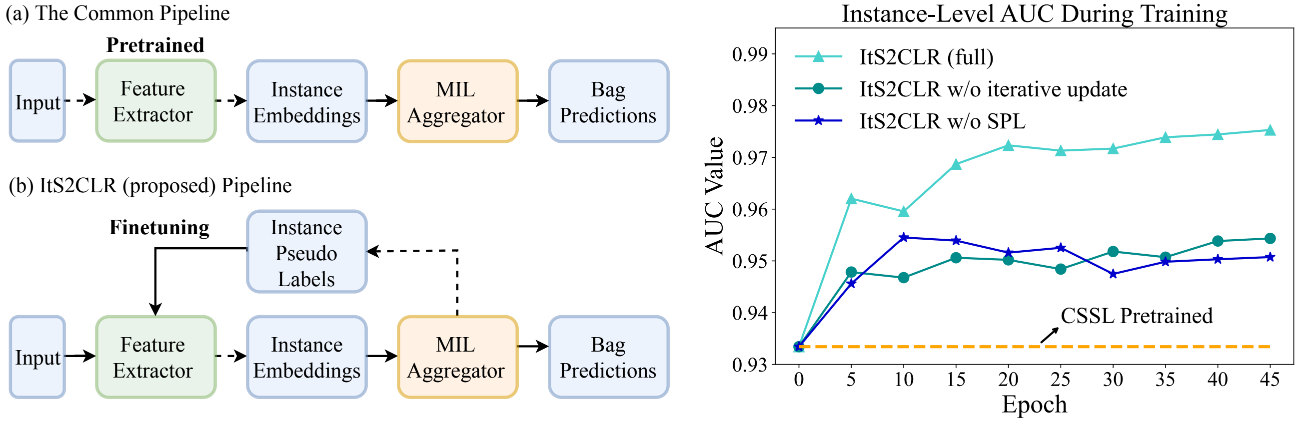

Figure 1. **Left:** (a) Commonly used deep MIL models first pretrain a feature extractor and then train an aggregator that maps the representations to a bag-level prediction. (b) Our proposed framework, ItS2CLR, uses instance-level pseudo labels obtained from the aggregator to finetune the feature extractor. ItS2CLR updates the features iteratively based on the pseudo label of a subset of instances selected according to a self-paced learning (SPL) strategy. **Right:** The dashed line is the instance-level AUC of the MIL model trained on instance feature extracted by the CSSL pretrained feature extractor. On a benchmark dataset (Camelyon16 [5]), the iterative finetuning process gradually improves the instance-level AUC during training, which results in more accurate pseudo labels. Both the iterative updates and SPL are important to achieve this.

labels from trained MIL models, which are more accurate than assigning the bag-level labels to all instances within a positive bag. Second, we can use the pseudo labels to finetune the feature extractor, improving the instance-level representations. Third, these improved representations result in improved bag-level classification and more accurate instance-level pseudo labels. These observations are utilized in our proposed framework, Iterative Self-Paced Supervised Contrastive Learning for MIL Representation (ItS2CLR), as illustrated in Figure 1. After initializing the features with CSSL, we iteratively improve them via supervised contrastive learning [32] using pseudo labels inferred by the aggregator. This feature refinement utilizes pseudo labels sampled according to a novel self-paced strategy, which ensures that they are sufficiently accurate (see Section 3.2). In summary, our contributions are the following:

1. We propose ItS2CLR – a novel MIL framework where instance features are iteratively improved using pseudo labels extracted from the MIL aggregator. The framework combines supervised contrastive learning with a self-paced sampling scheme to ensure that pseudo labels are accurate.
2. We demonstrate that the proposed approach outperforms existing MIL methods in terms of bag- and instance-level accuracy on three real-world medical datasets relevant to cancer diagnosis: two histopathology datasets and a breast ultrasound dataset. It also outperforms alternative finetuning methods, such as instance-level cross-entropy minimization and end-to-end training.
3. In a series of controlled experiments, we show that ItS2CLR is effective when applied to different feature-extraction architectures and when combined with different aggregators.

2. CSSL May Not Learn Discriminative Representations In MIL

Recent MIL approaches use contrastive self-supervised learning (CSSL) to train the feature extractor [18, 35, 43]. In this section, we show that CSSL (e.g. SimCLR [10], MoCo [27]) has a crucial limitation in realistic MIL settings, which precludes it from learning discriminative features. CSSL aims to learn a representation space where samples from the same class are close to each other, and samples from different classes are far from each other, without access to class labels. This is achieved by minimizing the InfoNCE loss [40].

$$\mathcal{L}_{\text{CSSL}} = \mathop{\mathbb{E}}_{\substack{x, x^{\text{aug}} \\ \{x_i^{\text{diff}}\}_{i=1}^{n}}} \left[-\log \frac{sim(x, x^{\text{aug}})}{sim(x, x^{\text{aug}}) + \sum_{i=1}^{n} sim(x, x_i^{\text{diff}})} \right]. \tag{1}$$

The similarity score $sim\left(\cdot,\cdot\right) : \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}$ is defined as $sim\left(x, x'\right) = \exp\left(f_\psi(x) \cdot f_\psi(x')/\tau\right)$ for any $x, x' \in \mathbb{R}^m$, where $f_\psi = \psi \circ f$, in which $f : \mathbb{R}^m \to \mathbb{R}^d$ is the feature extractor mapping the input data to a representation, $\psi : \mathbb{R}^d \to \mathbb{R}^{d'}$ is a projection head with a feed-forward

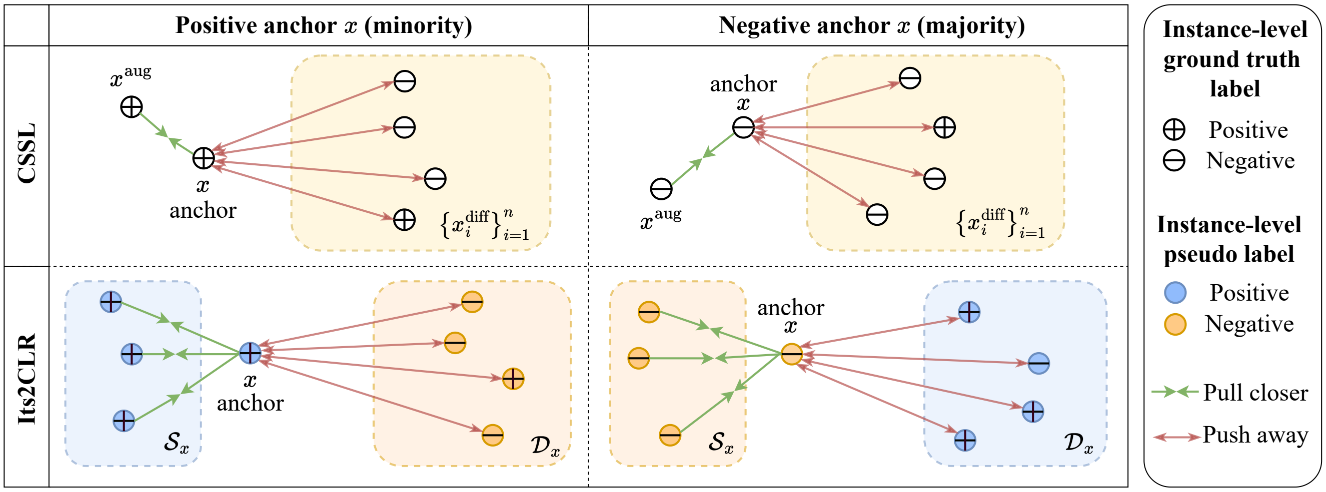

Figure 2. **Top**: In contrastive self-supervised learning (CSSL), the representation of an instance x is pulled closer to its random augmentation x^{aug} and pushed away from the representations of other randomly selected instances $\{x_i^{\text{diff}}\}_{i=1}^n$. In many MIL datasets relevant to medical diagnosis, most instances are negative, so CSSL mostly pushes apart representations of negative instances (right). **Bottom**: Our proposed framework ItS2CLR applies the supervised contrastive learning approach described in Section 3.1. Instance-level pseudo labels are used to build a set of positive pairs $\mathcal{S}_x$ and a set of negative pairs $\mathcal{D}_x$ corresponding to x. The representation of an instance x is pulled closer to those in $\mathcal{S}_x$ and pushed away from those in $\mathcal{D}_x$. The set of pseudo-labels is built iteratively following the self-paced sampling strategy in Section 3.2.

network and ℓ_2 normalization, and τ is a temperature hyperparameter. The expectation is taken over samples $x \in \mathbb{R}^m$ drawn uniformly from the training set. Minimizing the loss brings the representation of an instance x closer to the representation of its random augmentation, x^{aug}, and pushes the representation of x away from the representations of n other examples $\{x_i^{\text{diff}}\}_{i=1}^n$ in the training set.

A key assumption in CSSL is that x belongs to a different class than most of the randomly-sampled examples $x_1^{\text{diff}}, \ldots, x_n^{\text{diff}}$. This usually holds in standard classification datasets with many classes such as ImageNet [19], but *not in MIL tasks relevant to medical diagnosis*, where a majority of instances are negative (e.g. 95% in Camelyon16). Hence, most terms in the sum $\sum_{i=1}^n \exp\left(f_\psi(x) \cdot f_\psi(x_i^{\text{diff}})/\tau\right)$ in the loss in Equation 1 correspond to pairs of examples (x, x_i^{diff}) both belonging to the negative class. Therefore, minimizing the loss mostly pushes apart the representations of negative instances, as illustrated in the top panel of Figure 2. This is an example of *class collision* [2,13], a general problem in CSSL, which has been shown to impair performance on downstream tasks [3,56].

Class collision makes CSSL learn representations that are not discriminative between classes. In order to study this phenomenon, we report the average inter-class distances and intra-class deviations for representations learned by CSSL on Camelyon16 in Table 1. The inter-class distance reflects how far the instance representations from different classes are apart; the intra-class distance reflects the variation of instance representations within each class. As predicted, the intra-class deviation corresponding to the representations of negative instances learned by CSSL is large. Representations learned by ItS2CLR have larger inter-class distance (more separated classes) and smaller intra-class deviation (less variance among instances belonging to the same class) than those learned by CSSL. This suggests that the features learned by ItS2CLR are more discriminative, which is confirmed by the results in Section 4.

Note that using bag-level labels does not solve the problem of class collision. When x is negative, even if we select $\{x_i^{\text{diff}}\}_{i=1}^n$ from the positive bags in equation 1, most of the selected instances will still be negative. Overcoming the class-collision problem requires explicitly detecting positive instances. This motivates our proposed framework, described in the following section.

3. MIL via Iterative Self-paced Supervised Contrastive learning

Iterative **S**elf-paced **S**upervised **C**ontrastive **L**earning for MIL **R**epresentations (ItS2CLR) addresses the limitation of contrastive self-supervised learning (CSSL) described in Section 2. ItS2CLR relies on latent variables indicating whether each instance is positive or negative, which we call *instance-level pseudo labels*. To estimate pseudo labels, we use instance-level probabilities obtained from the MIL aggregator (specifically the aggregator from DS-MIL [35] but our framework is compatible with any aggre-

Table 1. Quantitative analysis of instance-level features learned from Camelyon16 [5]. The inter-class distance is the ℓ_2-distance between the representation mean of the positive instances and that of negative instances. The intra-class deviation is the square root of the spectral norm of the covariance matrix of features corresponding to each class. The spectral norm is the largest eigenvalue of the covariance matrix and is therefore equal to the variance in the direction of the greatest variance. Due to class collision among negative instances in CSSL (see Section 2), the intra-class deviation of the corresponding features is very large. In contrast, the features learned by the proposed framework ItS2CLR has smaller intra-class deviation among both negative and positive instances, and a larger inter-class distance.

	Training set			Test set		
	Inter-class distance	Intra-class deviation		Inter-class distance	Intra-class deviation	
		pos	*neg*		*pos*	*neg*
CSSL (SimCLR)	1.835	1.299	1.453	2.109	1.416	1.484
ItS2CLR (proposed)	**2.376**	**1.176**	**0.805**	**2.432**	**1.215**	**0.847**

gator that generates instance-level probabilities, such as the ones described in Appendix C). The pseudo labels are obtained by binarizing the probabilities according to a threshold $\eta \in (0, 1)$, which is a hyperparameter.

ItS2CLR uses the pseudo labels to finetune the feature extractor (initialized using CSSL). In the spirit of iterative self-training techniques [36, 49, 57], we alternate between refining the feature extractor, re-computing the pseudo labels, and training the aggregator, as described in Algorithm 1. A key challenge is that the pseudo labels are not completely accurate, especially at the beginning of the training process. To address the impact of incorrect pseudo labels, we apply a contrastive loss to finetune the feature extractor (see Section 3.1), where the contrastive pairs are selected according to a novel self-paced learning scheme (see Section 3.2). The right panel of Figure 1 shows that our approach iteratively improves the pseudo labels on the Camelyon16 dataset [5]. This finetuning only requires a modest increment in computational time (see Section 4.5).

3.1. Supervised contrastive learning with pseudo labels

To address the class collision problem described in Section 2, we leverage *supervised* contrastive learning [21, 32, 41] combined with the pseudo labels estimated by the aggregator. The goal is to learn discriminative representations by pulling together the representations corresponding to instances in the same class, and pushing apart those belong to instances of different classes. For each *anchor* instance x selected for contrastive learning, we collect a set $\mathcal{S}_x$ believed to have the *same label* as x, and a set $\mathcal{D}_x$ believed to have a *different label* to x. These sets are depicted in the bottom panel of Figure 2. The supervised contrastive loss corresponding to x is defined as:

$$\mathcal{L}_{\text{sup}}(x) = \frac{1}{|\mathcal{S}_x|} \sum_{x_s \in \mathcal{S}_x} -\log \frac{sim(x, x_s)}{\sum\limits_{x_s \in \mathcal{S}_x} sim(x, x_s) + \sum\limits_{x_d \in \mathcal{D}_x} sim(x, x_d)}. \tag{2}$$

In Section 3.2, we explain how to select x, $\mathcal{S}_x$ and $\mathcal{D}_x$ to ensure that the selected samples have high-quality pseudo labels.

A tempting alternative to supervised constrastive learning is to train the feature extractor on pseudo labels using standard cross-entropy loss. However, in Section 4.3 we show that this leads to substantially worse performance in the downstream MIL classification task due to memorization of incorrect pseudo labels.

Algorithm 1 Iterative Self-Paced Supervised Contrastive Learning (It2SCLR)

```
Require: Feature extractor $f$, projection head $\psi$;
Require: MIL aggregator $g_\phi$, where $\phi$ is an instance classifier;
Require: Bags $\{X_b\}_{b=1}^B$, bag-level labels $\{Y_b\}_{b=1}^B$;
$f^{(0)} \leftarrow f_{\text{SSL}}$  # Initialize $f$ with SSL-pretrained weights
for $t = 0$ to T do
    # Extract instance representation
    $h_k^b \leftarrow f^{(t)}(x_k^b), \forall x_k^b \in X_b, \forall b$
    # Group instance embedding into bags
    $H_b \leftarrow \{h_k^b\}_{k=1}^{K_b}, \forall b$
    # Train the MIL aggregator
    $g_\phi^{(t)} \leftarrow$ Train with $\{H_b\}_{b=1}^B$ and $\{Y_b\}_{b=1}^B$
    $\text{AUC}_{val}^{(t)} \leftarrow$ bag-level AUC on the validation set
    # If the bag prediction improves on the validation set
    if $\text{AUC}_{val}^{(t)} \geq \max_{t' \leq t} \left\{\text{AUC}_{val}^{(t')}\right\}$ then
        # Update instance pseudo labels
        $\hat{y}_k^b \leftarrow \mathbb{1}_{\{\phi^{(t)}(h_k^b) > \eta\}}, \forall x_k^b \in X_b, \forall b$
    end if
    # Optimize feature extractor via Eq.(2)
    $f_\psi^{(t+1)} \leftarrow \text{argmin}_{f_\psi} \mathcal{L}_{\text{sup}}(f_\psi^{(t)})$
end for
```

3.2. Sampling via self-paced learning

A key challenge in It2SCLR is to improve the accuracy of instance-level pseudo labels without ground-truth labels. This is achieved by finetuning the feature extractor on a carefully-selected subset of instances. We select the anchor instance x and the corresponding sets $\mathcal{S}_x$ and $\mathcal{D}_x$ (defined in Section 3.1 and Figure 2) building upon two key insights: (1) The negative bags only contain negative instances. (2) The probabilities used to build the pseudo labels are indicative of their quality; instances with higher predicted probabilities usually have more accurate pseudo labels [36, 37, 60].

Let $\mathcal{X}_{\text{neg}}^{-}$ denote all instances within the negative bags. By definition of MIL, we can safely assume that all instances in $\mathcal{X}_{\text{neg}}^{-}$ are negative. In contrast, positive bags contain both positive and negative instances. Let $\mathcal{X}_{\text{pos}}^{+}$ and $\mathcal{X}_{\text{pos}}^{-}$ denote the sets of instances in positive bags with positive and negative pseudo labels respectively. During an initial warm-up lasting $T_{\text{warm-up}}$ epochs, we sample anchor instances x only from $\mathcal{X}_{\text{neg}}^{-}$ to ensure that they are indeed all negative. For each such instance, $\mathcal{S}_x$ is built by sampling instances from $\mathcal{X}_{\text{neg}}^{-}$, and $\mathcal{D}_x$ is built by sampling from $\mathcal{X}_{\text{pos}}^{+}$.

After the warm-up phase, we start sampling anchor instances from $\mathcal{X}_{\text{pos}}^{+}$ and $\mathcal{X}_{\text{pos}}^{-}$. To ensure that these instances have accurate pseudo labels, we only consider the top-$r\%$ instances with the highest confidence in each of these sets (i.e. the highest probabilities in $\mathcal{X}_{\text{pos}}^{+}$ and lowest probabilities in $\mathcal{X}_{\text{pos}}^{-}$), which we call $\mathcal{X}_{\text{pos}}^{+}(r)$ and $\mathcal{X}_{\text{pos}}^{-}(r)$ respectively, as illustrated by Appendix Figure 5. The ratio of positive-to-negative anchors is a fixed hyperparameter p_+. For each anchor x, the same-label set $\mathcal{S}_x$ is sampled from $\mathcal{X}_{\text{pos}}^{+}(r)$ if x is positive and from $\mathcal{X}_{\text{neg}}^{-} \cup \mathcal{X}_{\text{pos}}^{-}(r)$ if x is negative. The different-label set $\mathcal{D}_x$ is sampled from $\mathcal{X}_{\text{neg}}^{-} \cup \mathcal{X}_{\text{pos}}^{-}(r)$ if x is positive, and from $\mathcal{X}_{\text{pos}}^{+}(r)$ if x is negative.

To exploit the improvement of the instance representations during training, we gradually increase r to include more instances from positive bags, which can be interpreted as a self-paced *easy-to-hard* learning scheme [30, 33, 59]. Let t and T denote the current epoch, and the total number of epochs respectively. For $T_{\text{warmup}} < t \leq T$. we set:

$$r := r_0 + \alpha_r \left(t - T_{\text{warm-up}}\right), \tag{3}$$

where $\alpha_r = (r_T - r_0)/(T - T_{\text{warm-up}})$, r_0 and r_T are hyperparameters. Details on tuning these hyperparameters are provided in Appendix *A.3*. As demonstrated in the right panel of Figure 1 (see also Appendix B.1), this scheme indeed results in an improvement of the pseudo labels (and hence of the underlying representations).

4. Experiments

We evaluate ItS2CLR on three MIL datasets described in Section 4.1. In Section 4.2 we show that ItS2CLR consistently outperforms approaches that use CSSL feature-extraction by a substantial margin on all three datasets for different choices of aggregators. In Section 4.3 we show that ItS2CLR outperforms alternative finetuning approaches based on cross-entropy loss minimization and end-to-end training across a wide range of settings where the prevalence of positive instances and bag size vary. In Section 4.4, we show that ItS2CLR is able to improve features obtained from a variety of pretraining schemes and network architectures.

4.1. Datasets

We evaluate the proposed framework on three cancer diagnosis tasks. When training our models, we select the model with the highest bag-level performance on the validation set and report the performance on a held-out test set. More information about the datasets, experimental setup, and implementation is provided in Appendix A.

Camelyon16 [5] is a popular benchmark for MIL [35, 44, 53] where the goal is to detect breast-cancer metastasis in lymph node sections. It consists of 400 whole-slide histopathology images. Each whole slide image (WSI) corresponds to a bag with a binary label indicating the presence of cancer. Each WSI is divided into an average of 625 tiles at 5x magnification, which correspond to individual instances. The dataset also contains pixel-wise annotations indicating the presence of cancer, which can be used to derive ground-truth instance-level labels.

TCGA-LUAD is a dataset from The Cancer Genome Atlas (TCGA) [1], a landmark cancer genomics program, where the associated task is to detect genetic mutations in cancer cells. We build models to detect four mutations - EGFR, KRAS, STK11, and TP53, which are important to determine treatment options for LUAD [16, 23]. The data contains 800 labeled tumorous frozen WSIs from lung adenocarcinoma (LUAD). Each WSI is divided into an average of 633 tiles at 10x magnification corresponding to unlabeled instances.

The **Breast Ultrasound Dataset** contains 28,914 B-mode breast ultrasound exams [45]. The associated task is to detect breast cancer. Each exam contains between 4-70 images (18.8 images per exam on average) corresponding to individual instances, but only a bag-level label indicating the presence of cancer is available per exam. Additionally, a subset of images is annotated, which makes it possible to also evaluate instance-level performance. This dataset is imbalanced at the bag level: only 5,593 of 28,914 exams contain cancer.

4.2. Comparison with contrastive self-supervised learning

In this section, we compare the performance of ItS2CLR to a baseline that performs feature-extraction via the CSSL method SimCLR [10]. This approach has achieved state-of-the-art performance on multiple WSI datasets [35]. To ensure a fair comparison, we initialize the feature extractor in ItS2CLR also using SimCLR. Table 2 shows that ItS2CLR clearly outperforms the SimCLR baseline on all three datasets. The performance improvement is particularly significant in Camelyon16 where it achieves a bag-level AUC of 0.943, outperforming the baseline by an absolute margin of 8.87%. ItS2CLR also outperforms an improved baseline reported by Li *et al.* [35] with an AUC of 0.917, which uses higher resolution tiles than in our experi-

Table 2. Bag-level AUC of ItS2CLR and a two-stage baseline using a SimCLR feature extractor and a MIL aggregator. ItS2CLR outperforms the baseline on all three datasets.

AUC ($\times 10^{-2}$)	Camelyon16	Breast Ultrasound	TCGA-LUAD mutation			
			EGFR	KRAS	STK11	TP53
SimCLR + Aggregator	85.38	80.79	67.51	68.79	70.40	62.15
ItS2CLR	**94.25**	**93.93**	**72.30**	**71.06**	**75.08**	**65.61**

Table 3. Bag-level AUC on Camelyon16 for ItS2CLR and different baselines for five aggregators. We retrain each aggregator 5 times to report the mean and standard deviation (reported as a subscript). All feature extractors are initialized using SimCLR. Ground-truth and cross-entropy (CE) finetuning use ground-truth instance-level labels and pseudo labels to optimize the feature extractor respectively. We also include versions of ItS2CLR without iterative updates (w/o iter.), self-paced learning (w/o SPL) and both (w/o both), and a version of CE finetuning without iterative updates (w/o iter). See Appendix A.5 for a detailed description of the ablated models.

AUC ($\times 10^{-2}$)	SimCLR (CSSL)	Ground-truth finetuning*	CE finetuning		ItS2CLR			
			w/o iter.	*iter.*	*w/o both*	*w/o iter.*	*w/o SPL*	*Full*
Max pooling	$86.69_{1.09}$	$98.25_{0.01}$	$85.48_{0.24}$	$88.05_{0.77}$	$85.38_{0.31}$	$91.96_{0.31}$	$90.85_{0.76}$	$\mathbf{94.69}_{0.07}$
Top-k pooling [46]	$85.39_{1.20}$	$98.39_{0.05}$	$85.96_{0.45}$	$87.26_{0.42}$	$85.46_{0.21}$	$91.73_{0.42}$	$91.69_{0.28}$	$\mathbf{95.07}_{0.09}$
Attention-MIL [29]	$79.49_{3.20}$	$99.06_{0.02}$	$88.50_{0.54}$	$90.46_{0.64}$	$85.21_{0.74}$	$93.13_{0.22}$	$86.20_{3.25}$	$\mathbf{94.45}_{0.05}$
DS-MIL [35]	$85.38_{1.32}$	$98.65_{0.08}$	$87.01_{0.82}$	$90.38_{0.67}$	$85.08_{0.38}$	$91.69_{0.54}$	$88.29_{0.99}$	$\mathbf{94.25}_{0.07}$
Transformer [9]	$87.25_{0.59}$	$98.85_{0.25}$	$89.02_{0.54}$	$92.13_{1.07}$	$87.13_{0.71}$	$93.52_{0.49}$	$92.12_{0.68}$	$\mathbf{95.74}_{0.27}$

ments (both 20x and 5x, as opposed to only 5x).

To perform a more exhaustive comparison of the features learned by SimCLR and ItS2CLR, we compare them in combination with several different popular MIL aggregators.[2]: max pooling, top-k pooling [46], attention-MIL pooling [29], DS-MIL pooling [35], and transformer [9] (see Appendix C for a detailed description). Table 3 shows that the ItS2CLR features outperform the SimCLR features by a large margin for all aggregators, and are substantially more stable (the standard deviation of the AUC over multiple trials is lower).

We also evaluate instance-level accuracy, which can be used to interpret the bag-level prediction (for instance, by revealing tumor locations). In Table 4, we report the instance-level AUC, F1 score, and Dice score of both ItS2CLR and the SimCLR-based baseline on Camelyon16. ItS2CLR again exhibits stronger performance. Figure 3 shows an example of instance-level predictions in the form of a tumor localization map.

4.3. Comparison with alternative approaches

In Tables 3, 4 and 5, we compare ItS2CLR with the approaches described below. Table 6 reports additional comparisons at different *witness rates* (the fraction of positive instances in positive bags), created synthetically by modifying the ratio between negative and positive instances in Camelyon16.

Finetuning with ground-truth instance labels provides an upper bound on the performance that can be achieved through feature improvement. ItS2CLR does not reach this gold standard, but substantially closes the gap.

Cross-entropy finetuning with pseudo labels, which we refer to as *CE finetuning*, consistently underperforms ItS2CLR when combined with different aggregators, except at high witness rates. We conjecture that this is due to the sensitivity of the cross-entropy loss to incorrect pseudo labels. We experiment with two settings: CE finetuning *without* and *with* iterative updating. In CE finetuning *without* iterative updating, we use the same initial pseudo labels and pretrained representations as our ItS2CLR framework. Concretely, we label all the instances in negative bags as negative, and the instances in positive bags using the instance prediction obtained from the initially trained MIL aggregator. When finetuning the feature extractor, pseudo labels are kept fixed. In CE finetuning *with* iterative updating, the pseudo labels are updated every few epochs.

Ablated versions of ItS2CLR where we do not apply iterative updates of the pseudo labels (w/o iter.), or our self-paced learning scheme (w/o SPL) or both (w/o both) achieve substantially worse performance than the full approach. This indicates that both of these ingredients are critical in learning discriminative features.

End-to-end training is often computationally infeasible in medical applications. We compare ItS2CLR to end-to-end models on a downsampled version of Camelyon16 (see Appendix A.4) and on the breast ultrasound dataset. For a fair comparison, all end-to-end models use the same CSSL-

---

[2] To be clear, the ItS2CLR features are learned using the DS-MIL aggregator, as described in Section 3, and then frozen before combining them with the different aggregators.

Table 4. Comparison of instance-level performance between the models in Table 3. All models use the DS-MIL aggregator. ItS2CLR achieves the best localization performance. Dice score is computed from a post-processed probability estimate described in Appendix B.2, which also includes further details and results for other aggregators.

($\times 10^{-2}$)	SimCLR	Ground-truth	CE finetuning		ItS2CLR			
	(CSSL)	finetuning	*w/o iter.*	*iter.*	*w/o both*	*w/o iter.*	*w/o SPL*	*Full*
AUC	94.01	97.94	95.69	96.06	95.13	95.90	96.12	**96.72**
F1-score	84.49	88.01	86.94	86.93	86.74	87.45	**87.95**	87.47
AUPRC	86.57	86.13	89.26	89.39	88.30	89.51	90.00	**91.12**
Dice (*)	31.79	62.17	49.11	49.41	43.74	51.70	53.03	**57.86**
IoU	39.53	50.24	44.98	44.88	41.37	44.56	45.41	**48.27**

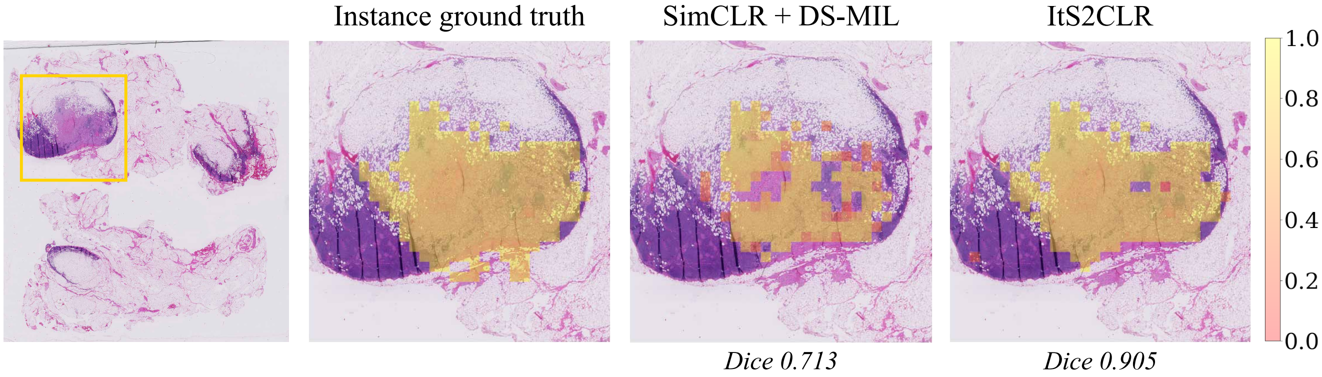

Figure 3. Tumor localization in a histopathology slide from the Camelyon16 test set. Instance-level predictions are generated by the instance-level classifier of the DS-MIL aggregator based on different instance representations. Yellow indicates higher probability of being cancerous. Transparent tiles are with probabilities less than 0.2. Appendix B.4 shows additional examples.

pretrained weights and aggregator as used in ItS2CLR. Table 5 shows that ItS2CLR achieves better instance- and bag-level performance than end-to-end training. In Appendix B.3 we show that end-to-end models suffer from overfitting.

4.4. Improving different pretrained representations

In this section, we show that ItS2CLR is capable of improving representations learned by different pretraining methods: supervised training on ImageNet and two non-contrastive SSL methods, BYOL [25] and DINO [8]. DINO is by default based on the ViT-S/16 architecture [20], whereas the other methods are based on ResNet-18.

Table 7a shows the result of initializing ItS2CLR with pretrained weights obtained from these different models (as well as from SimCLR). The non-contrastive SSL methods fail to learn better representations than SimCLR. Non-contrastive SSL methods do not use the negative samples, Wang *et al*. [48] report that this can make the representations under-clustering and result in different object categories overlapping in the representation space. The significant improvement of ItS2CLR on top of different pretraining methods demonstrates that the proposed framework is more effective in learning discriminative representations than altering SSL pretraining methods.

As shown in Table 7b, different initializations achieve varying degrees of pseudo label accuracy, but ItS2CLR improves the performance of all of them. This further demonstrates the robustness of the proposed framework.

4.5. Computational complexity

ItS2CLR only requires a small increment in computational time, with respect to existing approaches. For Camelyon16, it takes 600 epochs (approximately 90 hours) to train SimCLR. It only takes 50 extra epochs (approximately 10 hours) to finetune with ItS2CLR, which is only 1/10 of the pretraining time. Updating the pseudo labels is also efficient: it only takes around 10 minutes to update the instance features and training the MIL aggregator. These updates occur every 5 epochs. More training details are provided in Appendix A.2.

5. Related work

Self-supervised learning Contrastive learning methods have become popular in unsupervised representation learning, achieving state-of-the-art self-supervised learning performance for natural images [7, 8, 10, 25, 27, 51]. These methods have also shown promising results in medical

Table 5. Comparison to models trained end-to-end, initialized with the same pretrained weights as ItS2CLR, and use the same aggregator. ItS2CLR achieves better instance- and bag-level performance.

	Camelyon16 (downsampled synthetic)			Breast Ultrasound			
	Bag AUC	Instance AUC	Instance F-score	Bag AUC	Bag AUPRC	Instance AUC	Instance AUPRC
End-to-end	66.71	78.32	55.71	91.26	58.73	82.11	31.31
ItS2CLR	**88.65**	**95.58**	**87.01**	**93.93**	**70.30**	**88.63**	**43.71**

Table 6. Bag-level AUC on Camelyon16 across different witness rates (WR): the fraction of positive instances in positive bags. All methods use DS-MIL aggregator for a fair comparison. When the WR is low, ItS2CLR outperforms CE finetuning by a large margin. As the WR increases, CE finetuning becomes more effective.

Down. Instances	WR (%)	SimCLR (CSSL)	CE finetuning iterative	ItS2CLR	Finetuning w. instance GT
5% Neg.	71.2	94.52	**98.55**	97.58	99.11
10% Neg.	45.0	93.70	**97.88**	96.15	99.18
40% Neg.	23.5	90.38	93.32	**95.40**	97.68
Original	10.9	85.38	90.38	**94.25**	98.65
50% Pos.	5.8	82.47	86.96	**88.52**	91.81
33% Pos.	4.1	78.21	80.56	**86.02**	88.01

Table 7. The effects of finetuning from different initial weights on Camelyon16. **(a)** The bag-level prediction performance on **test** set; **(b)** the evaluation on instance-level prediction in **training** set.

Bag-level test AUC	ImageNet	BYOL	SimCLR	DINO (ViT)
Pretrained	0.712	0.704	0.854	0.857
ItS2CLR	**0.791**	**0.764**	**0.943**	**0.936**

(a)

Training ins. pred	ImageNet		BYOL		SimCLR		DINO (ViT)	
	AUC	*F-score*	*AUC*	*F-score*	*AUC*	*F-score*	*AUC*	*F-score*
Initial	0.772	0.314	0.729	0.225	0.934	0.746	0.906	0.716
Finetuned	**0.835**	**0.598**	**0.804**	**0.288**	**0.973**	**0.783**	**0.972**	**0.806**

(b)

imaging [4, 14, 31, 35, 58]. Recently, Li *et al.* [35] applied SimCLR [10] to extract instance-level features for WSI MIL tasks and achieved state-of-the-art performance. However, Arora *et al.* [2] point out the potential issue of class collision in contrastive learning, i.e. that some negative pairs may actually have the same class. Prior works on alleviating class collision problem include reweighting the negative and positive terms with class ratio [13], pulling closer additional similar pairs [21], and avoiding pushing apart negatives that are inferred to belong to the same class based on a similarity metric [56]. In contrast, we propose a framework that leverages information from the bag-level labels to iteratively resolve the class collision problem.

Multiple instance learning A major part of MIL works focuses on improving the MIL aggregator. Traditionally, non-learnable pooling operators such as mean-pooling and max-pooling were commonly used in MIL [22, 42]. More recent methods parameterize the aggregator using neural networks that employ attention mechanisms [9, 29, 35, 44]. This research direction is complementary to our proposed approach, which focuses on obtaining better instance representations, and can be combined with different types of aggregators (see Section 4.2).

Self-paced Learning The core idea of self-paced learning is the "easy-to-hard" training scheme, which has been used in semi-supervised learning [52], learning with noisy label, unsupervised clustering [26], domain adaptation [12, 24, 54, 55]. In this work, we apply self-paced learning to instance representation learning in MIL tasks.

6. Conclusion

In this work, we investigate how to improve feature extraction in multiple-instance learning models. We identify a limitation of contrastive self-supervised learning: class collision hinders it from learning discriminative features in class-imbalanced MIL problems. To address this, we propose a novel framework that iteratively refines the features with pseudo labels estimated by the aggregator. Our method outperforms the existing state-of-the-art MIL methods on three medical datasets, and can be combined with different aggregators and pretrained feature extractors.

The proposed method does not outperform a cross-entropy-based baseline at very high witness rates, suggesting that it is mostly suitable for low witness rates scenarios (however, it is worth noting that this is the regime more commonly encountered in medical applications such as cancer diagnosis). In addition, there is a performance gap between our method and finetuning using instance-level ground truth, suggesting further room for improvement.

Acknowledgments The authors gratefully acknowledge NSF grants OAC-2103936 (K.L.), DMS-2009752 (C.F.G., S.L.), NRT-HDR-1922658 (K.L., S.L., Y.S., W.Z.), the NYU Langone Health Predictive Analytics Unit (N.R., W.Z.), the Alzheimer's Association grant AARG-NTF-21-848627 (S.L.), the NIH grant P41EB017183, and the Gordon and Betty Moore Foundation (9683) (K.G., Y.S.).

References

[1] The cancer genome atlas program. `https://www.cancer.gov/tcga`, 2019. 5

[2] Sanjeev Arora, Hrishikesh Khandeparkar, Mikhail Khodak, Orestis Plevrakis, and Nikunj Saunshi. A theoretical analysis of contrastive unsupervised representation learning. *arXiv preprint arXiv:1902.09229*, 2019. 3, 8

[3] Jordan T Ash, Surbhi Goel, Akshay Krishnamurthy, and Dipendra Misra. Investigating the role of negatives in contrastive representation learning. *arXiv preprint arXiv:2106.09943*, 2021. 3

[4] Shekoofeh Azizi, Basil Mustafa, Fiona Ryan, Zachary Beaver, Jan Freyberg, Jonathan Deaton, Aaron Loh, Alan Karthikesalingam, Simon Kornblith, Ting Chen, et al. Big self-supervised models advance medical image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 8

[5] Babak Ehteshami Bejnordi, Mitko Veta, Paul Johannes Van Diest, Bram Van Ginneken, Nico Karssemeijer, Geert Litjens, Jeroen AWM Van Der Laak, Meyke Hermsen, Quirine F Manson, Maschenka Balkenhol, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *Jama*, 318(22):2199–2210, 2017. 1, 2, 4, 5

[6] Gabriele Campanella, Matthew G Hanna, Luke Geneslaw, Allen Miraflor, Vitor Werneck Krauss Silva, Klaus J Busam, Edi Brogi, Victor E Reuter, David S Klimstra, and Thomas J Fuchs. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature medicine*, 25(8):1301–1309, 2019. 1

[7] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *arXiv preprint arXiv:2006.09882*, 2020. 7

[8] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. *CoRR*, abs/2104.14294, 2021. 7

[9] Richard J. Chen and Rahul G. Krishnan. Self-supervised vision transformers learn visual concepts in histopathology, 2022. 1, 6, 8, 21

[10] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 2, 5, 7, 8, 12, 13, 18

[11] Philip Chikontwe, Meejeong Kim, Soo Jeong Nam, Heounjeong Go, and Sang Hyun Park. Multiple instance learning with center embeddings for histopathology classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 519–528. Springer, 2020. 1

[12] Jaehoon Choi, Minki Jeong, Taekyung Kim, and Changick Kim. Pseudo-labeling curriculum for unsupervised domain adaptation. *arXiv preprint arXiv:1908.00262*, 2019. 8

[13] Ching-Yao Chuang, Joshua Robinson, Yen-Chen Lin, Antonio Torralba, and Stefanie Jegelka. Debiased contrastive learning. *Advances in neural information processing systems*, 33:8765–8775, 2020. 3, 8

[14] Ozan Ciga, Tony Xu, and Anne Louise Martel. Self supervised contrastive learning for digital histopathology. *Machine Learning with Applications*, 7:100198, 2022. 1, 8

[15] Alex Clark. Pillow (pil fork) documentation, 2015. 12, 13

[16] Nicolas Coudray, Paolo Santiago Ocampo, Theodore Sakellaropoulos, Navneet Narula, Matija Snuderl, David Fenyö, Andre L. Moreira, Narges Razavian, and Aristotelis Tsirigos. Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. *Nature Medicine*, 24(10):1559–1567, 2018. 5, 12

[17] Pierre Courtiol, Eric W Tramel, Marc Sanselme, and Gilles Wainrib. Classification and disease localization in histopathology using only global labels: A weakly-supervised approach. *arXiv preprint arXiv:1802.02212*, 2018. 1

[18] Olivier Dehaene, Axel Camara, Olivier Moindrot, Axel de Lavergne, and Pierre Courtiol. Self-supervision closes the gap between weak and strong supervision in histology. *arXiv preprint arXiv:2012.03583*, 2020. 2

[19] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 3

[20] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020. 7

[21] Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. With a little help from my friends: Nearest-neighbor contrastive learning of visual representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9588–9597, 2021. 4, 8

[22] Ji Feng and Zhi-Hua Zhou. Deep miml network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017. 8

[23] Yu Fu, Alexander W. Jung, Ramon Viñas Torne, Santiago Gonzalez, Harald Vöhringer, Artem Shmatko, Lucy R. Yates, Mercedes Jimenez-Linan, Luiza Moore, and Moritz Gerstung. Pan-cancer computational histopathology reveals mutations, tumor composition and prognosis. *Nature Cancer*, 1(8):800–810, 2020. 5, 12

[24] Yixiao Ge, Feng Zhu, Dapeng Chen, Rui Zhao, et al. Self-paced contrastive learning with hybrid memory for domain adaptive object re-id. *Advances in Neural Information Processing Systems*, 33:11309–11321, 2020. 8

[25] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020. 7

[26] Xifeng Guo, Xinwang Liu, En Zhu, Xinzhong Zhu, Miaomiao Li, Xin Xu, and Jianping Yin. Adaptive self-paced deep clustering with data augmentation. *IEEE Transactions on Knowledge and Data Engineering*, 32(9):1680–1693, 2019. 8

[27] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020. 2, 7

[28] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 12, 13

[29] Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *International conference on machine learning*, pages 2127–2136. PMLR, 2018. 6, 8, 13, 21

[30] Lu Jiang, Deyu Meng, Shoou-I Yu, Zhenzhong Lan, Shiguang Shan, and Alexander Hauptmann. Self-paced learning with diversity. *Advances in neural information processing systems*, 27, 2014. 5

[31] Aakash Kaku, Sahana Upadhya, and Narges Razavian. Intermediate layers matter in momentum contrastive self supervised learning. In *Advances in Neural Information Processing Systems*, 2021. 8

[32] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33:18661–18673, 2020. 2, 4

[33] M Kumar, Benjamin Packer, and Daphne Koller. Self-paced learning for latent variable models. *Advances in neural information processing systems*, 23, 2010. 5

[34] Marvin Lerousseau, Maria Vakalopoulou, Marion Classe, Julien Adam, Enzo Battistella, Alexandre Carré, Théo Estienne, Théophraste Henry, Eric Deutsch, and Nikos Paragios. Weakly supervised multiple instance learning histopathological tumor segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 470–479. Springer, 2020. 1

[35] Bin Li, Yin Li, and Kevin W Eliceiri. Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14318–14328, 2021. 1, 2, 3, 5, 6, 8, 12, 13, 21

[36] Sheng Liu, Kangning Liu, Weicheng Zhu, Yiqiu Shen, and Carlos Fernandez-Granda. Adaptive early-learning correction for segmentation from noisy annotations. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 4, 5

[37] Sheng Liu, Jonathan Niles-Weed, Narges Razavian, and Carlos Fernandez-Granda. Early-learning regularization prevents memorization of noisy labels. *Advances in Neural Information Processing Systems*, 2020. 5

[38] Ming Y Lu, Richard J Chen, and Faisal Mahmood. Semi-supervised breast cancer histology classification using deep multiple instance learning and contrast predictive coding (conference presentation). In *Medical Imaging 2020: Digital Pathology*, volume 11320, page 113200J. International Society for Optics and Photonics, 2020. 1

[39] Ming Y Lu, Drew FK Williamson, Tiffany Y Chen, Richard J Chen, Matteo Barbieri, and Faisal Mahmood. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature biomedical engineering*, 5(6):555–570, 2021. 1

[40] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 2

[41] Omiros Pantazis, Gabriel J Brostow, Kate E Jones, and Oisin Mac Aodha. Focus on the positives: Self-supervised learning for biodiversity monitoring. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10583–10592, 2021. 4

[42] Pedro O Pinheiro and Ronan Collobert. From image-level to pixel-level labeling with convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1713–1721, 2015. 8

[43] Dawid Rymarczyk, Aneta Kaczy'nska, Jaroslaw Kraus, Adam Pardyl, and Bartosz Zieli'nski. Protomil: Multiple instance learning with prototypical parts for fine-grained interpretability. *ArXiv*, abs/2108.10612, 2021. 2

[44] Zhuchen Shao, Hao Bian, Yang Chen, Yifeng Wang, Jian Zhang, Xiangyang Ji, et al. Transmil: Transformer based correlated multiple instance learning for whole slide image classification. *Advances in Neural Information Processing Systems*, 34, 2021. 1, 5, 8

[45] Yiqiu Shen, Farah E Shamout, Jamie R Oliver, Jan Witowski, Kawshik Kannan, Jungkyu Park, Nan Wu, Connor Huddleston, Stacey Wolfson, Alexandra Millet, et al. Artificial intelligence system reduces false-positive findings in the interpretation of breast ultrasound exams. *Nature communications*, 12(1):1–13, 2021. 5, 12, 13

[46] Yiqiu Shen, Nan Wu, Jason Phang, Jungkyu Park, Kangning Liu, Sudarshini Tyagi, Laura Heacock, S Gene Kim, Linda Moy, Kyunghyun Cho, et al. An interpretable classifier for high-resolution breast cancer screening images utilizing weakly supervised localization. *Medical image analysis*, 68:101908, 2021. 6, 21

[47] Abhishek Vahadane, Tingying Peng, Amit Sethi, Shadi Albarqouni, Lichao Wang, Maximilian Baust, Katja Steiger, Anna Melissa Schlitter, Irene Esposito, and Nassir Navab. Structure-preserving color normalization and sparse stain separation for histological images. *IEEE Transactions on Medical Imaging*, 35(8):1962–1971, 2016. 13

[48] Guangrun Wang, Keze Wang, Guangcong Wang, Philip HS Torr, and Liang Lin. Solving inefficiency of self-supervised representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9505–9515, 2021. 7

[49] Colin Wei, Kendrick Shen, Yining Chen, and Tengyu Ma. Theoretical analysis of self-training with deep networks on unlabeled data. In *International Conference on Learning Representations*, 2020. 4

[50] Gang Xu, Zhigang Song, Zhuo Sun, Calvin Ku, Zhe Yang, Cancheng Liu, Shuhao Wang, Jianpeng Ma, and Wei Xu. Camel: A weakly supervised learning framework for histopathology image segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10682–10691, 2019. 1

[51] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. *arXiv preprint arXiv:2103.03230*, 2021. 7

[52] 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, 2021. 8

[53] Hongrun Zhang, Yanda Meng, Yitian Zhao, Yihong Qiao, Xiaoyun Yang, Sarah E Coupland, and Yalin Zheng. Dtfd-mil: Double-tier feature distillation multiple instance learning for histopathology whole slide image classification. *arXiv preprint arXiv:2203.12081*, 2022. 1, 5

[54] Yang Zhang, Philip David, and Boqing Gong. Curriculum domain adaptation for semantic segmentation of urban scenes. In *Proceedings of the IEEE international conference on computer vision*, pages 2020–2030, 2017. 8

[55] Fang Zhao, Shengcai Liao, Guo-Sen Xie, Jian Zhao, Kaihao Zhang, and Ling Shao. Unsupervised domain adaptation with noise resistible mutual-training for person re-identification. In *European Conference on Computer Vision*, pages 526–544. Springer, 2020. 8

[56] Mingkai Zheng, Fei Wang, Shan You, Chen Qian, Changshui Zhang, Xiaogang Wang, and Chang Xu. Weakly supervised contrastive learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10042–10051, 2021. 3, 8

[57] Jia-Xing Zhong, Nannan Li, Weijie Kong, Shan Liu, Thomas H Li, and Ge Li. Graph convolutional label noise cleaner: Train a plug-and-play action classifier for anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1237–1246, 2019. 4

[58] Weicheng Zhu, Carlos Fernandez-Granda, and Narges Razavian. Interpretable prediction of lung squamous cell carcinoma recurrence with self-supervised learning. In *Proceedings of The 5th International Conference on Medical Imaging with Deep Learning*, volume 172 of *Proceedings of Machine Learning Research*, pages 1504–1522. PMLR, 06–08 Jul 2022. 8

[59] Yang Zou, Zhiding Yu, BVK Kumar, and Jinsong Wang. Domain adaptation for semantic segmentation via class-balanced self-training. *arXiv preprint arXiv:1810.07911*, 2018. 5

[60] Yang Zou, Zhiding Yu, BVK Kumar, and Jinsong Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *Proceedings of the European conference on computer vision*, pages 289–305, 2018. 5