Improving Self-supervised Learning with Hardness-aware Dynamic Curriculum Learning: An Application to Digital Pathology

Chetan L Srinidhi, Anne L Martel
Physical Sciences, Sunnybrook Research Institute, Toronto, Canada
Department of Medical Biophysics, University of Toronto, Canada
{chetan.srinidhi, a.martel}@utoronto.ca

Abstract

*Self-supervised learning (SSL) has recently shown tremendous potential to learn generic visual representations useful for many image analysis tasks. Despite their notable success, the existing SSL methods fail to generalize to downstream tasks when the number of labeled training instances is small or if the domain shift between the transfer domains is significant. In this paper, we attempt to improve self-supervised pretrained representations through the lens of curriculum learning by proposing a hardness-aware dynamic curriculum learning (**HaDCL**) approach. To improve the robustness and generalizability of SSL, we dynamically leverage progressive harder examples via **easy-to-hard** and **hard-to-very-hard** samples during mini-batch downstream fine-tuning. We discover that by progressive stage-wise curriculum learning, the pretrained representations are significantly enhanced and adaptable to both **in-domain** and **out-of-domain** distribution data.*

*We performed extensive validation on **three** histology benchmark datasets on both **patch-wise** and **slide-level** classification problems. Our curriculum based fine-tuning yields a significant improvement over standard fine-tuning, with a minimum improvement in area-under-the-curve (AUC) score of 1.7% and 2.2% on in-domain and out-of-domain distribution data, respectively. Further, we empirically show that our approach is more generic and adaptable to any SSL methods and does not impose any additional overhead complexity. Besides, we also outline the role of patch-based versus slide-based curriculum learning in histopathology to provide practical insights into the success of curriculum based fine-tuning of SSL methods.*[1]

1. Introduction

Learning with limited human supervision is a longstanding goal in machine learning, especially in medical image analysis due to the expensive and time-consuming annotation process. Self-supervised learning (SSL) methods have gained increasing popularity due to their ability to learn general-purpose features that are competitive with representations generated by state-of-the-art (SoTA) fully-supervised methods [3, 9, 12, 31]. These methods involve two steps: unsupervised pretraining on unlabeled data in a task-agnostic way, followed by supervised fine-tuning in a task-specific way with limited labeled data. SSL methods, however, often struggle to perform well on downstream tasks and generalize poorly on out-of-distribution data due to limited downstream supervision [1, 40, 41].

Recent studies have focused on improving self-supervised pretrained representations with effective sampling strategies that mine informative hard examples via aggressive data augmentations [24] or with hard-negative mining techniques [28]. However, these methods are tailored to improve a specific family of contrastive based SSL methods (such as MoCo [20], and SimCLR [9]) and cannot be applied or generalized to other pretraining methods. In contrast, consistency based semi-supervised techniques [10, 31, 41] have been proposed to improve the SSL by utilizing the unlabeled data in a task-specific semi-supervised manner. However, despite their improved performance, the semi-supervised approaches typically suffer from the problem of confirmation bias [2], and as yet, their practical applicability to medical image analysis has been severely limited.

In this paper, we attempt to improve self-supervised pretrained representations through the lens of curriculum learning (CL). CL in machine learning paradigm [6, 29] is fundamentally inspired by the human learning process, where the easier concepts (examples) are presented first, and most difficult concepts are learned later on. Such meaningful ordering of samples (as opposed to random ordering) during training has shown to improve both convergence speed and accuracy of the neural network model. To this end, we extend the previous idea of leveraging hard examples to improve the self-supervised pretrained representa-

[1] Code will be released at https://github.com/srinidhiPY/ICCV-CDPATH2021-ID-8

tions [24, 28] further by combining SSL and CL in an elegant manner. In this study, we empirically investigate their inter-dependencies and present novel ways to combine them to achieve faster convergence, better generalization ability, and alleviate over-fitting to in-domain data. In relation to the CL paradigm introduced in [6, 19, 38], we first start by asking two very fundamental questions: (i) **how to determine the notion of example difficulty** (i.e., *scoring* function) that is made available to the network during training? (ii) **how to specify the order** (typically, easy to hard) **at which the examples are presented to the network?** - which depends on both the data and learning model.

To answer the above questions, we first attempt to determine the "***hardness***" or "***difficulty***" of each sample in the training data via curriculum by transfer learning approach, initially proposed in Weinshall *et al.* [37]. Here, we choose to rank the difficulty of training samples with the help of instantaneous feedback (i.e., *loss*) from the pretrained self-supervised model, while fine-tuning on the downstream task of interest. Unlike in the previous study in histopathology [36], where the ranking (i.e., hardness) of samples are determined with the aid of human teachers (i.e., pathologists), our proposed approach rather investigates the knowledge transfer to determine the hardness of each training sample, which provides more reliable scores for the target task and does not involve any additional human-intervention. This is particularly important in pathology, where obtaining multiple annotator agreements as a proxy for determining the sample difficulty is often time-consuming and challenging. Besides, it is also shown in previous studies [19, 37] that the ranking provided by the human teachers may not reflect the true underlying difficulty as it affects the neural network.

Second, we focus our attention on specifying the *"**order**"* at which the data is presented to the network. Typically, most studies in the CL literature [29, 38] either follow ordering of input examples from easy-to-hard (curriculum) or hard-to-easy (anti-curriculum) and sometimes random [38]. However, one significant limitation with the existing approaches is that they do not necessarily consider the learning dynamics of the neural network while estimating the sample hardness over the course of model training. Due to the stochastic mini-batch style nature of gradient-descent optimization, the instantaneous hardness of each training sample changes over time from the early part of training to the later part of training. i.e., the hardness of each sample decreases monotonically over the course of training, where the hard samples become easier, while easy samples stay easy throughout training. With this motivation, we choose to measure the sample hardness adaptively in each mini-batch during task-specific fine-tuning of self-supervised pretrained model by introducing "hardness-aware dynamic curriculum learning (HaDCL)" as a mini-batch instantaneous hardness measure of a training sample over time. Empirically, we show that our proposed HaDCL strategy significantly improves the SSL on a challenging downstream task, i.e., breast cancer lymph node metastases detection on both in-domain and out-of-domain data, supporting that the hard example mining is indeed crucial for improved model accuracy and generalizability.

Contributions. To summarise, we make the following contributions in this study:

- We propose a principled way of combining SSL with CL to improve the self-supervised pretrained representations on the downstream task on both in-domain and out-of-domain distribution data.
- We present a mini-batch hardness-aware dynamic curriculum learning (HaDCL) strategy to determine the instantaneous hardness of training samples with improved training convergence and better accuracy.
- We also conduct an empirical study to understand the boundaries within which the curriculum works to improve SSL on both patch-wise and slide-level classification tasks in histology. Further, we also probe the generalizability of our method on out-of-distribution data with significant domain shifts.

2. Related Work

Self-supervised Learning. Inspired by the recent success in SSL, the existing methods are categorized into context-based and contrastive-based learning methods. The early works focused on context-based methods to formulate an auxiliary task (i.e., *pretext* task) to pretrain the model on the unlabeled data [22]. These pretext tasks were handcrafted based on domain knowledge, which includes rotation [18], solving jigsaw puzzles [25], relative patch prediction [17], and so on. Many of these tasks are based on ad-hoc heuristics that limit the applicability of these approaches to broader domains. Consequently, a new family of SSL methods based on contrastive learning [9, 20] has emerged as the top-performing method that demonstrated excellent performance on many downstream tasks. More recently, these techniques have been extended to medical image analysis [3, 4, 12, 23, 30, 31] and have shown a promising viable alternative to fully supervised based methods.

In the context of histopathology, a few domain-specific pretext tasks [23, 31] have been proposed to leverage multi-resolution contextual features for learning representations in pathology images. Notably, the recently proposed resolution sequence prediction (RSP) [31] pretext task has shown promising results on three different histopathology tasks, including patch-wise and slide-level classification problems. Contrastive learning based methods such as SimCLR have also been extended to histology [12] and have shown SOTA performance on many diverse histology tasks. However, in recent studies [31, 40, 41], it is shown that the

representations learned by SSL methods are often overfitted to the pretraining objective and do not generalize well to downstream tasks. Furthermore, the improved efficiency of these methods is heavily dependent on the quantity of both labeled and unlabeled data [14, 27], and most importantly, the aggressive data augmentation strategies [24, 26, 39] that are used during pretraining.

Consequently, some recent works have attempted to improve the pretrained representations by leveraging hard examples either during the pretraining stage [24, 28] or during the fine-tuning stage [1, 11]. Inspired by these previous works, we propose to improve SSL on downstream tasks with a curriculum based hard example fine-tuning. We will show that our proposed technique improves robustness on both in-domain and out-of-domain distribution data, and furthermore, our method is generic and easily adaptable to any self-supervised pretrained objective.

Curriculum Learning. The human learning mechanism follows a curriculum to understand complex tasks by imposing the order at which the complexity of the data is presented to the learner. For instance, human teachers often divide complex tasks into smaller sub-tasks and teach easier concepts first, followed by difficult concepts to another human. However, in machine learning, the supervision is often random, and training models have no clue about the difficulty of the sample which is being presented. One of the early seminal works by Bengio et al. [6] demonstrated the applicability of CL to machine learning and showed that the learning improves if the data is presented in a meaningful order, with a gradual increase in complexity (typically, easier to hard). Following this intuition, several methods [19, 37, 38, 42] have been proposed to determine the difficulty of the data sample and also the order in which it is presented to the network. Most of these previous methods either depend on the confidence of a pretrained model [37, 19] or human-annotators to determine sample difficulty [36]. For instance, Wei et al. [36] explored the CL in histology based on multiple annotator agreements as a proxy to estimate the difficulty of a training sample. Notably, Wu et al. [38] investigated several benefits of CL and provided thorough insights on when and where curriculum works to improve machine learning models on standard benchmark datasets. Our work takes inspiration from [37] and extends the idea of the curriculum by transfer learning to improve SSL on downstream tasks by generalizing representations to both in-domain and out-of-domain distribution data.

3. Method

Our approach consists of the following steps. First, we perform self-supervised pretraining on unlabeled data to learn histology specific visual representations. Second, we fine-tune the pretrained representations using hard examples via hardness-aware dynamic curriculum learning (HaDCL) approach. The HaDCL comprises two following stages: i) we first fine-tune the model with easy-to-hard examples (i.e., Curriculum-I stage), and ii) we then initialize the Curriculum-I model to fine-tune with hard-to-very-hard samples in the Curriculum-II stage. The details are presented next.

3.1. Self-supervised Pretraining

The goal of SSL is to first learn general visual representations with task-agnostic pretraining using unlabeled data. The pretraining is performed via solving a pretext objective, where the labels needed to train a convolutional neural network are generated within the data itself. These pretrained representations are transferred to downstream tasks by supervised fine-tuning on limited label data. In this work, we consider two prominent SSL techniques: a context-based Resolution Sequence Prediction (RSP) [31] and a contrastive learning based Momentum Contrastive Coding (MoCo) [20] approach. Our motivation behind adopting RSP and MoCo is because these techniques have shown consistent and reliable performance across a variety of histopathology tasks based on a recent study in [31].

3.2. Hardness-aware Dynamic Curriculum Learning (HaDCL)

In this section, we begin by answering the two following questions in the context of CL: **Q1.** *How to measure the hardness or difficulty of a training sample?* and **Q2.** *How to specify the order at which the training data is presented to the network?*. Before we begin, we shall setup some basic notations and definitions of CL.

Let $D = \{(x_i, y_i)\}_{i=1}^{N}$ denote the training data, where $x_i \in \mathbb{R}^d$ denotes an input sample and $y_i \in \mathcal{C}$ its corresponding label. In CL, the common approach is to train a target model $f_\theta : \mathcal{X} \rightarrow \mathcal{Y}$ with a set of non-uniformly sampled mini-batches $[\mathbb{B}_1, ..., \mathbb{B}_M] \subseteq D$ using a Stochastic Gradient Descent (SGD) optimization. To measure the instantaneous hardness of a training sample (**Q1**), we define a scoring function via curriculum by transfer learning approach [19]: $s(x_i, y_i) \in \mathbb{R}$ based on the loss value $\ell = (f_\theta(x_i), y_i)$ obtained from a pretrained self-supervised model ($f_{pre}(\cdot; \theta)$). We measure this instantaneous hardness of training samples during downstream fine-tuning by initializing f_θ with $f_{pre}(\cdot; \theta)$. We say that a sample x_j is more difficult/hard than x_i, if $\ell(f_\theta(x_j), y_j) > \ell(f_\theta(x_i), y_i)$. In this work, we consider $\ell(.,.)$ as the standard categorical cross-entropy loss to measure the instantaneous hardness of a training sample.

Unlike in the previous work [36], our proposed approach is more reliable to the training dynamics of a neural network; since it makes use of a powerful pretrained SSL model to examine the sample difficulty, which reflects the

true underlying hardness of a training sample as it is experienced by the machine learner rather than a human-teacher. Such model-based ranking of training samples is of paramount importance in histopathology, where measuring hardness level by multiple annotator agreements is costly and sometimes infeasible for large-scale applications.

Next, we focus on the order in which the training data is presented to the network (**Q2**). Typically, an easy to hard (i.e., lowest to highest score (s)) strategy is followed to determine the ordering of samples during training. However, in the context of CL, we argue that there exist two main limitations: (i) due to randomness of SGD optimization, the instantaneous hardness of each training sample can vary significantly over consecutive epochs, which may not reflect the true hardness level of a sample over time with the model being trained. This is because the easier samples stay easy throughout training since their loss value is more likely to stay at samples minima; while for hard examples, the loss value is relatively less stable during the early part of the training and gradually stabilizes as we train more on them. Thus the instantaneous hardness level of a sample tends to decrease monotonically during training and cannot be at a fixed level; (ii) further, keeping track of the instantaneous hardness of each sample up-to-date requires extra inference computation over all training samples, which can be computationally challenging for neural networks [21].

The aforementioned limitations motivated us to propose a "*hardness-aware dynamic curriculum learning* (**HaDCL**)" approach to dynamically determine the sample's instantaneous hardness level over the gradual course of training. Our proposed approach consists of a ***dual-stage*** curriculum training strategy, which we apply during downstream fine-tuning. In the first stage, we focus on *easy-to-hard* samples, and in the second stage, we focus on *hard-to-very-hard* samples for fine-tuning the pretrained SSL model.

In **Curriculum-I** (i.e., ***easy-to-hard***) stage, we first initialize the downstream fine-tuning model $f_{ft}(\cdot;\theta)$ with the pretrained SSL model $f_{pre}(\cdot;\theta)$, and compute loss for all input samples $\ell = (f_{ft}(x_i), y_i)_{i=1}^{B}$ in a mini-batch $\mathbb{B}$ using categorical cross-entropy. Next, all B samples within a mini-batch $\mathbb{B}$ are sorted in descending order by their loss value ℓ to obtain a set $\tilde{D}$. From the sorted set $\tilde{D}$, we select the top-K samples that constitute the hard examples: **top-$\boldsymbol{K}$ = $\boldsymbol{\alpha} \times \boldsymbol{B}$**, where α is parameter ($0 \leq \alpha \leq 1$) which denotes the portion of hard samples in a set $\tilde{D}$. However, relying only on the portion of hard samples in each mini-batch does not always necessarily consider varying hardness levels between mini-batches. In other words, treating all mini-batches equally may lead to sub-optimal performance as different batches will have a varying number of hard examples. Thus, the level of hardness must be smoothly adjusted according to the training dynamics of neural network to account for varying instantaneous hardness levels of training samples over time. Therefore, we choose to *dynamically* determine the mini-batch instantaneous hardness level using an adaptive **threshold** as

$$\textit{thres} = a(1 - \frac{t}{T}) + b, \tag{1}$$

where, a and b are hyperparameters (such that, $a \gg b$) which controls *thres* such that its value changes from $a + b \rightarrow b$ at uniform speed over the gradual course of training. The term t denotes the current iteration, and T indicates the total number of iterations within an epoch.

Next, we dynamically update the model weights θ in a mini-batch $\mathbb{B}$ based on the top-K samples in set $\tilde{D}$ for which the **sum of top-$\boldsymbol{K}$ loss** (i.e., $\sum_{k=1}^{K} \ell_k$) exceeds the threshold ***'thres'*** as

$$\sum_{k=1}^{K} \ell_k > \textit{thres} \times \sum_{i=1}^{B} \ell_{total}, \tag{2}$$

where, $\sum_{i=1}^{B} \ell_{total}$ is the **total loss** value over all B samples within a mini-batch $\mathbb{B}$. By doing so, we can avoid an extra inference step for keeping track of the instantaneous hardness of each sample up-to-date, which can be computationally expensive. Further, this dynamic way of updating the model parameters based on mini-batch hardness level can simultaneously alleviate both under-fitting (for hard samples) and over-fitting (for easy samples) problems. From Eq. (1), during the early phase of training (i.e., at $t \geq 1$), the model is oriented to learn with a larger number of easier examples - due to a larger threshold value *thres* $\approx a + b$; while, during the later part of the training (i.e., at $t \approx T$) fewer but hard samples are learned - due to lower threshold value of *thres* $\approx b$. This dynamic way of CL allows the model to revisit more frequently those samples that have been historically hard, while making less frequent revisits to those easier samples that have been already learned.

In **Curriculum-II** training, we start by initializing the model (θ_2) with Curriculum-I fine-tuned model (θ_1) and focus on ***hard-to-very-hard*** examples for CL. Here, we determine the instantaneous hardness of hard to very-hard samples by dynamically choosing a subset of top-K' samples, within a pre-defined set of top-K samples as: **top-$\boldsymbol{K'}$ = *thres* $\times$ top-$\boldsymbol{K}$**; where, *thres* is an adaptive threshold (see, Eq. (1)) to estimate the mini-batch instantaneous hardness level over top-K' samples.

In this stage, we dynamically update the fine-tuned Curriculum-I model weights θ_1 based on top-K' samples in a set $\tilde{D}$, for which the **sum of top-$\boldsymbol{K'}$ loss** (i.e., $\sum_{k'=1}^{K'} \ell_{k'}$) exceeds the threshold ***'thres'*** as

$$\sum_{k'=1}^{K'} \ell_{k'} > \textit{thres} \times \sum_{k=1}^{K} \ell_k, \tag{3}$$

where, $\sum_{k=1}^{K} \ell_k$ is the sum of loss values over top-K samples within a mini-batch $\mathbb{B}$, as defined in Eq. (2). The pseudocode for our proposed dual-stage HaDCL strategy is illustrated in Algorithm 1.

Algorithm 1: HaDCL method

Inputs: $D = \{(x_i, y_i)\}_{i=1}^{B}$ = training samples in mini-batch $\mathbb{B}$
$f_{pre}(\cdot;\theta)$ = SSL pretrained model
$\ell = (f_\theta(x_i), y_i) \in \mathbb{R}$ = loss / scoring function
$o \in \{$"descending"$\}$ = order
α = portion of hard samples in a set $\tilde{D}$
a, b = hyperparameters such that $a \gg b$
t = current iteration
T = total number of iterations within an epoch
$f_{ft}(\cdot;\theta)$ = fine-tuning model

$\tilde{D} = (x_1, ..., x_B) \leftarrow$ **sort**$(\{x_1, ..., x_B\}, \ell, o)$

Curriculum-I stage:

Initialize: $f_{ft}(\cdot;\theta_1) \leftarrow f_{pre}(\cdot;\theta)$, with weights θ unfrozen across entire network + a 2-layer MLP (Fc1, ReLU, Fc2) that predicts class logits

for *epoch* in $[1, ..., num\_epochs]$ **do**
  for each minibatch $\mathbb{B}_t$, where $t \in [1, ..., T]$ **do**
    top-K = $\alpha \times B$
    $thres = a(1 - \frac{t}{T}) + b$; adaptive hardness threshold
    $D' = \tilde{D}[0 : \text{top-}K]$; top-K ***hard*** samples
    $\ell_{total} = \ell(D) = \ell(f_{ft}(x_i), y_i)_{i=1}^{B}$; **total loss**
    $\ell_k = \ell(D') = \ell(f_{ft}(x_k), y_k)_{k=1}^{K}$; **top-K loss**
    if $\sum_{k=1}^{K} \ell_k > thres \times \sum_{i=1}^{B} \ell_{total}$ **then**
      Update $f_{ft}(\cdot;\theta_1)$ with ℓ_k
    else
      Update $f_{ft}(\cdot;\theta_1)$ with ℓ_{total}
    end
  end
end
return θ_1;

Curriculum-II stage:

Initialize: $f_{ft}(\cdot;\theta_2) \leftarrow \theta_1$

for *epoch* in $[1, ..., num\_epochs]$ **do**
  for each minibatch $\mathbb{B}_t$, where $t \in [1, ..., T]$ **do**
    $thres = a(1 - \frac{t}{T}) + b$; adaptive hardness threshold
    top-K = $\alpha \times B$
    top-K' = $thres \times$ top-K
    $D' = \tilde{D}[0 : \text{top-}K]$; top-K ***hard*** samples
    $D'' = D'[0 : \text{top-}K']$; top-K' ***very-hard*** samples
    $\ell_k = \ell(D') = \ell(f_{ft}(x_k), y_k)_{k=1}^{K}$; **top-$K$ loss**
    $\ell_{k'} = \ell(D'') = \ell(f_{ft}(x_{k'}), y_{k'})_{k'=1}^{K'}$; **top-$K'$ loss**
    if $\sum_{k'=1}^{K'} \ell_{k'} > thres \times \sum_{k=1}^{K} \ell_k$ **then**
      Update $f_{ft}(\cdot;\theta_2)$ with $\ell_{k'}$
    else
      Update $f_{ft}(\cdot;\theta_2)$ with ℓ_k
    end
  end
end
return θ_2;

4. Experiments

In this section, we validate our method on **three** standard benchmark datasets for breast cancer metastasis detection in lymph nodes at whole-slide-image (WSI)-level (**Camelyon16, MSK**) [5, 8] and patch-level colorectal polyps classification (**MHIST**) [36]. We choose these three datasets to investigate the relative benefits of CL on standard high and low-data training regimes. In addition, the chosen tasks embody both patch-wise and slide-level classification in histopathology and explore the problem of domain shift when training data from the target domain is entirely absent. This helps to understand the generalizability of our proposed approach and the boundaries within which the CL works to improve SSL in practice.

4.1. Datasets

We first perform self-supervised pretraining on the Camelyon16 dataset, followed by fine-tuning the pretrained model with our proposed HaDCL approach on Camelyon16 and MHIST datasets, respectively; and finally, evaluated on test sets of three datasets: Camelyon16, MSK, and MHIST. We will next introduce the datasets in detail.

Camelyon16 dataset [5]. Camelyon16 consists of 399 hematoxylin and eosin (H&E) stained WSIs (from 399 patients) of lymph nodes in the breast, divided into 270 for training and 129 for testing. The WSIs were acquired from two different centers using two different scanners with specimen level pixel sizes of $(0.226\mu m/pixel)$ and $(0.243\mu m/pixel)$. For self-supervised pretraining, we only considered 60 WSIs (slide id: normal set (1-35); tumor set (1-25)) from the total 270 training images discarding their labels (we refer to this as an ***unlabeled* set**). While, the downstream fine-tuning is performed with 228 WSIs (85%) (slide id: normal set (1-135); tumor set (1-93)) and validation with the rest 42 WSIs (15%) (slide id: normal set (136-160); tumor set (94-110)). Further, the fine-tuning set contains 400K patches (200K tumor and 200K normal), and the validation set contains 40K patches (20K tumor and 20K normal). The test set contains an independent set of 129 WSIs (49 with nodal metastases and 80 normal WSIs).

MSK dataset [7]. MSK set was released as part of a previous study in [8], which contains an independent test set of 130 H&E stained WSIs of axillary lymph nodes from 78 breast cancer patients. The nodal metastasis is present in 36 images from 27 patients with corresponding slide-level labels. The WSIs were scanned at $20\times$ magnification $(0.5\mu m/pixel)$. Note: the publicly released dataset[2] is only an independent test set and does not contain training images. MSK is considered out-of-distribution (OOD) to Camelyon because of three reasons: i) image resolution difference ($20\times$ in MSK vs. $40\times$ magnification in Camelyon); ii) technical variability in slide preparation [8]; iii) presence of cases with signs showing the effect of treatment response from neoadjuvant chemotherapy in MSK vs.

[2] https://doi.org/10.7937/tcia.2019.3xbn2jcc

no treatment response cases in Camelyon. Therefore, we choose the MSK dataset to test the generalizability of our approach to domain shift.

MHIST dataset [36]. MHIST contains a total of 3,152 images (with 224×224 pixels) for classifying colorectal polyps as between hyperplastic polyps (HPs) and sessile serrated adenomas (SSAs). This dataset is split into a training set consisting of 2,175 images, whereas the test set contains 977 images. We further divide the train set into fine-tuning set with 1740 images (80%) and a validation set of 435 images (20%). Multiple annotators annotated the images, and majority voting of labels was performed to obtain the final ground truth.

4.2. Implementation Details

We first perform **self-supervised pretraining** of RSP and MoCo on Camelyon16 *unlabeled* set using ResNet-18 as our base encoder network. We adopt similar hyper-parameter settings and domain-specific data augmentation strategies for RSP and MoCo pretraining as reported in [31]. After pretraining, we only use the ResNet encoder f_θ (that maps output to a 512-dimensional embedding) for downstream fine-tuning, discarding the project head (2-layer MLP in previous design [31]) following the suggestion in SimCLR [9]. Further, we choose to fine-tune from the first layer in the encoder f_θ with a newly initialized 2-layer MLP (Fc1, ReLU, Fc2) that predicts the class logits for final classification using all labeled samples and a standard supervised cross-entropy loss.

In our experiments, we fine-tune the pretrained model with a patch size of 256×256 pixels on Camelyon16 and MHIST datasets (224×224 resized to 256×256), respectively, followed by evaluation on Camelyon16, MSK, and MHIST test sets. Note: to account for the input resolution differences between MSK ($0.5\mu m/pixel$) and Camelyon16 datasets ($0.23-0.24\mu m/pixel$), we choose to test the camelyon16 fine-tuned model on MSK by upsampling the input patch from 256×256 to 512×512 pixels, followed by centre cropping to 256×256 pixels. For fine-tuning, we use the following sets of domain-specific data augmentations [34]: perturbations of hue and saturation values between (-0.1, 0.1) and (-1, 1), respectively in HSV color space, additive Gaussian noise with $\mu = 0$ and $\sigma = (0, 0.1)$, shifting brightness and contrast intensity ratios between (-0.2, 0.2), blurring with a random-sized kernel (3, 7), affine transformation with translation, scale and rotation limit of $(0.0625, 0.5, 45°)$, rotation with centre crop of (-90°, +90°) and finally, we scale with a factor of $(0.8, 1.2)$ and randomly resize and crop the image patch to its original size. We apply these augmentations in sequence by randomly selecting 2 of total 7 augmentations in each mini-batch, similar to RandAugment technique [15].

The **fine-tuning** is performed with three different strategies: **supervised fine-tuning** (*vanilla baseline*), **curriculum-I** and **curriculum-II** fine-tuning as described in Section. 3.2. We first list the hyperparameters common to all three strategies for **Camelyon16** dataset: we set the batch size to 512 and optimize the network with Adam optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.999$) with a weight decay of $1e-4$. Next, we train the model for 250 epochs, with an initial learning rate (lr) of $5e^{-4}$ and a multi-step decay at (60, 120, 180) epochs by 0.1 for supervised and Curriculum-I fine-tuning; while we train for 60 epochs with lr=$1e^{-5}$ and a multi-step decay at 30^{th} epoch by 0.1 for Curriculum-II stage. We set empirically the parameters k as 0.10 and a and b in Eq. 1 as (0.7, 0.2) in both Curriculum-I and II stages (refer, Section 4.3.1 for ablations). For the **MHIST** dataset, we adopted the same settings as Camelyon16, except the following parameters: we set the batch size as 32 and trained for 1000 epochs with lr=$1e^{-5}$ and a multi-step decay at (200, 400, 600, 800) epochs by 0.95 for supervised and Curriculum-I fine-tuning; while for Curriculum-II, we trained for 60 epochs with lr=$1e^{-5}$ and a multi-step decay at 30^{th} epoch by 0.95. Finally, we saved the best model based on the highest validation accuracy to test on the test set. We implemented our approach in PyTorch and trained with Nvidia V100 GPUs.

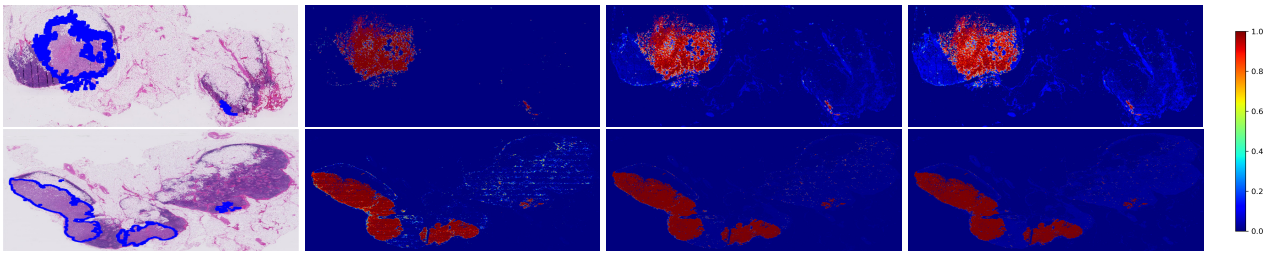
4.3. Results and Discussion

We validate the performance of our HaDCL approach with strong set of baselines: (i) **pretraining** with **RSP** [31] and **MoCo** [20] based SSL methods, along with **fully-supervised** method (***randomly*** initialized); (ii) **fine-tuning** with 3 different strategies: **Supervised (*'Baseline'***, as depicted in Table 1), **Curriculum-I** (with *easy-to-hard* examples) and **Curriculum-II** (with *hard-to-very-hard* examples). We evaluate these baselines for breast cancer metastasis detection at WSI-level (Camelyon16, MSK) and patch-level colorectal polyps classification (MHIST) tasks. For WSI-level classification, a random-forest-based slide-level classifier was used to obtain the final slide-level predictions. Similar to Wang et al. [35], we extract several geometrical features from the heatmap predictions (connected component analysis with threshold of 0.5 and 0.95) to train a final slide-level classifier. We used accuracy (Acc) and area under the receiver operating characteristic curve (AUC) as evaluation metrics for accessing both WSI-level and patch-level classification performance. Further, to check whether our HaDCL approach significantly improved the performance, we also computed statistical significance test using Delong's test [33] for pairs of AUCs between supervised (baseline) and HaDCL based fine-tuning methods. The 95% CIs were computed to access the significance at p-value < 0.05.

WSI-level Classification. The quantitative results are summarized in Table 1 and qualitative results are shown in

Table 1. WSI classification results on Camelyon16 and MSK set evaluated with WSI-level accuracy and area under the curve (AUC) with 95% CIs shown in square brackets; followed by patch-wise colorectal polyps classification on MHIST dataset evaluated with patch-level accuracy and AUC. The DeLong method [33] was used to construct 95% CIs. The best results for each dataset are highlighted in bold.

Pretraining	Fine-tuning	**Camelyon16** (slide-level)		**MSK** (slide-level)		**MHIST** (patch-level)	
		Accuracy	**AUC**	**Accuracy**	**AUC**	**Accuracy**	**AUC**
Random	Baseline	0.760	0.780 [0.595-0.804]	0.285	0.518 [0.403-0.632]	0.803	0.880
	Curriculum-I	0.853	0.814 [0.735-0.892]	0.400	0.685 [0.571-0.798]	0.802	0.889
	Curriculum-II	0.822	0.845 [0.768-0.922]	0.800	0.744 [0.645-0.842]	0.795	0.874
RSP [31]	Baseline	0.752	0.806 [0.724-0.887]	0.285	0.542 [0.428-0.654]	0.816	0.888
	Curriculum-I	0.860	0.891 [0.824-0.958]	0.654	0.743 [0.650-0.835]	0.805	0.880
	Curriculum-II	**0.891**	**0.942 [0.897-0.987]**	0.669	0.771 [0.670-0.871]	0.793	0.872
MoCo [20]	Baseline	0.744	0.837 [0.759-0.915]	**0.846**	0.749 [0.645-0.852]	0.815	0.884
	Curriculum-I	0.729	0.829 [0.751-0.906]	0.823	**0.771 [0.667-0.874]**	**0.825**	**0.896**
	Curriculum-II	0.744	0.854 [0.783-0.925]	0.808	0.771 [0.676-0.864]	0.815	0.887



(a) Original (b) Baseline (c) Curriculum-I (d) Curriculum-II

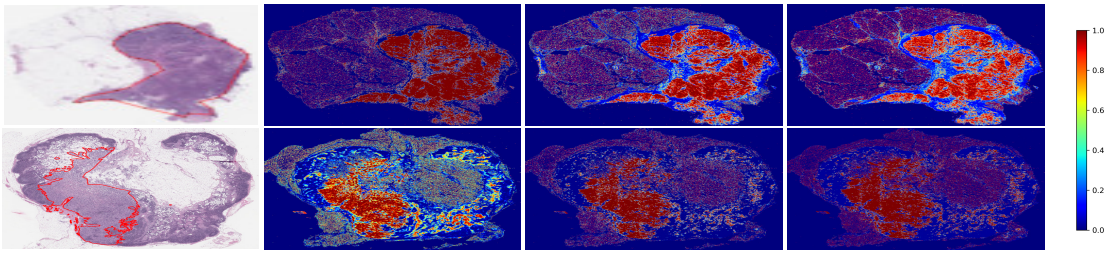
Figure 1. Predicted tumor probability heat-maps on Camelyon16 test set with RSP (**top** row) and MoCo (**bottom** row) methods. Note. (a) Original WSI's with overlaid manual ground truth (shown in blue) depicting the region containing both macro and micro-metastases.

Figure 1, 2. On the Camelyon16 dataset, we achieved statistically significant improvement in accuracy (Acc) and AUC, with a minimum score of 9.3% and 3.4%, respectively, with Curriculum-I stage against the standard baseline; while the performance of Curriculum-II improved with a minimum Acc and AUC score of 6.2% and 6.5%, respectively, over the baseline, using Random and RSP pretrained methods. On the other hand, the MoCo performance improved marginally with a 1.7% increase in AUC with Curriculum-II vs. baseline. Notably, our proposed HaDCL method achieves the best AUC score of 0.942 with 400K labeled samples compared to an AUC of 0.925 of the top-1 winning method of Camelyon16 [35], which was trained in a fully-supervised manner with millions of image patches.

We conducted further experiments to evaluate the effective robustness of SSL methods to out-of-distribution (OOD) data. For this, we first pretrain followed by fine-tuning the model on Camelyon16 but tested on MSK dataset. We observed significant improvement in SSL methods on OOD data, particularly when fine-tuned with curriculum-I and II approaches over the standard baseline, as shown in Table 1. We observed larger gains with minimum improvement in Acc and AUC score of 11.5% and 16.7%, respectively, with Curriculum-I stage vs. standard baseline; whereas Curriculum-II's performance further improved over baseline, with a minimum increase in Acc and AUC score of 38.4% and 22.6%, respectively, using Random and RSP methods. Furthermore, the performance with MoCo also improved with a 2.2% increase in Acc and AUC score with both Curriculum-I and -II over baseline approach. This significant improvement under domain shift is of paramount importance in real clinical settings [16, 32], where the model trained on Camelyon16 with images acquired with higher resolution ($0.25\mu m/pixel$) can generalize satisfactorily to the OOD MSK test set, which was acquired with a lower resolution ($0.5\mu m/pixel$).

Overall, our results provide evidence that representations learned by SSL methods can be further enhanced and made more generalizable to out-of-domain distribution by effectively leveraging difficult examples during fine-tuning. This observation is also **consistent** with **recent studies** in [1, 38]; where the authors have shown that the effective robustness of pretraining models can be further enhanced with a more extensive and diverse set of pretraining samples followed by fine-tuning with more difficult and noisy samples. Thus, our experimental findings clearly demonstrate that the hardness-aware curriculum learning has a superior advantage over the standard fine-tuning in improving SSL methods. Further, it is interesting to explore the effect of Curriculum fine-tuning of SSL methods under a limited labeled regime, which has significant opportunities for further enhancements as shown in a recent study in [31].

Patch-level Classification. Table 1 presents the colorectal polyps patch-wise classification results on MHIST dataset.



(a) Original (b) Baseline (c) Curriculum-I (d) Curriculum-II

Figure 2. Out-of-distribution prediction results showing tumor probability heat-maps on MSK test set (**target**) trained from Camelyon16 (**source**) with RSP (**top** row) and MoCo (**bottom** row) methods. Note. (a) Original WSI's with overlaid manual ground truth (shown in red) annotated by our in-house pathologist.

Table 2. Impact of parameter α which denotes the portion of hard samples in each mini-batch. These experiments were performed on the Camelyon16 validation set with RSP based SSL method [31] using the Curriculum-I approach.

α	**0.05**	**0.10**	**0.15**	**0.20**
Accuracy	0.8511	0.8837	0.7441	0.6511
AUC	0.8920	0.9176	0.7269	0.6892

On this task, we didn't observe any significant improvement with the HaDCL approach over the standard baseline. However, we obtained a marginal improvement in AUC of 0.895 compared with the recent CL based method [36] with AUC of 0.882. Unlike the previous method [36], our approach doesn't depend on the annotator agreement to determine the sample hardness but rather estimates the sample hardness via knowledge transfer from a powerful pre-trained SSL model. One of the main reasons for no significant improvement is because the patch-wise dataset usually does not capture all diversity of hardness that is presented in the data compare to the level of hardness that is present in the WSI. Further, most of the curated patches are often very clean and carefully hand-picked, which lacks the level of difficulty/hardness suitable for training a model. This phenomenon has also been studied in recent work [13], where the authors show evidence that injecting hard negatives samples for patch-wise classification has been shown to degrade performance, whilst the performance improves significantly for slide-level classification tasks. Notably, this phenomenon was also shown to be consistent on vision tasks [38, 37], where CL has shown almost no improvement on standard benchmark datasets such as CIFAR10 and CIFAR100; while, it improves only when the task is made more difficult.

4.3.1 Ablation Study

Table 2 shows the effect of parameter α that selects the portion of hard samples in each mini-batch in our formulation. We observe that varying α to large values (> 0.10) will lead to a large selection of easy samples, thus deteriorating the performance; on the other hand, selecting α to small values (< 0.10) over exaggerates the hard samples leading to under-fitting. Thus, we empirically found $\alpha = 0.10$ as the optimal choice based on the validation performance of the Camelyon16 set that selects sufficient hard examples to balance between over-fitting to easy samples or under-fitting to hard samples. Next, we empirically chose the value of (a, b) as $(0.7, 0.2)$ in a reasonable range such that the threshold $thres$ in Eq. 1 changes at uniform speed from $a + b \rightarrow b$ which is $0.9 \rightarrow 0.2$ during the gradual course of training. More intuition on selection of (a, b) with respect to training dynamics of neural network is discussed in Section 3 (Curriculum-I). However, note that we fixed these parameters constant across all three datasets, and we find that the above choices of (a, b) are less sensitive to data distribution.

5. Conclusion

We introduce HaDCL, a method for improving self-supervised learning to both in-domain and out-of-domain distribution data, and also slide-level and patch-wise classification tasks in histopathology. By dynamically leveraging the hard examples during downstream mini-batch fine-tuning, we learn robust features that are adaptable to different domains with significant domain shifts. Our approach is more generic and adaptable to different SSL methods and does not involve any additional overhead complexity. Through experiments, we demonstrated state-of-the-art classification results on three histology benchmark datasets with a significant performance improvement on an external test set with notable domain-shift. We believe HaDCL may prove to be a useful stepping stone in generalizing the pre-trained representations to various downstream tasks under a limited annotation setting. Future research will focus on extending the approach to mixed supervision to simultaneously exploit both pixel-level and image-level annotations for slide-level prediction tasks.

References

[1] Anders Andreassen, Yasaman Bahri, Behnam Neyshabur, and Rebecca Roelofs. The evolution of out-of-distribution robustness throughout fine-tuning. *arXiv preprint arXiv:2106.15831*, 2021. 1, 3, 7

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

[3] 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. *arXiv preprint arXiv:2101.05224*, 2021. 1, 2

[4] Wenjia Bai, Chen Chen, Giacomo Tarroni, Jinming Duan, Florian Guitton, Steffen E Petersen, Yike Guo, Paul M Matthews, and Daniel Rueckert. Self-supervised learning for cardiac mr image segmentation by anatomical position prediction. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 541–549, 2019. 2

[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. 5

[6] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th International Conference on Machine Learning*, pages 41–48, 2009. 1, 2, 3

[7] Gabriele Campanella, Matthew G. Hanna, Edi Brogi, and Thomas J. Fuchs. Breast metastases to axillary lymph nodes, 2019. 5

[8] 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. 5

[9] 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, 2020. 1, 2, 6

[10] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey Hinton. Big self-supervised models are strong semi-supervised learners. *arXiv preprint arXiv:2006.10029*, 2020. 1

[11] Tianlong Chen, Sijia Liu, Shiyu Chang, Yu Cheng, Lisa Amini, and Zhangyang Wang. Adversarial robustness: From self-supervised pre-training to fine-tuning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 699–708, 2020. 3

[12] Ozan Ciga, Anne L Martel, and Tony Xu. Self supervised contrastive learning for digital histopathology. *arXiv preprint arXiv:2011.13971*, 2020. 1, 2

[13] Ozan Ciga, Tony Xu, Sharon Nofech-Mozes, Shawna Noy, Fang-I Lu, and Anne L Martel. Overcoming the limitations of patch-based learning to detect cancer in whole slide images. *Scientific Reports*, 11(1):1–10, 2021. 8

[14] Elijah Cole, Xuan Yang, Kimberly Wilber, Oisin Mac Aodha, and Serge Belongie. When does contrastive visual representation learning work? *arXiv preprint arXiv:2105.05837*, 2021. 3

[15] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 702–703, 2020. 6

[16] Thomas de Bel, John-Melle Bokhorst, Jeroen van der Laak, and Geert Litjens. Residual cyclegan for robust domain transformation of histopathological tissue slides. *Medical Image Analysis*, 70:102004, 2021. 7

[17] Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1422–1430, 2015. 2

[18] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*, 2018. 2

[19] Guy Hacohen and Daphna Weinshall. On the power of curriculum learning in training deep networks. In *International Conference on Machine Learning*, pages 2535–2544, 2019. 2, 3

[20] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020. 1, 2, 3, 6, 7

[21] Angela H Jiang, Daniel L-K Wong, Giulio Zhou, David G Andersen, Jeffrey Dean, Gregory R Ganger, Gauri Joshi, Michael Kaminksy, Michael Kozuch, Zachary C Lipton, et al. Accelerating deep learning by focusing on the biggest losers. *arXiv preprint arXiv:1910.00762*, 2019. 4

[22] Longlong Jing and Yingli Tian. Self-supervised visual feature learning with deep neural networks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 2

[23] Navid Alemi Koohbanani, Balagopal Unnikrishnan, Syed Ali Khurram, Pavitra Krishnaswamy, and Nasir Rajpoot. Self-path: Self-supervision for classification of pathology images with limited annotations. *IEEE Transactions on Medical Imaging*, 2021. 2

[24] Chunyuan Li, Xiujun Li, Lei Zhang, Baolin Peng, Mingyuan Zhou, and Jianfeng Gao. Self-supervised pre-training with hard examples improves visual representations. *arXiv preprint arXiv:2012.13493*, 2020. 1, 2, 3

[25] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European Conference on Computer Vision*, pages 69–84, 2016. 2

[26] Senthil Purushwalkam and Abhinav Gupta. Demystifying contrastive self-supervised learning: Invariances, augmentations and dataset biases. *arXiv preprint arXiv:2007.13916*, 2020. 3

[27] Colorado J Reed, Xiangyu Yue, Ani Nrusimha, Sayna Ebrahimi, Vivek Vijaykumar, Richard Mao, Bo Li, Shanghang Zhang, Devin Guillory, Sean Metzger, et al. Self-supervised pretraining improves self-supervised pretraining. *arXiv preprint arXiv:2103.12718*, 2021. 3

[28] Joshua Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. Contrastive learning with hard negative samples. *arXiv preprint arXiv:2010.04592*, 2020. 1, 2, 3

[29] Petru Soviany, Radu Tudor Ionescu, Paolo Rota, and Nicu Sebe. Curriculum learning: A survey. *arXiv preprint arXiv:2101.10382*, 2021. 1, 2

[30] Hari Sowrirajan, Jingbo Yang, Andrew Y Ng, and Pranav Rajpurkar. Moco-cxr: Moco pretraining improves representation and transferability of chest x-ray models. *arXiv preprint arXiv:2010.05352*, 2020. 2

[31] Chetan L Srinidhi, Seung Wook Kim, Fu-Der Chen, and Anne L Martel. Self-supervised driven consistency training for annotation efficient histopathology image analysis. *arXiv preprint arXiv:2102.03897*, 2021. 1, 2, 3, 6, 7, 8

[32] Karin Stacke, Gabriel Eilertsen, Jonas Unger, and Claes Lundström. A closer look at domain shift for deep learning in histopathology. *arXiv preprint arXiv:1909.11575*, 2019. 7

[33] Xu Sun and Weichao Xu. Fast implementation of delong's algorithm for comparing the areas under correlated receiver operating characteristic curves. *IEEE Signal Processing Letters*, 21(11):1389–1393, 2014. 6, 7

[34] David Tellez, Geert Litjens, Péter Bándi, Wouter Bulten, John-Melle Bokhorst, Francesco Ciompi, and Jeroen van der Laak. Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology. *Medical Image Analysis*, 58:101544, 2019. 6

[35] Dayong Wang, Aditya Khosla, Rishab Gargeya, Humayun Irshad, and Andrew H Beck. Deep learning for identifying metastatic breast cancer. *arXiv preprint arXiv:1606.05718*, 2016. 6, 7

[36] Jerry Wei, Arief Suriawinata, Bing Ren, Xiaoying Liu, Mikhail Lisovsky, Louis Vaickus, Charles Brown, Michael Baker, Mustafa Nasir-Moin, Naofumi Tomita, et al. Learn like a pathologist: curriculum learning by annotator agreement for histopathology image classification. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, pages 2473–2483, 2021. 2, 3, 5, 6, 8

[37] Daphna Weinshall, Gad Cohen, and Dan Amir. Curriculum learning by transfer learning: Theory and experiments with deep networks. In *International Conference on Machine Learning*, pages 5238–5246, 2018. 2, 3, 8

[38] Xiaoxia Wu, Ethan Dyer, and Behnam Neyshabur. When do curricula work? *arXiv preprint arXiv:2012.03107*, 2020. 2, 3, 7, 8

[39] Tete Xiao, Xiaolong Wang, Alexei A Efros, and Trevor Darrell. What should not be contrastive in contrastive learning. *arXiv preprint arXiv:2008.05659*, 2020. 3

[40] Xueting Yan, Ishan Misra, Abhinav Gupta, Deepti Ghadiyaram, and Dhruv Mahajan. Clusterfit: Improving generalization of visual representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6509–6518, 2020. 1, 2

[41] Xiaohua Zhai, Avital Oliver, Alexander Kolesnikov, and Lucas Beyer. S4l: Self-supervised semi-supervised learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1476–1485, 2019. 1, 2

[42] Tianyi Zhou, Shengjie Wang, and Jeff A Bilmes. Curriculum learning by dynamic instance hardness. *Advances in Neural Information Processing Systems*, 33, 2020. 3