Benchmarking Self-Supervised Learning on Diverse Pathology Datasets

Mingu Kang* Heon Song* Seonwook Park Donggeun Yoo Sérgio Pereira
Lunit Inc.
{jeffkang, heon.song, spark, dgyoo, sergio}@lunit.io

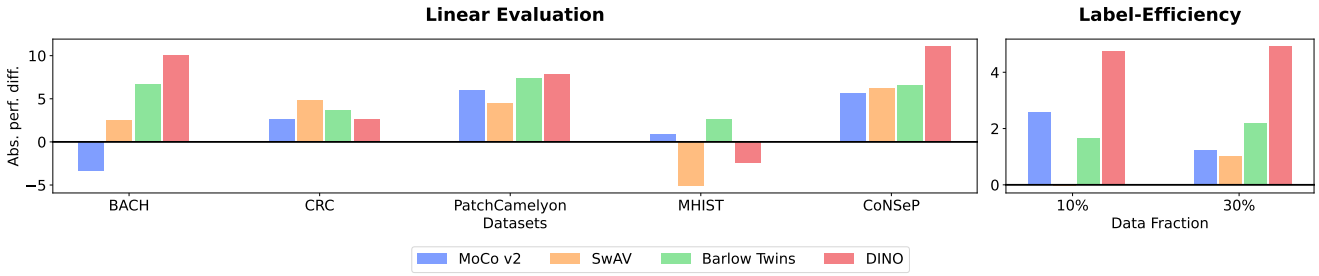

Figure 1. **Self-supervised pre-training on pathology data improves performance on pathology downstream tasks** compared to ImageNet-supervised baselines. The *y*-axes show absolute differences in downstream task performance (Top-1 Acc. or mPQ Score). Linear evaluation (**left**) is performed on 4 classification tasks (BACH, CRC, PatchCamelyon, and MHIST) and 1 nuclei instance segmentation task (CoNSeP). Label-efficiency (**right**) is assessed by fine-tuning using small fractions of labeled data from the CoNSeP dataset.

Abstract

Computational pathology can lead to saving human lives, but models are annotation hungry and pathology images are notoriously expensive to annotate. Self-supervised learning (SSL) has shown to be an effective method for utilizing unlabeled data, and its application to pathology could greatly benefit its downstream tasks. Yet, there are no principled studies that compare SSL methods and discuss how to adapt them for pathology. To address this need, we execute the largest-scale study of SSL pre-training on pathology image data, to date. Our study is conducted using 4 representative SSL methods on diverse downstream tasks. We establish that large-scale domain-aligned pre-training in pathology consistently out-performs ImageNet pre-training in standard SSL settings such as linear and fine-tuning evaluations, as well as in low-label regimes. Moreover, we propose a set of domain-specific techniques that we experimentally show leads to a performance boost. Lastly, for the first time, we apply SSL to the challenging task of nuclei instance segmentation and show large and consistent performance improvements. We release the pre-trained model weights[1].

1. Introduction

The computational analysis of microscopic images of human tissue – also known as computational pathology – has emerged as an important topic of research, as its clinical implementations can result in the saving of human lives by improving cancer diagnosis [49] and treatment [42]. Deep Learning and Computer Vision methods in pathology allow for objectivity [15], large-scale analysis [20], and triaging [5] but often require large amounts of annotated data [52]. However, the annotation of pathology images requires specialists with many years of clinical residency [37], resulting in scarce labeled public datasets and the need for methods to train effectively on them.

When annotated data is scarce for a given Computer Vision task, one common and practical solution is to fine-tune a model that was pre-trained in a supervised manner using the ImageNet dataset [19, 34]. This paradigm of transfer learning [34] was recently challenged by self-supervised learning (SSL), which trains on large amounts of unlabeled data only, yet out-performs supervised pre-training on ImageNet [8, 10, 26]. In the field of pathology, large unlabeled datasets are abundant [4, 37, 38, 57] in contrast to the lack of annotated datasets [52]. If we were to apply SSL effectively to this huge amount of unlabeled data, downstream pathology tasks could benefit greatly even if they contain limited amount of annotated training data. Naturally, we ask the question: *How well does self-supervised learning help in improving the performance of pathology tasks?*

ImageNet pre-trained weights are commonly used in medical imaging and are known to be helpful in attaining high task performance [30, 32, 43, 59]. Due to the difference between natural images and medical images, large-scale domain-aligned pre-training has the potential to push performance beyond ImageNet pre-training [39]. Accordingly, recent works show that SSL pre-training on pathol-

*The first two authors contributed equally.

[1] https://lunit-io.github.io/research/publications/pathology_ssl

ogy data can improve performance on downstream pathology tasks [3, 16, 23, 55]. Our study aims to expand on these previous works by evaluating multiple SSL methods on diverse downstream pathology tasks. In addition, we propose techniques to adapt SSL methods that were designed for natural image data, to better learn from pathology data.

To understand how to adapt existing SSL methods to work on pathology image data, we must identify several key differences between natural and pathology imagery. Unlike natural images, pathology images can be rotated arbitrarily (impossible to determine a canonical orientation) and exhibit fewer variations in color. Also, pathology images can be interpreted differently depending on the field-of-view (FoV) due to the multiple hierarchies and contextual differences involved in each task. We propose to overcome these differences when adapting SSL methods for pathology data, via changes to the training data augmentation scheme in particular, during pre-training.

In this paper, we carry out an in-depth analysis of 4 recent and representative SSL methods; MoCo v2 [12], SwAV [7], Barlow Twins [61], and DINO [8], when applied to large-scale pathology data. For this purpose, we source 19 million image patches from Whole Slide Images (WSI) in The Cancer Genome Atlas (TCGA) dataset [57] and apply our domain-specific techniques in training the SSL methods on this data. The evaluations are conducted for 2 different downstream tasks over 5 datasets: (1) pathological image classification using BACH [1], CRC [31], MHIST [56], and PatchCamelyon [54] datasets, and (2) nuclei instance segmentation and classification using the CoNSeP dataset [25].

Our large-scale study yields several useful contributions: (a) we conduct the largest-scale study of SSL pre-training on pathology image data to date, and show its benefit over using ImageNet pre-trained weights on diverse downstream tasks (see Fig. 1), (b) we propose a set of carefully designed data curation and data augmentation techniques that can further improve downstream performance, (c) we demonstrate that SSL is label-efficient, and is therefore a practical solution in pathology where gathering annotation is particularly expensive, and (d) for the first time, we apply SSL to the dense prediction task of nuclei instance segmentation and show its value under diverse evaluation settings. We release our pre-trained model weights at https://lunit-io.github.io/research/publications/pathology_ssl to further contribute to the research community.

2. Related Work

2.1. Self-supervised Representation Learning

SSL methods learn representations through pre-text tasks designed to exploit supervisory signals obtained from the unlabeled data itself. We describe the 4 major paradigms of SSL as commonly discussed in literature.

Contrastive Learning. Contrastive methods [27, 40, 41] such as SimCLR [10] and MoCo v2 [12] learn to discriminate each training data instance from all the others. The objective is to learn similar representations of positive pairs (perturbations by data augmentation) and discriminative representations in relation to negative pairs (other instances). A limitation is the need for diverse negative pairs, which is mitigated through large batch sizes [10] or memory banks [12]. In this work, we explore MoCo v2 [12].

Non-contrastive Learning. Methods such as BYOL [26], SimSiam [13], and Barlow Twins [61], share similarities with *contrastive learning* methods in that they learn representations of images under different augmented views. The fundamental difference is that these approaches do not rely on negative pairs, which allows them to work with small batch sizes. In this work, we explore Barlow Twins [61].

Clustering. This paradigm uses the concept of clustering and is shown in DeepCluster [6] and SwAV [7]. Clustering-based SSL discriminates between clusters of image representations instead of explicit pairs of images. In this work, we explore SwAV [7].

SSL with VisionTransformer. The effectiveness of Vision Transformers (ViT) [21] has been demonstrated on various computer vision tasks. Thus, the paradigm shift from CNN to ViT has recently emerged in the field of self-supervised learning. Consequently, recent studies [8, 14, 36] try to investigate techniques that facilitate SSL with ViT-based architectures. In this work, we explore DINO [8].

2.2. SSL in Medical Imaging

Recently, [39] investigates transfer learning in medical imaging and observes that using domain-aligned datasets for pre-training improves the transferability of models. Moreover, domain-specific SSL methods can further improve the performance of models fine-tuned on downstream medical image-related tasks [2, 3, 16, 23, 48, 55, 60]. In pathology, [55] employs BYOL and evaluates pre-trained weights learned from pathology data on image classification tasks. [23] adopts SimSiam, showing that SSL improves pathology image retrieval. Also recently, [16] uses SimCLR and observes that SSL consistently improves on downstream pathology tasks compared to ImageNet pre-training.

Unlike previous works that focus on just a single SSL approach [9, 16, 35], or either CNNs or ViTs only [60], we explore one representative method from each of the aforementioned SSL paradigms including ViT-based SSL. In this way, we establish a common and fair benchmark for comparing these methods in pathology. Furthermore, we evaluate the domain-specific pre-trained weights on various downstream tasks, including the challenging task of nuclei instance segmentation. Finally, we devise techniques for data augmentation that are specifically tailored to tackle

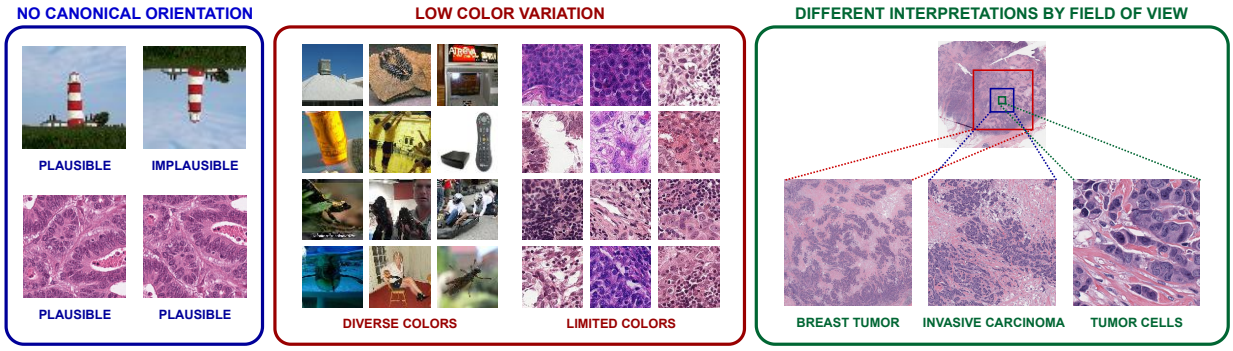

Figure 2. **Pathology images vs natural images.** Pathology images are different from natural images in 3 major ways. They have no canonical orientation (no "right way up"), have low color variation, and can be interpreted differently depending on field-of-view. Consequently, self-supervised learning methods need to be implemented differently when working in the domain of pathology.

pathology-specific challenges, thus leading to better representations and performance in the downstream tasks.

3. Self-supervised Pre-training for Pathology

The performance of SSL methods can vary greatly depending on the composition of training data and the selected set of data augmentation methods. SSL methods in the literature are commonly designed and evaluated in settings involving natural images and may benefit from further adaptation when applied to different domains, such as pathology. In this section, we discuss the differences between natural images and pathology images. We also propose a set of techniques that can be easily adopted to improve the performance of models pre-trained on pathology image data.

3.1. Differences to Natural Images

Images contained in popular Computer Vision datasets (e.g. ImageNet [19]) are often denoted as "natural images". Pathology images have several unique characteristics that make them distinct from natural images. We discuss these differences in this section and summarize them in Fig. 2.

No canonical orientation. Objects or scenes contained in natural images are oriented based on plausibility, i.e. how a human expects the objects to be oriented. Methods in Computer Vision can take advantage of such assumptions or patterns (e.g. Manhattan World assumption [17]) and thus SSL methods do not randomly augment the orientation of images at training time. However, pathology images can be oriented in any way and still remain plausible. Furthermore, objects (e.g. cells) are many and dispersed at arbitrary locations, making it impossible to define a "canonical orientation", i.e. the correct standard orientation.

Low color variation. While natural images contain a large range of colors due to the diversity of represented objects, pathology images tend to display similar color distributions (e.g. purple and pink staining). Though the staining can vary across institutions and the same biological structures have different appearances depending on the cancer type, pathology images are more consistent than natural images.

Different FoVs. To correctly analyze pathology images, different Field of Views (FoVs) must be considered. A larger FoV allows pathologists and algorithms to better understand the larger context of the tissue regions and cell classes to make high-level predictions such as the grading of prostate cancer [4]. In other tasks that require the classification of individual cells or communities of cells, a smaller FoV is required to increase the resolution on the objects of interest [25]. Thus, a pre-trained model for pathology should ideally be able to handle tasks from diverse FoVs.

3.2. Techniques to Adapt SSL for Pathology

In this section, we introduce our techniques for adapting SSL methods for pathology imagery.

Random vertical flips. Unlike natural images, pathology images are no less plausible or realistic when they are vertically flipped. We therefore propose to randomly apply vertical flips during SSL pre-training.

Stain augmentations. The typical color distortion employed by SSL methods applies a strong amount of jitter to brightness, contrast, and saturation, resulting in pathology images that look highly unrealistic. [50] proposes to apply this jitter in pathology-specific color spaces such as the HED-space [46]. [47] points out that naive jittering can produce unrealistic images and proposes RandStainNA. RandStainNA fits unimodal Gaussian distributions to the channel-wise statistics of 3 color spaces (HSV, Lab, HED) using images from the training dataset. At training time, a color space is randomly chosen, then the target channel-wise mean and standard deviations for that color space are sampled from the fitted Gaussian distributions. Reinhard's method [44] is used to re-normalize the input image to match the target statistics. RandStainNA is shown to improve supervised learning performance for pathology, and therefore we adopt it for our SSL pre-training.

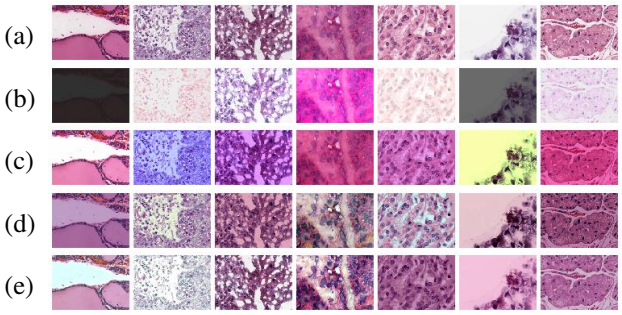

Figure 3. **Color augmentations on pathology images.** (a) input images, (b) color jitter as used in typical SSL methods [10], (c) HED-light [50], (d) RandStainNA [47], and (e) RandStainNA$_{GMM}$. RandStainNA and RandStainNA$_{GMM}$ can produce more realistic and plausible pathology images.

Data source	No. of WSIs	No. of patches		
		20x	40x	Total
TCGA	20,994	9,497,768	9,502,301	19,000,069
TULIP	15,672	7,084,130	6,494,358	13,578,488
Total	36,666	16,581,898	15,996,659	32,578,557

Table 1. **Unlabeled data for pre-training**. Amount of data used for pre-training in terms of the number of WSIs and the actual number of extracted patches (per FoV).

Furthermore, we attempt to improve the realism of RandStainNA by fitting a Gaussian Mixture Model (GMM) with 10 components, to the channel-wise statistics of each color space. The GMM can fit the covariances between variables and respect the multi-modality of the channel-wise mean and standard deviation values. We denote this as RandStainNA$_{GMM}$ and show the visual differences against alternative methods in Fig. 3.

Lastly, we observe that previous works in SSL [10, 13, 26, 61] highlight the importance of color distortion. We therefore propose to apply color distortion with a weaker jittering strength as done in [16]. Our main experiments adopt RandStainNA$_{GMM}$ along with random grayscale and a weaker color jittering augmentation.

Using multiple FoVs. As aforementioned, some pathology tasks require high FoV while others benefit from low FoV. We identify that pathology tasks (e.g., image classification and instance segmentation) are commonly defined at approximately 20× [1,31] or 40× [25,54,56] objective magnification. Therefore, we build our large-scale unlabeled dataset using image patches from both magnifications.

4. Experiment Setup

4.1. Pre-training Dataset

Tab. 1 presents the scale of unlabeled data used for pre-training. We first collect 20,994 WSIs from The Cancer Genome Atlas (TCGA) and 15,672 WSIs from TULIP. Both datasets consist of Hematoxylin & Eosin (H&E) stained WSIs of various cancers. TCGA is publicly available and widely used for training deep learning models [18, 20, 42]. TULIP is an internally collected dataset. To increase diversity and keep our experimental setting practical, we extract at most 1,000 patches of resolution 512 × 512 pixels from each slide, resulting in a total of 32.6M patches (19M from TCGA and 13.6M from TULIP). The pre-training data covers two different FoVs; 20× (0.5μm/px) and 40× (0.25μm/px) objective magnification. All experiments, unless specified otherwise, present the results of pre-training on the TCGA dataset only.

Dataset	# Classes	# Patches	Patch size	FoV	Task
BACH	4	400	2048×1536	20×	Cls
CRC	9	107,180	224×224	20×	Cls
PCam	2	327,680	96×96	40×	Cls
MHIST	2	3,152	224×224	40×	Cls
CoNSeP	7	41	1000×1000	40×	Seg

Table 2. **Datasets for downstream tasks**. Note that, Cls indicates "image classification" and Seg is "nuclei instance segmentation".

4.2. Downstream Datasets

We validate the pre-trained models under classification and segmentation tasks using various downstream datasets described in Tab. 2. For image classification, the following four datasets are used: BACH (four-class breast cancer type classification) [1], CRC (nine-class human colorectal cancer type classification) [31], MHIST (binary class colorectal polyp type classification) [56], and PCam (binary class breast cancer type classification) [54]. The patches of the datasets are labeled according to the predominant cancer type or the presence of cancers. For nuclei instance segmentation, we use CoNSeP [25] which contains segmentation masks for each cell nucleus along with nuclei types. Further details of the downstream datasets are shown in the supplementary materials.

4.3. Pre-training Details

We learn representations using 4 different SSL methods. Unless otherwise mentioned, we use the ResNet-50 (1×) [29] architecture for MoCo v2 [12], Barlow Twins [61], and SwAV [7]. For DINO [8], we use ViT-Small [21] with different patch sizes, 16×16 and 8×8 (denoted DINO$_{p=16}$ and DINO$_{p=8}$, respectively), as it has a comparable number of parameters to ResNet-50 (1×). We follow the proposed recipe of each SSL method and launch the pre-training, distributed across 64 NVIDIA V100 GPUs. The linear scaling rule [24] is applied to adjust the learning rate: $lr = lr_{method} * batchsize/256$. We adopt the concept of the *"ImageNet epoch"* from [51] for ease of analysis and train models for 200 ImageNet epochs, across all experiments. We

Arch.	Method	BACH		CRC		PCam		MHIST	
		Linear	Fine-tune	Linear	Fine-tune	Linear	Fine-tune	Linear	Fine-tune
ResNet-50	*Random*	51.67	61.67	68.91	89.99	76.52	75.71	63.15	75.54
	Supervised	80.83	86.67	90.93	92.09	80.79	80.63	76.25	78.92
	MoCo v2	77.50	**90.83**	93.52	**96.21**	86.78	**87.62**	77.07	**85.88**
	SwAV	83.33	82.50	**95.78**	93.31	85.28	87.60	71.14	77.99
	BT	**87.50**	85.00	94.60	93.23	**88.15**	86.92	**78.81**	81.27
ViT-S	$Random_{p=16}$	45.00	57.50	69.90	86.10	74.43	75.42	63.46	62.13
	$Supervised_{p=16}$	75.83	85.83	91.56	95.81	80.96	88.30	**78.51**	**81.68**
	$DINO_{p=16}$	**85.83**	87.50	94.19	95.81	88.78	90.40	76.15	79.43
	$DINO_{p=8}$	83.33	**93.33**	**95.29**	**97.13**	**90.12**	**90.76**	77.89	78.40

Table 3. **Downstream evaluation of the image classification tasks**. We report Top-1 accuracy for both linear and fine-tuning experiment protocols for models trained using the TCGA data source. Note that, p represents the patch size used in ViT.

define 10 ImageNet epochs for the warmup, and the cosine scheduler is followed. The details of configurations can be found in supplementary materials.

4.4. Downstream Training Details

For the downstream tasks, we follow the standard practice as introduced in various SSL papers [10, 26, 61]. For image classification tasks, the datasets are split into training, validation, and test sets. We perform the hyper-parameter search based on the validation set, reporting the performance on the test set. For the segmentation task, the Hover-Net [25] architecture – a standard architecture in the nuclei instance segmentation task – is adopted with the pre-trained backbone. We follow the same data pre-processing and training schemes as in [25] to enable reproducibility and fair comparison of results. Further details of downstream tasks can be found in the supplementary materials.

4.5. Evaluation Metrics

For image classification, we report top-1 accuracy, while using panoptic quality (PQ) [33] for nuclei instance segmentation. PQ is a standard metric for assessing the performance of nuclear instance segmentation [25] that accounts for both detection and segmentation quality with respect to each instance. The PQ metric is defined as,

$$PQ = \frac{\sum_{(p,g)\in TP} \text{IoU}(p,g)}{|TP| + \frac{1}{2}|FP| + \frac{1}{2}|FN|}, \quad (1)$$

where p denotes a predicted mask for each nuclei class and g denotes a corresponding ground truth. The numerator $\sum_{(p,g)\in TP} \text{IoU}(p,g)$ represents the summation of correctly matched Intersection Over Union (IoU) over all pairs between predictions and ground truth. We count pairs between predictions and ground truths with IoU of more than 0.5 as True Positives (TP). False Positives (FP) and False Negatives (FP) represent wrongly predicted predictions and ground truth, respectively. Note that we use multi-class PQ (mPQ) to measure the performance of instance segmentation and classification simultaneously.

5. Experimental Results

In this section, we carry out various experiments on the downstream tasks of image classification and nuclei instance segmentation. Through these experiments, we compare the utility of various SSL pre-training methods in the light of downstream pathology task performance. First, we present our quantitative results using well-established evaluation protocols. We then demonstrate the benefit of SSL pre-training with a limited number of labeled data and under different fine-tuning schedules. Finally, we perform an ablation study to quantitatively verify the efficacy of our techniques for adapting SSL methods to pathology data.

In evaluating downstream task performance, we stick to well-established evaluation protocols in SSL. The first is *linear evaluation* (denoted *Linear*), where we freeze the backbone and train the remaining parts of the model (e.g., linear classifier or decoders). The second is *full fine-tuning* (denoted *Fine-tune*), where all layers including the backbone are fine-tuned. The former protocol assesses the quality of learned representations, whereas the latter evaluates the transferability of learned weights. In our experiments, we compare against the ImageNet-supervised (denoted *Supervised*) pre-training baseline of the corresponding backbone type as well as a random initialization (denoted *Random*) baseline.

5.1. Image Classification

We present our linear evaluation and fine-tuning results for 4 image classification benchmarks in Tab. 3.

Linear evaluation. In linear evaluation results, we find that self-supervised TCGA pre-training typically outperforms supervised ImageNet pre-training. Of the ResNet-50 based SSL methods, Barlow Twins performs consistently well, out-performing other methods on the BACH,

Arch.	Method	CoNSeP	
		Linear	Fine-tune
ResNet-50	*Random*	22.29	46.72
	Supervised	34.25	49.60
	MoCo v2	39.85	**51.75**
	SwAV	40.45	51.16
	BT	**40.79**	51.61
ViT-S	$Random_{p=16}$	20.55	27.19
	$Supervised_{p=16}$	21.43	36.70
	$DINO_{p=16}$	32.54	38.43
	$DINO_{p=8}$	**42.71**	**46.70**

Table 4. **Downstream evaluation for the nuclei instance segmentation task**. We report the mPQ score for both linear and fine-tuning experiment protocols for models trained using the TCGA data source.

PCam, and MHIST datasets. Of the ViT-based SSL methods, $DINO_{p=16}$ achieves competitive results, and $DINO_{p=8}$ performs even better on the CRC and PCam datasets. The improved performance of $DINO_{p=8}$ is in line with observations from [8] which shows that performance improves at the cost of computation with smaller patch size. One exception is on the MHIST dataset, where the supervised baseline shows good performance. Based on the linear evaluation results, we can certainly claim that domain-aligned pre-training improves representation quality.

Fine-tuning. Under fine-tuning, we find that the trends are similar but with notable changes. Firstly, as shown in other SSL works, the performance gap between ImageNet-supervised and TCGA-SSL reduces. Furthermore, MoCo v2 shows consistently high performance among CNN methods, showing that it may be the architecture of choice for transfer learning settings using CNNs. Regarding ViTs, we find that the trends are almost identical to linear evaluation except that the fine-tuned performances are often better than CNN counterparts trained using SSL on TCGA data. For classification tasks, then, SSL using ViT on large-scale pathology data is beneficial.

5.2. Nuclei Instance Segmentation

To the best of our knowledge, we show for the first time, the effect of self-supervised domain-aligned pre-training on a downstream dense prediction task. We run experiments on the CoNSeP dataset for the task of nuclei instance segmentation and report the mPQ score in Tab. 4.

CNN experiments. The performance of SSL using ResNet-50 shows similar trends to the case of image classification, where Barlow Twins performs well on the linear evaluation protocol and MoCo v2 performs well in fine-tuning. More consistently than in the case of classification, SSL pre-trained models out-perform supervised ImageNet pre-training by a large margin, especially considering the difficulty of increasing the mPQ score.

Method	Pre. Data	Top-1 Acc. (%)				mPQ
		BACH	CRC	PCam	MHIST	CoNSeP
Random	-	51.67	68.91	76.52	63.15	22.29
Supervised	IN	80.83	90.93	80.79	76.25	33.49
MoCo v2	IN	71.67	92.86	82.37	79.73	39.13
MoCo v2	TCGA	77.50	93.52	**86.78**	77.07	39.85
MoCo v2	TC+TU	**85.00**	**93.94**	86.53	**82.29**	**41.40**

Table 5. **Varying pre-training datasets under the linear evaluation protocol.** We consider ImageNet (IN), TCGA, and TCGA and TULIP combined (TC+TU) as pre-training datasets. Training with TC+TU results in consistent performance improvements.

ViT experiments. To the best of our knowledge, we are the first to integrate ViT backbones into the HoVer-Net architecture for nuclei instance segmentation. We find that DINO trained on TCGA data out-performs ImageNet-trained weights for $DINO_{p=16}$ by a large margin, showing again the importance of domain-aligned SSL. While $DINO_{p=16}$ does not work well in neither linear nor fine-tuning evaluations, $DINO_{p=8}$ out-performs even CNN-based methods in linear evaluation and performs reasonably well with fine-tuning. Future work may be able to further unlock the power of transformers as a backbone for nuclei instance segmentation.

5.3. Pre-training on Different Datasets

The experiment aims to investigate the impact on the downstream task in accordance with the pre-training data. We select MoCo v2 as it has previously shown robust performance in relation to various domain-specific data [53]. We show our linear evaluation results in Tab. 5 where we compare against MoCo v2 pre-training on ImageNet[2] as well as on the combined TCGA and TULIP data. We note that TCGA-pretraining out-performs supervised/SSL pre-training with ImageNet on BACH, CRC, PCAM, and CoNSeP. When adding TULIP data into the mix, we can use a total of 36K slides and 32.6M patches for pre-training, and we see that this results in the overall best performance. Through these experiments, we conclude that using a domain-aligned dataset such as TCGA is useful, and increasing the scale and diversity of data can further boost performance.

5.4. Fine-tuning with Limited Labeled Data

In the pathology domain, acquiring high-quality annotations require expert-derived labor and exhaustive refinement to maximize consensus across annotators, hindering the establishment of large-scale annotated datasets. Prior findings in computer vision and medical imaging show that SSL pre-trained models are label-efficient under fine-tuning evaluations [2, 11]. To evaluate the label-efficiency of pre-trained models in the pathology domain, we perform similar eval-

[2]We use a publicly available ImageNet pre-trained model for MoCo v2.

Method	CRC (Top-1 Acc.)			CoNSeP (mPQ)		
	1%	10%	100%	10%	30%	100%
ResNet-50						
Supervised	90.28	93.87	92.09	40.01	41.92	49.60
MoCo v2	**91.73**	**95.10**	**96.21**	**42.59**	43.15	**51.75**
SwAV	89.26	92.84	93.31	39.97	42.94	51.16
BT	91.23	92.84	93.23	41.66	**44.10**	51.61
ViT-S						
Supervised$_{p=16}$	93.15	94.76	95.81	18.49	20.92	36.70
DINO$_{p=16}$	94.03	94.92	95.81	23.22	25.85	38.43
DINO$_{p=8}$	**95.03**	**96.27**	**97.13**	**35.53**	**37.82**	**46.70**

Table 6. **Label-efficiency.** Full fine-tuning results when using a limited number of training samples for the CRC and CoNSeP downstream benchmarks.

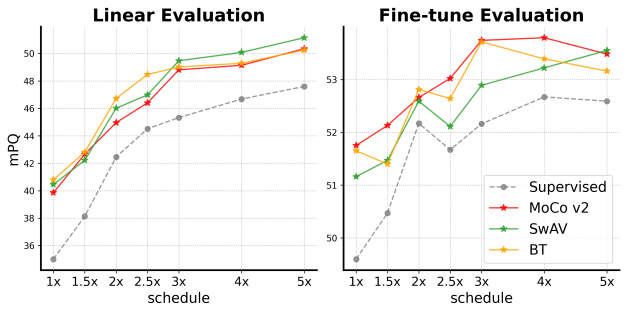

Figure 4. **Different learning schedules for the nuclei instance segmentation task using CoNSeP**. We scale up the learning schedule from 1× (20 epochs) to 5× (100 epochs).

uations and fine-tune our models while varying the labeled data fraction. Following prior works [10,61], subsets of size 1% and 10% are sampled for the image classification dataset. We pick CRC dataset since it has sufficient amounts of data as well as a number of classes to conduct the experiment. On the other hand, the CoNSeP dataset has insufficient data to conduct the same experiment, and therefore, we use subsets of size 10% and 30% for nuclei instance segmentation. Further details can be found in our supplementary materials.

Tab. 6 presents fine-tuning evaluation results with varying amounts of downstream labeled data. Compared to the *Supervised* baselines, SSL pre-trained weights out-perform the ImageNet pre-trained weights. In particular, MoCo v2 and DINO$_{p=8}$ show the best performances for ResNet-50 and ViT-S backbones respectively, maintaining the performance gap to *Supervised* baselines even with increasing amounts of labeled data.

5.5. Longer Learning Schedules

When evaluating SSL methods on downstream dense prediction tasks, it is desirable to show results with longer fine-tuning schedules [27, 45, 58] but this is rarely shown in papers due to the diminishing performance gap between methods when fine-tuning for longer [28]. To better demonstrate the real-world implications of SSL pre-training on pathology data for nuclei instance segmentation, we scale the training schedule from 1× to 5× and show mPQ scores computed on the CoNSeP dataset in Fig. 4. In both linear and fine-tuning evaluations, we clearly observe the benefit of TCGA pre-trained weights compared to ImageNet-supervised weights and note that the performance gap is maintained even at longer training schedules.

	S	V	G	ColorJitter		mPQ
				weak	strong	
Baseline	✓		✓		✓	50.71
	✓	✓	✓		✓	51.03
		✓	✓		✓	51.07
HED-light		✓				50.48
RandStainNA		✓				50.86
RandStainNA$_{GMM}$		✓				50.71
RandStainNA		✓	✓			51.07
		✓	✓	✓		51.13
		✓	✓		✓	50.27
RandStainNA$_{GMM}$		✓	✓			50.99
		✓	✓	✓		**51.61**
		✓	✓		✓	50.51

where **S**: Solarization, **V**: Vertical Flip, **G**: Grayscale

Table 7. **Augmentation ablation study on CoNSeP**. mPQ scores are computed after fine-tuning with a 1× schedule. *Baseline* refers to the original Barlow Twins setting.

5.6. Ablation Study

As described in Sec 3.2, we propose to use data augmentation techniques tailored for pathology data. We empirically assess whether these domain-adapted techniques are beneficial to SSL in pathology through an ablation study. We select Barlow Twins [61] to conduct our experiments and pre-train our models for 200 ImageNet epochs on the TCGA dataset. The fine-tuning evaluation protocol is adopted to better indicate real-world implications.

Augmentation Ablation. We select nuclei instance segmentation as a target task, as it is one of the most practical and challenging tasks in pathology. Starting from the default data augmentation of Barlow Twins [61] denoted as *Baseline* in Tab. 7, we add a random vertical flip augmentation in order to take advantage of the nature wherein pathology images have no canonical orientation. Based on prior work that claims that solarization can produce unrealistic and harmful images for pathology model training [22], we exclude the random solarization augmentation. With these two changes, we observe a gain of 0.36 in mPQ score.

We expect that stain augmentations serve to generate more domain-relevant augmented views, particularly, in terms of color variations. However, stain augmentation clashes with color distortion, as both alter the color statis-

FoV	Top-1 Acc. (%)				mPQ
	BACH	CRC	PCam	MHIST	CoNSeP
20×	79.17	**95.04**	85.24	79.32	50.66
40×	79.17	91.88	80.82	74.21	48.83
20×, 40×	**85.00**	93.23	**86.92**	**81.27**	**51.61**

Table 8. **Magnification ablation study.** Fine-tuning performance when using different FoVs during pre-training.

tics of images. Thus, we begin by disabling color distortion and then compare key stain augmentation methods first. We find that RandStainNA [47] and RandStainNA$_{GMM}$ outperform HED-light [50], confirming insights from supervised image classification [47].

Next, we bring back the components of color distortion (consisting of grayscale and color jitter) and evaluate them in detail. We find that random grayscale augmentation is surprisingly effective, given that the produced images may not be considered plausible in pathology. As the standard strength of color jittering produces highly unrealistic outputs, we evaluate a weaker jitter strength as well. Indeed, we find that while the performance drops substantially when using strong color jitter, using weak color jitter together with random grayscale results in the best performances. In particular, RandStainNA$_{GMM}$ shows high performance, motivating us to adopt it in our main experiments.

Through these observations, we substantiate our claim that existing augmentation schemes designed for SSL using natural images are sub-optimal for pathology data, necessitating pathology-specific augmentation schemes when training on pathology data such as TCGA and TULIP.

Magnification Ablation. In Sec. 3.2, we argue that pre-training using multiple magnifications or FoVs is beneficial as downstream pathology tasks occur at various magnifications. We do find experimentally that using multiple FoVs in the pre-training dataset is beneficial for overall downstream task performance (see Tab. 8).

Interestingly, we observe that using both 20× and 40× is best, while using only 20× is typically second-best. This is the case even for datasets such as PCam, MHIST, and CoNSeP which consist of images collected at approximately 40×. We hypothesize that the use of multiple magnifications is not valuable due to the matching of magnifications between upstream and downstream training, but rather due to the diversity of image appearances. Specifically, 20× images, due to the wider field-of-view, are visually and texture-wise more diverse than 40× images. Combining the two magnifications results in an even more diverse set of images. The more diverse data also results in better convergence during pre-training (see supplementary materials).

6. Discussion

In this section, we answer a few key questions that computational pathology researchers may naturally ask when considering self-supervised pre-training for their research.

Should we pre-train on pathology data? Yes – We have consistently demonstrated that pre-training on pathology data out-performs supervised pre-training on ImageNet by performing comprehensive experiments on many SSL methods and datasets. Interestingly, SSL pre-trained weights can maintain the performance gap on CoNSeP even for longer training schedules. Our experiments demystify and confirm the effectiveness of domain-aligned SSL pre-training on the pathology domain.

Which SSL method is best? We find that there is *no clear winner*. All SSL methods applied with domain-aligned pre-training generally perform well. Thus, instead of focusing on selecting a specific SSL method, we recommend that practitioners focus on curating large-scale domain-aligned datasets for SSL pre-training. Yet, some initial observations may be useful to future research. For example, (a) Barlow Twins tends to perform well in linear evaluations and MoCo v2 in fine-tuning evaluations, and (b) ViTs benefit more from domain-aligned SSL compared to CNNs.

What is a key ingredient for successful self-supervised pre-training? Domain knowledge – our proposed set of techniques are fully based on observations in pathology, and are experimentally shown to be effective. By incorporating domain-specific knowledge into the pre-training step, e.g., using stain augmentation and extracting patches from multiple FoVs, we go beyond the performance one can get from naively applying SSL to a new dataset.

7. Conclusion and Future Work

In this paper, we conduct the largest and most comprehensive study of SSL in the pathology domain, to date, using up to 33 million image patches during pre-training and evaluating 4 representative SSL methods (both CNNs and ViTs) on 2 downstream tasks and 5 datasets. Our study confirms that large-scale domain-aligned pre-training is helpful for pathology, showing its value in scenarios with limited labeled data, longer fine-tuning schedules, and when using larger and more diverse datasets for pre-training (such as TCGA + TULIP). Furthermore, we propose a set of techniques that are carefully designed by leveraging pathology-specific knowledge, and integrate them into the self-supervised pre-training stage, resulting in performance improvements. We believe that further exploration of domain-specific augmentation strategies will yield improved techniques for pathology-specific SSL in the future.

References

[1] Guilherme Aresta, Teresa Araújo, Scotty Kwok, Sai Saketh Chennamsetty, Mohammed Safwan, Varghese Alex, Bahram Marami, Marcel Prastawa, Monica Chan, Michael Donovan, et al. Bach: Grand challenge on breast cancer histology images. *Medical Image Analysis*, 56:122–139, 2019. 2, 4

[2] 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 *ICCV*, pages 3478–3488, 2021. 2, 6

[3] Joseph Boyd, Mykola Liashuha, Eric Deutsch, Nikos Paragios, Stergios Christodoulidis, and Maria Vakalopoulou. Self-supervised representation learning using visual field expansion on digital pathology. In *ICCV Workshops*, pages 639–647, 2021. 2

[4] Wouter Bulten, Kimmo Kartasalo, Po-Hsuan Cameron Chen, Peter Ström, Hans Pinckaers, Kunal Nagpal, Yuannan Cai, David F Steiner, Hester van Boven, Robert Vink, et al. Artificial intelligence for diagnosis and gleason grading of prostate cancer: the panda challenge. *Nature medicine*, 28(1):154–163, 2022. 1, 3

[5] 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

[6] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *ECCV*, pages 132–149, 2018. 2

[7] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *NeurIPS*, 33:9912–9924, 2020. 2, 4

[8] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, pages 9650–9660, 2021. 1, 2, 4, 6

[9] Richard J Chen, Chengkuan Chen, Yicong Li, Tiffany Y Chen, Andrew D Trister, Rahul G Krishnan, and Faisal Mahmood. Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In *CVPR*, pages 16144–16155, 2022. 2

[10] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, pages 1597–1607. PMLR, 2020. 1, 2, 4, 5, 7

[11] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E Hinton. Big self-supervised models are strong semi-supervised learners. *NeurIPS*, 33:22243–22255, 2020. 6

[12] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 2, 4

[13] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *CVPR*, pages 15750–15758, 2021. 2, 4

[14] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *ICCV*, pages 9640–9649, 2021. 2

[15] Sangjoon Choi, Soo Ick Cho, Minuk Ma, Seonwook Park, Sergio Pereira, Brian Jaehong Aum, Seunghwan Shin, Kyunghyun Paeng, Donggeun Yoo, Wonkyung Jung, et al. Artificial intelligence–powered programmed death ligand 1 analyser reduces interobserver variation in tumour proportion score for non–small cell lung cancer with better prediction of immunotherapy response. *European Journal of Cancer*, 170:17–26, 2022. 1

[16] Ozan Ciga, Tony Xu, and Anne Louise Martel. Self supervised contrastive learning for digital histopathology. *Machine Learning with Applications*, 7:100198, 2022. 2, 4

[17] James Coughlan and Alan L Yuille. The manhattan world assumption: Regularities in scene statistics which enable bayesian inference. In *NeurIPS*, volume 13, 2000. 3

[18] Angel Cruz-Roa, Hannah Gilmore, Ajay Basavanhally, Michael Feldman, Shridar Ganesan, Natalie NC Shih, John Tomaszewski, Fabio A González, and Anant Madabhushi. Accurate and reproducible invasive breast cancer detection in whole-slide images: A deep learning approach for quantifying tumor extent. *Scientific reports*, 7(1):1–14, 2017. 4

[19] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. IEEE, 2009. 1, 3

[20] James A Diao, Jason K Wang, Wan Fung Chui, Victoria Mountain, Sai Chowdary Gullapally, Ramprakash Srinivasan, Richard N Mitchell, Benjamin Glass, Sara Hoffman, Sudha K Rao, et al. Human-interpretable image features derived from densely mapped cancer pathology slides predict diverse molecular phenotypes. *Nature communications*, 12(1):1–15, 2021. 1, 4

[21] 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 *ICLR*, 2020. 2, 4

[22] Khrystyna Faryna, Jeroen van der Laak, and Geert Litjens. Tailoring automated data augmentation to h&e-stained histopathology. In *Medical Imaging with Deep Learning*, 2021. 7

[23] Jacob Gildenblat and Eldad Klaiman. Self-supervised similarity learning for digital pathology. In *MICCAI Workshops*, 2019. 2

[24] Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017. 4

[25] Simon Graham, Quoc Dang Vu, Shan E Ahmed Raza, Ayesha Azam, Yee Wah Tsang, Jin Tae Kwak, and Nasir Rajpoot. Hover-net: Simultaneous segmentation and classification of

nuclei in multi-tissue histology images. *Medical Image Analysis*, 58:101563, 2019. 2, 3, 4, 5

[26] 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. *NeurIPS*, 33:21271–21284, 2020. 1, 2, 4, 5

[27] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, pages 9729–9738, 2020. 2, 7

[28] Kaiming He, Ross Girshick, and Piotr Dollár. Rethinking imagenet pre-training. In *ICCV*, pages 4918–4927, 2019. 7

[29] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 4

[30] Vladimir Iglovikov and Alexey Shvets. Ternausnet: U-net with vgg11 encoder pre-trained on imagenet for image segmentation. *arXiv preprint arXiv:1801.05746*, 2018. 1

[31] Jakob Nikolas Kather, Niels Halama, and Alexander Marx. 100,000 histological images of human colorectal cancer and healthy tissue, Apr. 2018. 2, 4

[32] Alexander Ke, William Ellsworth, Oishi Banerjee, Andrew Y Ng, and Pranav Rajpurkar. Chextransfer: performance and parameter efficiency of imagenet models for chest x-ray interpretation. In *Proceedings of the Conference on Health, Inference, and Learning*, pages 116–124, 2021. 1

[33] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. Panoptic segmentation. In *CVPR*, pages 9404–9413, 2019. 5

[34] Simon Kornblith, Jonathon Shlens, and Quoc V Le. Do better imagenet models transfer better? In *CVPR*, pages 2661–2671, 2019. 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 *CVPR*, pages 14318–14328, 2021. 2

[36] Chunyuan Li, Jianwei Yang, Pengchuan Zhang, Mei Gao, Bin Xiao, Xiyang Dai, Lu Yuan, and Jianfeng Gao. Efficient self-supervised vision transformers for representation learning. *arXiv preprint arXiv:2106.09785*, 2021. 2

[37] Geert Litjens, Peter Bandi, Babak Ehteshami Bejnordi, Oscar Geessink, Maschenka Balkenhol, Peter Bult, Altuna Halilovic, Meyke Hermsen, Rob van de Loo, Rob Vogels, et al. 1399 h&e-stained sentinel lymph node sections of breast cancer patients: the camelyon dataset. *GigaScience*, 7(6):giy065, 2018. 1

[38] John Lonsdale, Jeffrey Thomas, Mike Salvatore, Rebecca Phillips, Edmund Lo, Saboor Shad, Richard Hasz, Gary Walters, Fernando Garcia, Nancy Young, et al. The genotype-tissue expression (gtex) project. *Nature genetics*, 45(6):580–585, 2013. 1

[39] Christos Matsoukas, Johan Fredin Haslum, Moein Sorkhei, Magnus Söderberg, and Kevin Smith. What makes transfer learning work for medical images: Feature reuse & other factors. In *CVPR*, pages 9225–9234, 2022. 1, 2

[40] Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. In *CVPR*, pages 6707–6717, 2020. 2

[41] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. In *NeurIPS*, 2018. 2

[42] Sehhoon Park, Chan-Young Ock, Hyojin Kim, Sergio Pereira, Seonwook Park, Minuk Ma, Sangjoon Choi, Seokhwi Kim, Seunghwan Shin, Brian Jaehong Aum, et al. Artificial intelligence–powered spatial analysis of tumor-infiltrating lymphocytes as complementary biomarker for immune checkpoint inhibition in non–small-cell lung cancer. *Journal of Clinical Oncology*, 40(17):1916, 2022. 1, 4

[43] Maithra Raghu, Chiyuan Zhang, Jon Kleinberg, and Samy Bengio. Transfusion: Understanding transfer learning for medical imaging. *NeurIPS*, 32, 2019. 1

[44] Erik Reinhard, Michael Adhikhmin, Bruce Gooch, and Peter Shirley. Color transfer between images. *IEEE Computer graphics and applications*, 21(5):34–41, 2001. 3

[45] Byungseok Roh, Wuhyun Shin, Ildoo Kim, and Sungwoong Kim. Spatially consistent representation learning. In *CVPR*, pages 1144–1153, 2021. 7

[46] Arnout C Ruifrok, Dennis A Johnston, et al. Quantification of histochemical staining by color deconvolution. *Analytical and quantitative cytology and histology*, 23(4):291–299, 2001. 3

[47] Yiqing Shen, Yulin Luo, Dinggang Shen, and Jing Ke. Randstainna: Learning stain-agnostic features from histology slides by bridging stain augmentation and normalization. In *MICCAI*, pages 212–221. Springer, 2022. 3, 4, 8

[48] Hari Sowrirajan, Jingbo Yang, Andrew Y Ng, and Pranav Rajpurkar. Moco pretraining improves representation and transferability of chest x-ray models. In *Medical Imaging with Deep Learning*, pages 728–744. PMLR, 2021. 2

[49] Fabio A Spanhol, Luiz S Oliveira, Caroline Petitjean, and Laurent Heutte. A dataset for breast cancer histopathological image classification. *Ieee transactions on biomedical engineering*, 63(7):1455–1462, 2015. 1

[50] 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. 3, 4, 8

[51] Yonglong Tian, Olivier J Henaff, and Aäron van den Oord. Divide and contrast: Self-supervised learning from uncurated data. In *ICCV*, pages 10063–10074, 2021. 4

[52] Jeroen Van der Laak, Geert Litjens, and Francesco Ciompi. Deep learning in histopathology: the path to the clinic. *Nature medicine*, 27(5):775–784, 2021. 1

[53] Wouter Van Gansbeke, Simon Vandenhende, Stamatios Georgoulis, and Luc V Gool. Revisiting contrastive methods for unsupervised learning of visual representations. *NeurIPS*, 34:16238–16250, 2021. 6

[54] Bastiaan S Veeling, Jasper Linmans, Jim Winkens, Taco Cohen, and Max Welling. Rotation equivariant cnns for digital pathology. In *MICCAI*, pages 210–218. Springer, 2018. 2, 4

[55] Xiyue Wang, Sen Yang, Jun Zhang, Minghui Wang, Jing Zhang, Junzhou Huang, Wei Yang, and Xiao Han. Transpath: Transformer-based self-supervised learning for histopathological image classification. In *MICCAI*, pages 186–195. Springer, 2021. 2

[56] Jerry Wei, Arief Suriawinata, Bing Ren, Xiaoying Liu, Mikhail Lisovsky, Louis Vaickus, Charles Brown, Michael Baker, Naofumi Tomita, Lorenzo Torresani, et al. A petri dish for histopathology image analysis. In *International Conference on Artificial Intelligence in Medicine*, pages 11–24. Springer, 2021. 2, 4

[57] John N Weinstein, Eric A Collisson, Gordon B Mills, Kenna R Shaw, Brad A Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, and Joshua M Stuart. The cancer genome atlas pan-cancer analysis project. *Nature genetics*, 45(10):1113–1120, 2013. 1, 2

[58] Tete Xiao, Colorado J Reed, Xiaolong Wang, Kurt Keutzer, and Trevor Darrell. Region similarity representation learning. In *ICCV*, pages 10539–10548, 2021. 7

[59] Yiting Xie and David Richmond. Pre-training on grayscale imagenet improves medical image classification. In *ECCV Workshops*, pages 0–0, 2018. 1

[60] Jiawei Yang, Hanbo Chen, Yuan Liang, Junzhou Huang, Lei He, and Jianhua Yao. Concl: Concept contrastive learning for dense prediction pre-training in pathology images. In *ECCV*, pages 523–539, 2022. 2

[61] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *ICML*, pages 12310–12320. PMLR, 2021. 2, 4, 5, 7