Learning Representations with Contrastive Self-Supervised Learning for Histopathology Applications

Karin Stacke   karin.stacke@liu.se
Department of Science and Technology (ITN), Linköping University, Sweden
Sectra AB, Linköping

Jonas Unger   jonas.unger@liu.se
Department of Science and Technology (ITN), Linköping University, Sweden
Center for Medical Image Science and Visualization (CMIV), Linköping University, Sweden

Claes Lundström   claes.lundstrom@liu.se
Department of Science and Technology (ITN), Linköping University, Sweden
Center for Medical Image Science and Visualization (CMIV), Linköping University, Sweden
Sectra AB, Linköping

Gabriel Eilertsen   gabriel.eilertsen@liu.se
Department of Science and Technology (ITN), Linköping University, Sweden
Center for Medical Image Science and Visualization (CMIV), Linköping University, Sweden

Abstract

Unsupervised learning has made substantial progress over the last few years, especially by means of contrastive self-supervised learning. The dominating dataset for benchmarking self-supervised learning has been ImageNet, for which recent methods are approaching the performance achieved by fully supervised training. The ImageNet dataset is however largely object-centric, and it is not clear yet what potential those methods have on widely different datasets and tasks that are not object-centric, such as in digital pathology. While self-supervised learning has started to be explored within this area with encouraging results, there is reason to look closer at how this setting differs from natural images and ImageNet. In this paper we make an in-depth analysis of contrastive learning for histopathology, pin-pointing how the contrastive objective will behave differently due to the characteristics of histopathology data. Using SimCLR and H&E stained images as a representative setting for contrastive self-supervised learning in histopathology, we bring forward a number of considerations, such as view generation for the contrastive objective and hyper-parameter tuning. In a large battery of experiments, we analyze how the down-stream performance in tissue classification will be affected by these considerations. The results point to how contrastive learning can reduce the annotation effort within digital pathology, but that the specific dataset characteristics need to be considered. To take full advantage of the contrastive learning objective, different calibrations of view generation and hyper-parameters are required. Our results pave the way for realizing the full potential of self-supervised learning for histopathology applications. Code and trained models are available at `https://github.com/k-stacke/ssl-pathology`.

Keywords: Deep Learning, Histopathology, Self-Supervised Learning, Transfer learning, Contrastive Learning, H&E, Whole-slide image analysis

1. Introduction

Deep learning has in the last decade shown great potential for medical image-analysis applications (Litjens et al., 2017). However, the transition from research results to clinically deployed applications is slow. One of the main bottlenecks is the lack of high-quality labeled data needed for training models with high accuracy and robustness, where annotations are cumbersome to acquire and relies on medical expertise (Stadler et al., 2021). An active research field has therefore been focused on reducing the dependency of labeled data. This can, for example, be accomplished through *transfer learning* (Yosinski et al., 2014), where training on a widely different dataset, such as ImageNet (Deng et al., 2009), can reduce the amount of training data needed in a targeted downstream medical imaging application (Truong et al., 2021). However, as ImageNet contains natural images, there is a large discrepancy between the source and target domains in terms of colors, intensities, contrasts, image features, class distribution, etc. It has been shown that a closer resemblance between the source and target datasets is preferable (Cui et al., 2018; Cole et al., 2021; Li et al., 2020), and the usefulness of ImageNet for pre-training in medical imaging has been questioned (Raghu et al., 2019). In addition, ImageNet pre-training has recently also been questioned because of the biased nature of the dataset, intensifying the need to find alternative pre-training methods that are better tailored for the target application (Birhane and Prabhu, 2021; Yang et al., 2021a).

Self-supervised learning (SSL) has recently emerged as a viable technique for creating pre-trained models without the need for large, annotated datasets. Instead, pre-training is performed on unlabeled data by means of a proxy objective, for which the labels can be automatically generated. The objective is formulated such that the model learns a general understanding of image content. It can include predicting image rotation (Gidaris et al., 2018), solving jigsaw puzzles (Noroozi and Favaro, 2016), re-coloring gray-scale images (Larsson et al., 2017), to mention a few. One family of SSL methods are those based on a contrastive training objective, which is formulated to contract the representation of two positive views, while simultaneously distracting the representations of negative views. This training strategy has shown great promise in the last few years, and include methods such as CPC (van den Oord et al., 2019), SimCLR (Chen et al., 2020a), CMC (Tian et al., 2020a), and MoCo (He et al., 2020). However, successful results have primarily been presented on ImageNet, despite the above-mentioned need for moving away from this dataset, and it is still unclear how well the results generalize to datasets with different characteristics.

In this paper, we investigate how SimCLR (Chen et al., 2020a) can be extended to learn representations for histopathology applications. We consider H&E stained images, and take a holistic approach, comparing how differences between the ImageNet dataset and histopathology data influence the SSL objective, as well as pin-pointing how the different components of the objective contribute to the learning outcome. We show that the heuristics that have been demonstrated to work well for natural images do not necessarily apply in histopathology scenarios. To explore the differences, we setup a rigorous experimental study, which includes:

- three different pathology datasets, for investigating different scenarios with in-domain and out-of-domain histology data,

- comparison between ImageNet pre-training and domain-specific SSL, for evaluating the added value of unsupervised pre-training on histology data,

- evaluation of different data availability scenarios in training of the target application, as we expect pre-training to add different value in the different scenarios,

- evaluation of the impact of different hyper-parameters, as these have been demonstrated important for the success of contrastive SSL,

- a study of the convergence behavior of SimCLR on pathology data, for demonstrating how well contrastive pre-training contribute to the target downstream application of tissue classification.

Our main motive is to clarify how contrastive SSL for histopathology cannot be considered under the same assumptions as for natural image data. Our results lead to a number of important conclusions. For example, we show that:

- In pathology, SimCLR pre-training gives substantial benefits, if used correctly.

- Different types of positive/negative views are optimal for contrastive SSL in histopathology compared to natural images, and the optimal views can be dataset dependent even within the pathology domain.

- Parameter tuning, such as the batch size used for SSL, does not have the same influence as for natural images, due to the differences in data characteristics.

- Pre-training data aligned with the target pathology sub-domain is better suited compared to more diverse pathology data.

The paper is organized as follows: In Section 2, we discuss and position our work in relation to previous work on SSL, contrastive SSL, and the aspects of SSL specific to digital pathology. In Section 3, we briefly explain contrastive SSL, followed by a thorough discussion around different aspects related to the view generation that we believe to be important when comparing natural images to digital pathology. This discussion sets the stage for the experimental study performed in Section 4, with results presented in Section 5. Finally, in Section 6, we conclude with an outlook on what needs to be considered for further improving contrastive SSL in histopathology, where we emphasize how the differences in data characteristics require a significantly different approach to formulating the contrastive learning objective. We believe that this work is important for broadening the understanding of self-supervised methods, and how the intrinsic properties of the data affect the representations. Our hope is that this will be a stepping stone towards pre-trained models better tailored for histopathology applications.

2. Related Work

A large body of literature has been devoted to unsupervised and self-supervised learning. For self-supervised learning, multiple creative methods have been presented for defining proxy objectives and performing self-labelling. These include, but are not limited to, colorization of grayscale images (Zhang et al., 2016; Larsson et al., 2017), solving of jigsaw puzzles (Noroozi and Favaro, 2016), and prediction of rotation (Gidaris et al., 2018).

Within self-supervised learning, significant attention has recently been given to a specific family of methods, which performs instance discrimination (Dosovitskiy et al., 2014; Wu et al., 2018) through contrastive learning with multiple views. Bachman et al. (2019) presented a contrastive self-supervised method (AMDIM) based on creating multiple views using augmentation. van den Oord et al. (2019) presented the InfoMax objective, and showed that by minimizing it you can maximize the mutual information between views. Building on these works, constrastive self-supervised methods such as CMC (Tian et al., 2020a), MoCo(v2) (He et al., 2020; Chen et al., 2020b), and SimCLR (Chen et al., 2020a) have recently shown improved results on ImageNet benchmarks, closing the gap between supervised and unsupervised training. Falcon and Cho (2020) actually showed that many of these methods (such as AMDIM, CPC and SimCLR) are special cases of a general framework for contrastive SSL. As a continuation of this development, methods such as BYOL (Grill et al., 2020) and SwAV (Caron et al., 2020) have been presented, expanding the concept to either avoiding contrastive negatives or to doing cluster assignments instead of instance discrimination. In this work, we use SimCLR as a representative method of the contrastive self-supervised methodology. The choice of method is motivated by its simplicity and high level of adoption since its introduction. We believe that the integral components of contrastive learning can be justly studied through SimCLR, and anticipate that the results will generalize to other state-of-the-art contrastive SSL methods.

Self-supervised methods have also been applied to medical images in general, and histopathology in specific. A number of methods for self-supervised learning have been presented for data such as volumetric CT and MRI, X-ray images and dermatological digital images, with application within classification, localization and segmentation (Liu et al., 2019; Yan et al., 2020; Chaitanya et al., 2020; Xie et al., 2020; Zhou et al., 2020; Azizi et al., 2021; Li et al., 2021a; Sowrirajan et al., 2021; You et al., 2021). Many of the works shows that in-domain pre-training is superior to ImageNet pre-training, and that domain-specific selections of positive and negatives views boosts performance. This motivates us to further understand what (if any) considerations that are needed for the domain of histopathology.

Self-supervised methods developed for histopathology has been presented. For example, incorporation of the spatial information of patches (Gildenblat and Klaiman, 2020; Li et al., 2021b), using augmentations based on stain separation (Yang et al., 2021b), using transformer architectures to capture global information (Wang et al., 2021), or utilizing the multi-resolution structure of whole-slide images (Koohbanani et al., 2020; Srinidhi et al., 2022). Furthermore, a number of previous work have evaluated contrastive SSL methods that were designed for natural images (Lu et al., 2019; Stacke et al., 2020; Dehaene et al., 2020; Ciga et al., 2022), all showing promising results. Among these, Ciga et al. (2022) are the one closes to this work, as SimCLR is the method used in both studies. Complimentary to Ciga et al. (2022)'s large battery of experiments, we deepen the understanding through experiments regarding optimal view generation and hyper-parameters, in junction with a in-depth discussion on the unique characteristics of histopathology datasets that impact the learned representations.

Some works give a more rigorous theoretical background to the contrastive methods, such as Arora et al. (2019), Tsai et al. (2021), Wu et al. (2020) and Tschannen et al. (2020). However, much of the success of the previously mentioned methods is derived from heuristics that are still left to be explained theoretically. It is not clear how well the performance

4

showed on one domain transfers to new and different ones. As Torralba and Efros (2011) pointed out some time ago, all datasets, ImageNet included, encompasses specific biases that may be inherited by a model trained on the data. Purushwalkam and Gupta (2020), for example, argued that the object-centric nature of ImageNet is the reason for why SSL methods based on heavy scale augmentations perform well, but that this approach does not work for object recognition tasks. Cole et al. (2021) showed that contrastive learning is less suited for tasks requiring more fine-grained details, and that pre-training on out-of-domain data gives little benefit. Therefore, we have reason to look closely on how contrastive learning methods transfer to the domain of histopathology.

3. Background

This section gives an overview of contrastive multi-view learning as well as view generation, with the goal of giving an conceptual description of how different design choices affect the learned representation. The descriptions will facilitate the experimental design and result analysis in Section 4-6, for identifying the differences in contrastive SSL for histopathology compared to object-centered datasets with natural images.

3.1 Contrastive learning

The general idea of contrastive learning is that an *anchor* data point x_i (sometimes referred to as query) together with a *positive* data point x_j (key) form a set of positive *views* of the same object. The goal is to map these views to a shared representation space, such that the representations contain underlying information (features) shared between them, while simultaneously discarding other (nuisance) information. A positive view pair could therefore share the information of depicting the same object, but may differ in view angle, lighting conditions, or occlusion level.

For images, this shared information is high-dimensional, which makes its estimation challenging. The views are therefore encoded to a more compact representation using a non-linear encoder, g, $\{z_i = g(x_i), z_j = g(x_j)\}$, such that the mutual information between z_i and z_j is maximized. Maximization of the mutual information can be estimated by using a contrastive loss function (van den Oord et al., 2019), defined to be minimized by assigning high values of positive pairs of data ($\{z_i, z_j\}$) and low values to all other ($\{z_i, z_k\}, k \neq \{j, i\}$ (denoted "negatives"). A popular such loss function is the InfoNCE loss, defined as:

$$\mathcal{L}_i = -\log \frac{\exp\left(z_i^T z_j\right)}{\sum_k \exp\left(z_i^T z_k\right)}. \tag{1}$$

The choice of positives can be done either in a supervised way, where coupled data is collected (such as multiple staining of the same tissue sample), or in a self-supervised manner, where the views are automatically generated. One popular approach of the latter kind is to create two views from the same data point, x, by applying random transformations, $\{t_i \sim T, t_j \sim T, \}$, such that two views of the data sample are created, $\{\widetilde{x}_i, \widetilde{x}_j\}$. Negative samples are typically taken as randomly selected samples from the training data. *View generation*, that is, how the views are chosen, has a direct impact of what features the model will learn.

3.2 View Generation

Due to the way the contrastive objective is formulated, the choices of how positive and negative views are selected will largely impact the learned representation. If they are chosen correctly, the learned representation will separate data in such a way that is useful for the target *downstream* task. Incorrectly chosen, the model may learn *nuisance* features that result in poor separation of the data with respect to the downstream task. The choice of optimally selecting positives and negatives thus depends on the intended downstream task, since it needs to take into account what is considered *task relevant* and *task irrelevant* [1] (Tian et al., 2020b). For example, color may be considered a nuisance variable in the downstream task of tumor classification in H&E slides and should therefore not be shared between positives, but may be an important feature if the downstream task is scoring of immunohistochemical staining. As the SSL is task-agnostic, that is, the downstream task is unknown with regard to the self-supervised objective, the view generation is critical for controlling what features the model learns, such that they tailored to the downstream task.

Figure 1 shows the relationship between shared mutual information (SSL objective) and view generation. Positive views (top) can be selected/created such that 1) no task-relevant features are shared, resulting in that the model will use only task-irrelevant features to solve the self-supervised objective, 2) only task-relevant features are shared, resulting in the model learning (a subset) of these, or 3) both task-relevant and irrelevant features are shared, increasing the risk of the model learning so called shortcut features (Geirhos et al., 2020), often low-level features (such as color). If the views are created with augmentations (which is what we will consider in this work), this is the result of 1) too strong, 2) just right or 3) too little transformations. To achieve optimal performance on the downstream task, the model should learn the minimally sufficient solution (Tsai et al., 2021), such that two positive views share as much task-relevant information as possible, and as little task-irrelevant information (middle column, Figure 1).

As highlighted by Arora et al. (2019), the choice of negatives (which generally are randomly selected from the mini-batch) is also important for the learning outcome. In the bottom row of Figure 1, the relationship between an anchor and negative is shown. If no information is shared, the model does not have to learn any task-relevant features as any feature may solve the pre-training objective (left). If too much information is shared, the model will not learn task-relevant features as these cannot be used to distinguish between positive and negatives (right). This is typically the case when negatives belong to the same (latent) class as the positive, making them so called *false negatives*. It is important to distinguish between *false* negatives and *hard* negatives, where hard negatives are true negatives which share similar features with the anchor. This is shown in the middle column, where the shared information between target and negative is composed of substantial amount of nuisance information, but no task related information. Hard negatives are generally beneficial for the learning outcome (Robinson et al., 2021), as they hinder the model to rely on task-irrelevant features to solve the contrastive objective.

To further illustrate the relationship between the different views and the self-supervised objective, an example is shown in Figure 2. In this example, view generation resulted in some features shared between the anchor and positive views, of which a subset are task-

---

1. the notation "task irrelevant" and "nuisance" features will be used interchangeably
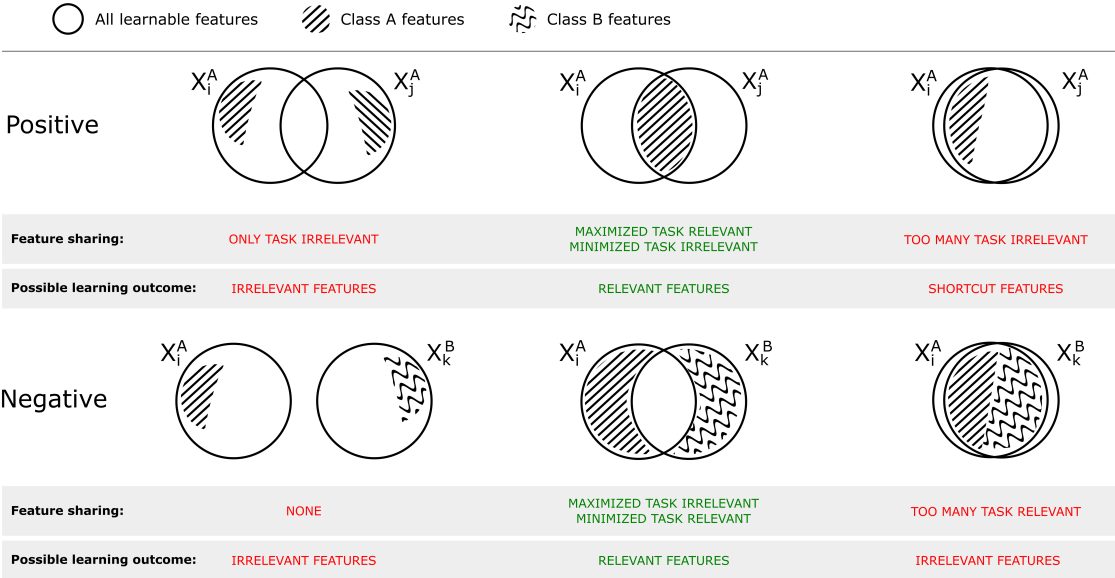
View Generation

Figure 1: How view generation is done will affect the learning outcome. Circles represent all learnable features for one data sample, where shaded areas denote task-relevant features. Between anchor (x_i^A) and positive (x_j^A), which belong to the same class (A), sharing of task-relevant features should be maximized, while simultaneously minimizing task-irrelevant features. Between the anchor (x_i^A) and the negative (x_k^B) (not belonging to same class, $A \neq B$), the opposite should be true.

relevant. Some task-relevant features are, however, not shared, existing in only one of the views. In addition, some features (both task-relevant and irrelevant) are also shared by a negative view. This means that out of all available features, some exist in one, two or all of the views. During model training, the model learns to represent each data point such that the contrastive objective is fulfilled: the anchor-positive views are attracted, and the anchor-negative views are repelled. The *attracting features* are found in the intersection between the anchor and positive but not in the negative. The *repelling features* are found in the negative, but not in the anchor. As discussed in the previous section, the region of attracting features should therefore contain primarily task-relevant features, and the repelling region should only contain task-irrelevant features. It should, however, be noted that there is no guarantee that the model will learn all features in these regions, but only the subset of features that is enough to solve the contrastive objective, as observed by Tian et al. (2020b). In the end, as the contrastive objective is task-agnostic and completely relies on this distinction of features over the training dataset, the degree of task relevance of the learned representation will depend on how the views were generated.
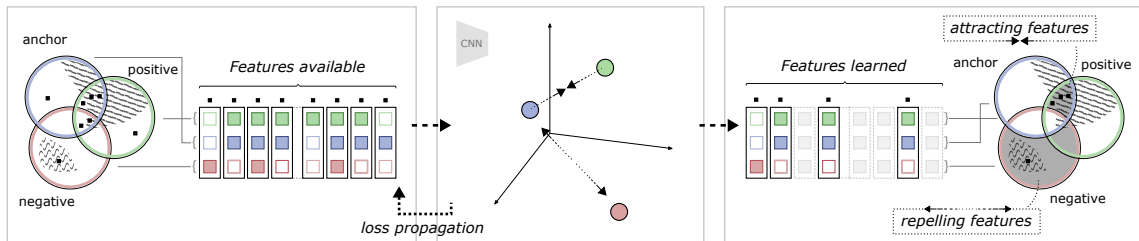
Figure 2: View generation results in shared features between anchor, positive and negative views. Due to the contrastive objective, only a subset of all available features are learned by the model. These features are either *attracting* features or *repelling* features, where view generation is the only means of controlling to what extent these features are task-relevant.

4. Method

The goal of this paper is to better understand how contrastive self-supervised learning (SSL) can be used for clinical applications where labeled data is scarce. We do this by evaluating different pre-training methods for the target downstream application, i.e., classification with varying amounts of labeled data. In doing so, it is necessary to systematically analyze and understand the impact of different pre-training methods and training strategies, and how this relates to the type of data used and the target task. In this section, we present the datasets, training details and evaluation metrics.

4.1 Experiment design

By using SimCLR as a representative method for contrastive learning, we build our investigation on a series of experiments where we vary the parameters and methodologies relevant to the analysis.

- *Training SimCLR models on in-domain and out-of-domain histology data.* It is of key interest to understand how pre-training from different histopathology sources and with different augmentations, affect the resulting learned outcome. The results from this analysis will form the basis of our discussion on self-supervised learning for histopathology data.

- *Compare domain-specific SimCLR with ImageNet pre-training and no pre-training.* Previous works often rely on transfer learning from pre-training using ImageNet. Motivated by the strong differences between pathology data and ImageNet data in terms of e.g., image content, number of classes, and overall composition, a systematic evaluation is done of whether a domain-specific pre-training using SimCLR is more beneficial in this context and if so why.

- *Evaluation of different amounts of supervised data for downstream-task training.* Pre-trained models are evaluated both with respect to linear and fine-tuning performance with varying amounts of supervised training data.

- *Batch size, learning rate and temperature scaling impact.* Tuning of hyper-parameters such as batch size, learning rate, and temperature scaling have been shown to play an

(a) ImageNet   (b) Camelyon16   (c) AIDA-LNSK   (d) Multidata
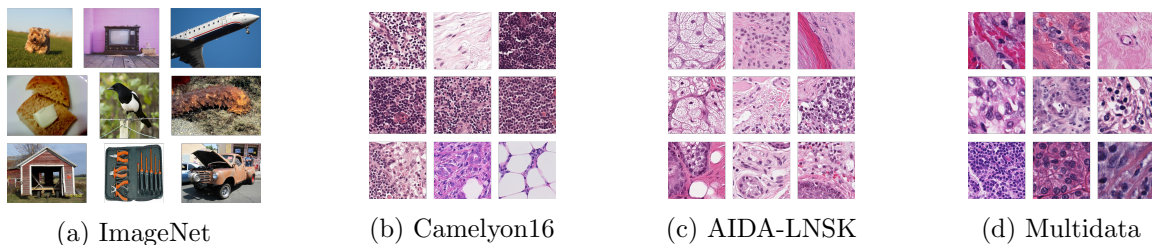
Figure 3: Example images from datasets.

important role in contrastive learning using ImageNet. This experiment explores the corresponding parameter tuning for histopathology data.

- *Evaluation of performance during training.* Training dynamics presents important information on model robustness, the optimization, and overall performance. This experiment investigates downstream task training and how the resulting models evolve over time.

These experiments, conducted with multiple datasets, form the basis for a detailed evaluation of contrastive self-supervised learning in general, and SimCLR in particular, in the context of histopathology.

4.2 Datasets

For this study, three different histopathology datasets were used, from sentinel breast lymph node tissue, skin tissue and one consisting of mixed tissues extracted as a subset from 60 different datasets. As reference dataset, ImageNet is used. Examples images are shown in Figure 3, with further details given below, and in Appendix A.

ImageNet ILSVRC2012   (Deng et al., 2009): a dataset constructed by searching on the internet using keywords listed in the WordNet database. A subset of the total dataset, approximately 1.2M images, was labeled as belonging to one of 1000 classes using a crowd sourcing technique. Despite the aim of being a representative dataset with a wide category of objects, the nature of the collection technique and annotation strategy has resulted in distinct characteristics and biases in the data, resulting in models trained on this dataset may inherit the biases (Torralba and Efros, 2011). SSL methods developed and tested on this dataset are therefore also likely to adhere to some inherit characteristics of the data (Cole et al., 2021). Using ImageNet pre-trained weights for transfer learning is a common approach for many medical image applications, which motivates us to use it as a baseline method. Pre-trained models (trained supervised) were accessed though the Pytorch library [2].

Camelyon16   (Litjens et al., 2018): 399 H&E-stained whole-slide images (WSIs) of sentinel lymph node tissue, annotated for breast cancer metastases. This dataset was sampled into smaller patches twice, to construct one dataset used for self-supervised training, and

---

2. Accessible here: `https://pytorch.org/vision/stable/models.html`

one for supervised training. For unsupervised training, the WSIs were sampled in an unsupervised way, i.e., no tissue annotations were used to guide sampling from the 270 slides selected as the training slides from the official split. Patches were sampled non-overlapping with patch size of 256x256 pixels with a resolution of 0.5 microns per pixel (mpp) (approximately 20x). Maximum 1000 samples were chosen per slide, resulting in a dataset consisted of slightly less than 270k images.

For supervised training, a downstream task was formulated as binary tumor classification task. For this dataset, which we denote *PatchCamelyon20x*, the patches were sampled in accordance to the PatchCamelyon (Veeling et al., 2018) dataset, a pre-defined probabilistic sampling of the Camleyon16 dataset using the pixel annotations, resulting in a class-balance between tumor and non-tumor labels in the dataset. The original PatchCamelyon dataset is sampled at 10x, with patch size of 96x96 pixels. In this study however, the data was resampled to match the unsupervised dataset at the target resolution of 0.5 microns per pixel (approx. 20x) with patch size 256x256 pixels (at the same coordinates as the original dataset). In line with the PatchCamelyon dataset, training/validation/test samples for *PatchCamelyon20x* were taken from 216/54/129 slides respectively. In addition, subsets (possibly overlapping) of the supervised training dataset was selected as taking all patches from 10, 20, 50, 100 random slides, respectively (the smaller subsets are subsets of the larger ones). This was repeated five times to create five folds for each subset. For more details about PatchCamelyon, see Veeling et al. (2018). Pre-training SimCLR models using the unsupervised dataset is therefore considered *in-domain* pre-training, as the same slides are re-sampled and used for training the supervised, downstream task.

AIDA-LNSK (Lindman et al., 2019): a dataset containing 96 WSIs from 71 unique patients of skin tissue. The data was split into train, validation, and test on patient level, such that 50, 6, and 15 patients were included in train, validation and test respectively. This resulted in 65, 8, and 23 WSIs in each dataset. In analogy with Camelyon16, the AIDA-LNSK is sampled to create two dataset, one for downstream task training, and a corresponding *in-domain* dataset for pre-training.

For unsupervised training, patches were extracted from tissue regions of slides in the training set (65 slides), found by Otsu threshold from WSI magnification 5x. From these regions, the data was sampled without overlap. This resulted in an unsupervised dataset size of approximately 270k patches, roughly the same size as the unsupervised Camelyon16 dataset. All patches were extracted with 0.5 (mpp) resolution at a size of 256x256 pixels.

From AIDA-LNSK, a downstream task was constructed as a five-class tissue classification task, using available pixel-level annotations. The five classes were formed as four classes representing healthy tissue types (dermis, epidermis, subcutaneous tissue and skin appendage structure) and one class representing "abnormal" (containing different types of cancer, inflammation, scaring and so on). The slides from the above mentioned split was sampled (same size and resolution as the unsupervised dataset) such that for the supervised training dataset, each class included *at least* 75'000 samples, resulting in approximately 320k patches. The training set was thereafter subdivided, by randomly selecting all patches from 10, 20 and 50 slides from the original 65. Smaller subsets are true subsets of larger ones. This was repeated 5 times, such that for each subset size, 5 (possibly overlapping) dataset were created. The validation and test set were sampled from the respective slides in a class-

balanced way. For more information about the data collection and annotations, please see Stadler et al. (2021).

Multidata : Ciga et al. (2022) constructed a multi-data dataset, consisting of samples from 60 publicly available datasets, originating from multiple tissue types. This pre-sampled dataset was sampled unsupervised, and will in this study be used for self-supervised training only. Patches were extracted with size 224x224 pixels, at the maximum available resolution per dataset, resulting in a variation of resolution between the patches (0.25–0.5 mpp). In this study, we use a 1% subset of this data, provided by the authors, consisting of 40k patches. With relation to the downstream tasks of breast tumor classification and skin tissue classification, this data is considered *out-of-domain*.

4.3 Training

For all experiments, the ResNet50 (He et al., 2016) model architecture was used. As self-supervised method, SimCLR was evaluated, and if nothing else is stated, the same training setup was used as in Chen et al. (2020a).

The SimCLR objective is to minimize the NT-Xent loss. For a positive pair this is defined as

$$\mathcal{L}_{i,j} = -\log \frac{\exp(\text{sim}(z_i, z_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp(\text{sim}(z_i, z_k)/\tau)}, \tag{2}$$

where (i, j) is a positive pair, and (i, k) a negative one, with the similarity function defined as $\text{sim}(u, v) = u^T v/(\|u\|\|v\|)$ (cosine similarity). The temperature scaling, τ, was set to 0.5 for all experiments, unless otherwise stated.

In SimCLR, augmentations are used to create the positive views from the same anchor sample. Henceforth, "original" augmentations will refer to the augmentations defined in the SimCLR paper (randomly applying resize and crop, horizontal flip, color jittering and Gaussian blur) with the modification of when training with histopathology data, additional vertical flip and random rotation of 90 degrees was added (due to the rotation invariance of histopathology data). For examples, please see Appendix B.

Following commonly used protocol, the self-supervised pre-training was done once (due to computational and time constraints). The resulting representation is evaluated primarily using a linear classifier on top of the frozen weights, but also using fine-tuning (linear classifier on top, without freezing any weights). The former method is a way of evaluating the quality of the pre-trained representations with regard to the target data and objective, while the latter is a more realistic use-case of the trained weights. For supervised training cases, training was repeated 5 times with different seeds, and when subsets of the training data is used, also different folds. All results are reported as patch-wise accuracy on class-balanced test sets.

All training was conducted on 4 NVIDIA V100 or NVIDIA A100 GPUs. For SimCLR training, effective batch size of 1024 was used for 200 epochs (training time approximately 24 hours). LARS (You et al., 2017) was used as optimizer, with an initial learning rate 1.2 regulated with the a cosine annealing scheduler.

For linear evaluation, models were trained in a supervised manner for 20 epochs using Adam optimizer with an initial learning rate of 0.01. For fine-tuning, models were trained for 50 epochs. For breast, Adam optimizer with an initial learning rate of 1e−3 was used,

with weight decay of 1e−4. For skin, SGD optimizer with Nesterov momentum was used with initial learning rate of 1e−4 and momentum parameter of 0.9. In addition, models were trained in a supervised manner with random initialization ("from scratch"), using Adam optimizer with learning rate 0.001 for 50 epochs. Common for all supervised training was the usage of cosine annealing scheduler to reduce learning rate and weighted sampling to mitigate effects of class imbalance. Augmentations applied during supervised training consisted of: random resize crop with scale variance between 0.95–1.0, color jittering, and rotation/flip.

5. Results

Below follows detailed description and results from the experiments outlined in Section 4.1.

5.1 Positive-view generation by augmentation

We here investigate the effect of different augmentations, and their ability to isolate the task-relevant features in order to improve the correlation between the SSL objective and the downstream performance.

SimCLR models were trained on the unsupervised datasets from either breast or skin, and evaluated after 50 epochs on in-domain data from 50 slides, respectively. Eight different augmentation combinations were evaluated in terms of the relative improvement over *Base* augmentations (using flip, rotate, color jitter and very low scale variance, $0.95 - 1.0$). As Chen et al. (2020a) found large scale variance (0.2–1.0) together with Gaussian blur beneficial for ImageNet, these augmentations were evaluated both together and individually. Furthermore, two additional augmentations were evaluated, *Grid Distort* and *Shuffle*. These were chosen as transformations that preserve the label of histopathology patches, but adds perturbations of the compositions of the cells. The results are shown in Table 1. Further details and examples of the augmentations can be found in Appendix B.

Choosing optimal augmentations for histopathology data depend on dataset and downstream task   Looking at Table 1, choosing the appropriate augmentations for view

Table 1: Relative improvement (percentage points) over using base augmentations only (rotate, flip, color jitter). SimCLR model trained for 50 epochs, linear evaluation on supervised training set from 50 slides.

Augmentations	Breast	Skin
Base	81.75%	68.3%
+ {*Gaussian blur*}	+ 1.26	+ 0.37
+ {*Scale*}	**+ 3.94**	+ 0.64
+ {*Gaussian blur, Scale*} *(SimCLR Orig.)*	+ 0.2	+ 1.25
+ {*Scale, Grid Distort*}	+ 2.44	+ 1.42
+ {*Scale, Grid Distort, Shuffle*}	+ 1.72	+ 1.39
+ {*Grid Shuffle*}	+ 0.37	+ 2.83
+ {*Grid Distort, Shuffle*}	+ 1.6	**+ 3.79**

generation makes it possible to boost performance with 3.94 and 3.79 percentage points for breast and skin respectably. However, it appears as there is no common set of augmentations which is optimal for both datasets. Furthermore, using the same augmentations that were presented in Chen et al. (2020a) as optimal for ImageNet gives sub-optimal performance for histology data. Using Gaussian Blur was found to be of negligible value, and scale was only substantially beneficial for breast data, not for skin. The best set of augmentations for breast data was to use *Base + Scale*, while for skin, *Base + Grid Distort + Shuffle* gave the highest performance.

Thus, different sets of augmentations are optimal for different datasets and different downstream task. This is not surprising, as the relevant information in the data depends both on the inherited features in the dataset, as well as what the downstream task is. Finding task- and data-specific augmentations are therefore needed.

5.2 Downstream Performance

In this section, different pre-training strategies are evaluated. By using a random initialized model trained in a supervised way as reference, we want to understand the gain of using a pre-trained model depending on the size of the supervised data. In Figure 4, the result of linear evaluation (frozen pre-trained weights) for breast and skin is shown, where the supervised reference is shown in black (solid line). For each tissue type, five different pre-trained models are evaluated, either ImageNet Supervised (gray, solid) or four different configurations of SimCLR, where color denotes dataset (in-domain in red, out-of-domain in blue) and markers denote augmentations applied (original SimCLR in dashed, best set from Table 5.1 as dotted). Table 2 shows the fine-tuning results comparing the pre-training method giving the best performance on the linear evaluation with ImageNet pre-training
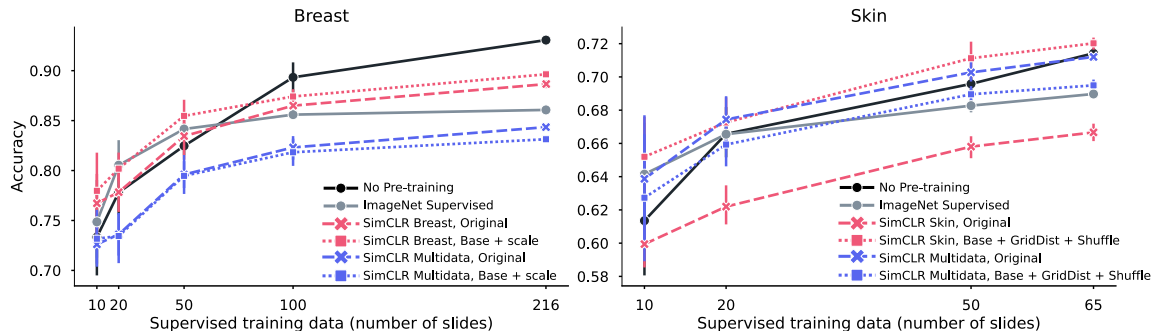

Figure 4: Patch-level performance of linear trainings of different pre-training strategies, at varying supervised training size (number of slides). The linear models are trained on representations learned by: ImageNet supervised pre-training (gray, solid), or using SimCLR pre-training with different augmentations (dashed/dotted), and on different datasets (in-domain pathology data: red, out-of-domain pathology data: blue). The reference (black, solid) is the full ResNet50 model trained in a supervised way on the (subset) training data. Left: breast data, right: skin data. Note different x-axes between subplots.

and no pre-training (random initialization). From the results in the figure and table, we can draw a number of conclusions.

In-domain pre-training boosts performance, especially in low-supervised data scenarios. The evaluation of linear training in Figure 4 shows that the best linear separation is given by pre-training using SimCLR on in-domain data with custom augmentations (red, dotted), exceeding ImageNet pre-training (gray, solid). These results are echoed also in the fine-tuning case, as shown in Table 2. Notably for smaller supervised training datasets (fewer than 65 slides), pre-training gives a substantial boost. When significantly more supervised training data is available (100 slides or more), the gain of using pre-trained weights is diminished. This is especially clear for breast data when doing fine-tuning (Table 2), where no pre-training on the full supervised dataset (216 slides) gave similar performance as using initalization from either ImageNet or SimCLR pre-trained weights. For skin, the size of the full supervised dataset (of 65 slides) is still small enough to make use of pre-trained weights a good idea.

Optimal dataset and view generation depend on downstream task. For breast data, we see in Figure 4 that the two in-domain models with different augmentations (red) gave similar performance, while for skin, different augmentations gave larger difference. Using a custom set of augmentations compared to the original SimCLR gave a significant boost in performance (dotted vs dashed). Furthermore, using Multidata as pre-training dataset that is out-of-domain pathology data (blue) gave for breast data the poorest performance, independent of augmentation, while for skin, Multidata with the original augmentations was on par or just slightly worse than the best in-domain model (blue, dashed). This corroborates the theory discussed in Section 3.2, that the features learned during pre-training are highly dependent on what data and how view generation was performed (i.e., what augmentations were applied), and that their usefulness/relevance are dependant on the downstream task.

Increasing diversity of pathology data is not beneficial *per se*. Both tissue types were evaluated on the Multidata dataset, with two different sets of augmentations each. These augmentations where chosen either as a general approach (SimCLR original) or a

Table 2: Fine-tuning performance (patch-level accuracy, %) using no pre-training (random initialized weights), ImageNet pre-training or SimCLR in-domain training.

| Data	Pre-training	Supervised size (# slides)					
		10	20	50	65	100	216
	None	73.33 ± 4.82	77.81 ± 2.27	82.47 ± 0.95	–	89.33 ± 1.73	93.06 ± 0.21
Breast	*ImageNet Supervised*	71.62 ± 5.74	82.67 ± 3.10	89.11 ± 2.43	–	91.35 ± 0.98	93.13 ± 0.42
	SimCLR Breast, Base + Scale	76.62 ± 6.92	84.58 ± 2.76	90.60 ± 1.81	–	91.63 ± 1.14	92.59 ± 0.36
	None	61.35 ± 3.88	66.57 ± 1.77	69.58 ± 0.66	71.42 ± 0.29	–	–
Skin	*ImageNet Supervised*	64.85 ± 4.22	67.64 ± 1.27	70.23 ± 0.14	70.96 ± 0.22	–	–
	SimCLR Skin, Base + Distort + Shuffle	66.27 ± 2.65	68.63 ± 0.95	72.65 ± 1.19	73.33 ± 0.22	–	–
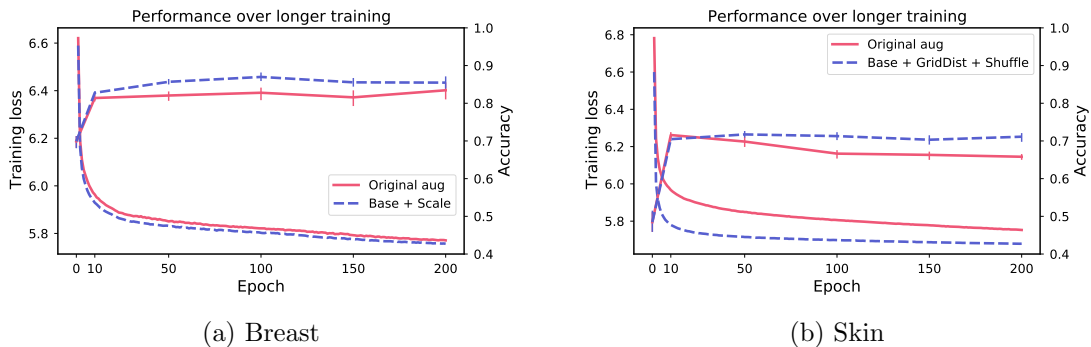
Figure 5: Training loss and downstream accuracy evaluated for two different models on breast and skin data, respectively. Loss continues to decrease continuously during training, while the performance on the downstream application evaluated at epochs {10, 20, 50, 100, 200} is after the initial epochs almost constant.

dataset specific augmentation. The Multidata dataset is smaller than the others, but has much larger diversity as it contains samples from a wide range of publicly available datasets. This reduces the risk of false negatives, and could potentially create more diverse sets of features. Looking at the linear performance in Figure 4, we see that the same model trained on Multidata with original augmentations (blue, dashed) performed poorly on the downstream task of breast tumor detection in sentinel lymph node tissue, but gave good results in skin tissue classification. This means that task-relevant features were to a higher degree extracted for skin as compared to breast. Either the dataset lack in the features needed for successful tumor detection in lymph node tissue, or the view generation could not isolate the relevant features. We conclude that a more diverse dataset does not guarantee a generalizable model *per se*, but acknowledge that diversity could potentially be beneficial if formulated correctly and used with the appropriate view generation strategy. That is, having a diverse dataset may increase the chance of including relevant features, but if those features are learned depends on the view generation.

5.3 Effects of hyper-parameters

Large batch sizes, long training times, and temperature scaling have been shown to play an important role in contrastive self-supervised learning for ImageNet (Chen et al., 2020a). Here, we investigate to see if this also holds true for histology data.

Longer training does not improve performance. Figure 5 shows evaluation of linear performance at intervals during training, for two models respectively for breast and skin. Despite continued reduction in training loss (the model is still learning to solve the SSL objective), the performance on the downstream task is changing very little after the first 10 epochs. This indicates that the view generation fails to isolate task-relevant features, making the model rely on task-irrelevant features to solve the SSL objective (scenario shown in Figure 2).

Large batch sizes are not needed. The motivation of large batch sizes is that this will form a better approximation of the true dataset distribution, wherein the separation of
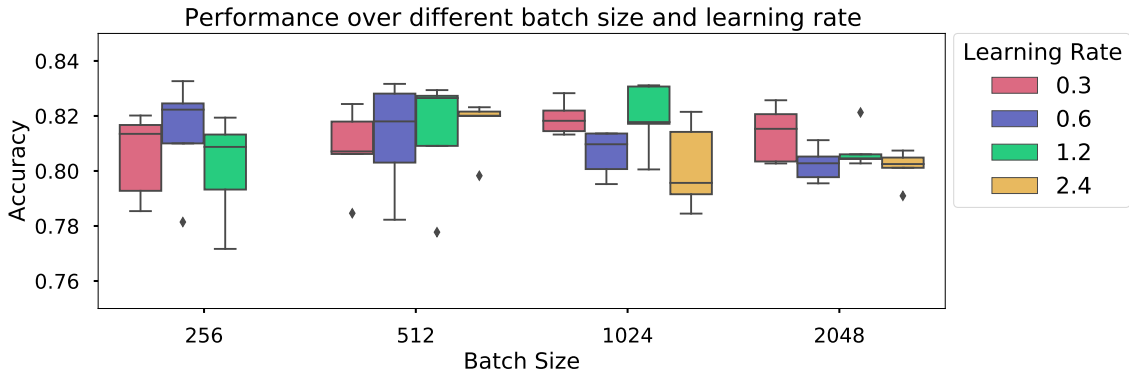
Figure 6: Performance for SimCLR trained with original augmentations for 50 epochs on breast data, using varying batch size and learning rate. Using learning rate of 2.4 with batch size 256 did not converge.
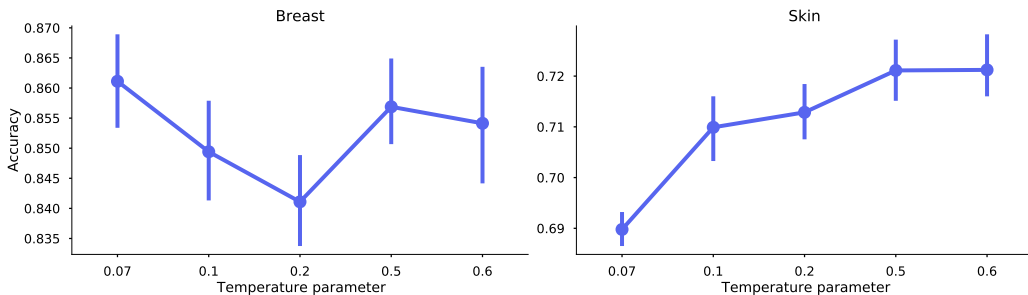

Figure 7: Linear performance (patch-level accuracy over five trainings) for SimCLR trained with best augmentations (from Table 1) for each datasets respectively, with varying temperature parameter. Note the different y-axes.

the positive and negatives will better reflect the true distribution. Figure 6 shows varying batch size for a model trained on breast data, with LARS optimizer. As long as the learning rate is updated according to the size of the batch (approximately following the relationship $LearningRate = 0.3 \cdot BatchSize/256$ as in Chen et al. (2020a)), increased batch size did not result in strictly better performance compared to smaller batch sizes. All batch sizes resulted in similar performance (at different learning rates), indicating that smaller batch sizes can be used to reach the same performance level as larger batch sizes. Similar results have been shown for volumetric medical data (Chaitanya et al., 2020).

Optimal temperature scaling is dataset dependant. Proper temperature scaling is important for the model to learn good representations (Ciga et al., 2022; Wang and Liu, 2021). In Figure 7, we investigate five different values of τ. The SimCLR models were trained on breast and skin data respectively, for 50 epochs and batch size 1024, using the best performing augmentations for each dataset (as of Table 1). Similarly as the results from Ciga et al. (2022), the different datasets may require different optimal values. However, the results show that the default value of 0.5 is a good compromise to achieve high levels of performance for both datasets.

16

6. Discussion

From the results in Section 5, we can make some interesting observations. Primarily, with correctly selected augmentations, in-domain contrastive SSL is beneficial as pre-training, especially in low-data regimes. In addition, experiments show that large batch sizes and long training times may not be needed to create pre-trained models, making model-creation more accessible. However, the results also raise concerns, which are discussed below.

6.1 Consequences of different dataset characteristics

The results presented in Section 5 show that the heuristics derived to be optimal for ImageNet does not transfer to the other datasets. This suggests that the method is tightly coupled with the dataset. We can identify a few important dataset characteristics that affect the learning outcome, as shown in Table 3 . In the table, ImageNet is compared with the histopathology datasets of breast sentinel lymph node tissue (Camelyon16) and skin tissue (AIDA-LNSK) (presented in Section 4.2), with respect to these characteristics, and the discussion is expanded below.

Number of classes and class balance affect the risk of false negatives. When negatives are chosen as random samples from the dataset, the distribution of the classes in the dataset affects the risk of getting a high number of false negatives. With many classes and perfect balance between them, the risk of drawing negatives belonging to the same class as the anchor data point is low. In the case of ImageNet, with 1000 classes,

Table 3: Dataset characteristics and the consequences they have for learning with contrastive self-supervised methods.

Dataset	Dataset characteristics				Consequence
	Number of classes	Class balance	Intra-/Inter-class variance	Downstream-target isolation	
ImageNet	1000	Medium/ Good	Good	Good	Many classes, good balance: **reduced risk of false negatives** High variance and good isolation: **easier view generation**
Camelyon16	2	Low	Low	Low/Medium	Few classes, poor balance: **higher risk of false negatives** Low variance with low/ medium isolation: **harder view generation**
AIDA-LNSK	5	Medium	Low	Low/Medium	Few classes, medium balance: **slightly higher risk of false negatives** Low variance with low/ medium isolation: **harder view generation**

drawing 1024 (or even up to 4096) random samples from the mini-batch as negatives, the likelihood of false negatives is low. Compare this with the histology datasets with 2 and 5 classes respectively, and where the poorer class balance further increases the risk for false negatives for the larger classes. In addition, ImageNet was collected in a supervised manner. For both histopathology data sources, the datasets used for SSL were sampled without knowledge of class labels, making the class distribution depend only on the natural occurrence of the class (in contrast to stratified sampling).

Diversity in/across classes and downstream-task isolation makes view generation easier. As contrastive SSL aims to do instance discrimination, having data samples that are distinct makes the objective easier. Variance across classes further helps the model to learn features that separate between classes. The nature of the datasets allows ImageNet to have higher variance both within and across classes, as attributes such as viewpoints, backgrounds, lighting, occlusions, etc, can vary in a natural dataset. Even within subgroups of labels, such as dogs, we have a larger variation due to more diversity of color, texture, shape and size – all of which, to a large extent, are constant between histopathology classes. In the ImageNet case, the augmentations applied can make use of the known variance, making the model invariant to them. The subtleties of the difference between classes in histology data makes it more challenging to find effective augmentations.

Downstream-target isolation relates to how isolated the downstream targets are on average in one image. ImageNet contains many object-centered images, containing only the class object, while both histopathology datasets have images that contain multiple classes (tissues and/or cell types). Having multiple objects in an image can introduce noise. For example, augmentations such as scale may in cases with low downstream-target isolation create a positive pair depicting two separate objects, instead of the same object in different scales.

6.2 How to do contrastive learning for histology?

From what we have seen so far, the cause of the success of contrastive self-supervised learning (SSL) methods on ImageNet has been highly dependent of intrinsic properties of the data. The intricate interplay between method and data raises questions on both how to adapt the method to better accommodate the data and how to better assemble datasets that fit the method.

Current positive-view generation is not sufficient. From the results shown in Table 1 we saw that a tailored set of augmentations gives substantial improvements in downstream performance. However, we also saw in Figure 5 that the representations learned are, to a large extent, based on features not relevant for the downstream task. As optimal augmentations depend on both dataset and task, finding a general approach that applies to all datasets and all tasks may not be feasible. The augmentations found in this study were sub-optimal even when tailored with a specific dataset and task in mind. Creating augmentations that are strong enough to retain only label information and remove all other is not trivial, and may require extensive domain knowledge. This is further exacerbated by the fact that pathologists generally are not used to describing diagnostic criteria in terms of features suitable to formulate as image transformations. Moreover, use of heavily tailored

augmentations could be criticized as a step towards unwanted feature engineering, in the sense that expert preconceptions could constrain the self-supervised deep learning approach A different direction would be to optimally learn what augmentation to apply, such as presented by Tamkin et al. (2021). However, contrary to what Tamkin et al. (2021) suggest, the conclusion from our results is that these augmentations need to be optimized for the downstream task, not the SSL objective. A semi-self-supervised approach could therefore be an interesting future research direction.

Many false-negatives gives conflicting signals   Datasets with few classes and/or large class imbalances suffers a large risk of introducing false negatives . As the mini-batch contains many samples of the same class, negatives picked from the same class are likely to occur. This risk also increases with larger batch sizes, as the ratio between number of classes and batch size increases. This could be one explanation why there is little benefit of increased batch size for histopathology data (Figure 6). Having a large portion of false negatives has consequences for the learning outcome, as this prevents the model from using class-specific features to discriminate between positives and negatives (as highlighted in Figure 2). We can further investigate this by looking at the cosine distance between samples in a mini-batch (1024 samples) of a SimCLR model trained on breast data. A significant number of negative samples have high or very high cosine similarity (> 0.9) with the anchor data point, an occurrence not seen for ImageNet data. Out of all negatives, 2% of the anchor-negative pairs had a similarity higher than 0.9, while the same number of ImageNet is 0.0003%. Minimizing the number of false negatives is an important part of getting better performance (Chuang et al., 2020; Chaitanya et al., 2020).

Methods for completely removing negatives have been presented, such as BYOL (Grill et al., 2020), which no longer uses the contrastive objective. Some exploratory experiments in this direction using BYOL are shown in Appendix C, Table 9. In these experiments, BYOL does not outperform SimCLR on either of the datasets. However, further research is needed to fully understand the role of negatives for optimal self-supervised learning for histopathology.

Intrinsic properties of histology datasets may be incompatible with current methods.   As discussed in Section 6.1, intrinsic properties of the datasets makes positive view generations challenging (low inter- and intra-class variance and lower target object isolation in individual images), and increases the risk of false-negatives (due to low number of classes and poor class distribution). Even if these problems could be addressed with new techniques such as better view generation and true negative sampling, questions regarding the suitability of SSL for these types of datasets remain. The SimCLR objective optimizes towards *instance discrimination*. This approach is intuitive when we have a dataset where the intra-class variance consists of multiple ways of describing the same object. Being able to separate each of these instance helps give a wider distribution of the possible appearances of the object in question. In histopathology, datasets are constructed as smaller patches from whole-slide images, and where the intra-class variance consists of images showing multiple cells, where the cells actually are more or less clones of each other. A different approach of constructing these datasets may be needed, such that the downstream-target isolation becomes higher. This is indeed challenging. Taking smaller patches depicting as little as individual cells suffers even more of false negatives, and macro structures may be

hard to learn. Taking larger patches/reduce resolution would include larger structures, but may include multiple tissues at once, reducing downstream-target isolation further (visual comparison of patches sampled at different resolutions are shown in Appendix 10).

If you understand your data, then contrastive self-supervised learning can still be useful. Despite the above mentioned limitations for contrastive SSL used on histology data, the current setup may still bring value for specific histology applications. For example, the reduced need for large batch sizes and long training times makes in-domain pre-training using contrastive SSL accessible to a larger community. Furthermore, by keeping the dataset characteristics from Table 3 in mind, risk factors for the dataset in question may be early identified. By understanding the inter- and intra-variance of the (latent) classes of the dataset, augmentations may be formulated that are better tailored to the specific downstream application in mind, compared to naively using those optimized for ImageNet. The risk of false negatives might be possible to mitigate during the data collection, for example by controlling the field of view by changing the resolution in which the patches are sampled. In combination with the results shown in Table 2 that shows that in-domain SimCLR pre-training does boost performance, contrastive self-supervised learning can indeed be a way to reduce the need for labeled data for histopathology applications.

6.3 Limitations and Future work

This paper aims to evaluate if and how contrastive self-supervised methods can be used to reduce the needed amount of labeled data for the target histopathology application. The scope of the study was limited to three datasets and two classification tasks, and where SimCLR was chosen as representative method among all contrastive self-supervised methods. Restricted by the challenges and limitations of pre-training models for *one* downstream task, we did not evaluate the generalization of the SSL models by evaluating one pre-trained model on multiple downstream tasks, with one exception (SimCLR Multidata with original augmentations was used as pre-trained model for both downstream tasks). Despite the restricted scope, we believe that the results may give guidance when applied to an extended domain.

The results from this study show that contrastive self-supervised methods have the potential, if applied correctly, to reduce the need for labeled target data. However, they also show that the method is still sub-optimal with respect to the specific data characteristics of histopathology. There is therefore room for improvement, but the challenges of creating informative positives and reduce false negatives are not trivial to solve. Creating informative positives may be easier with deepened understanding of what features should be considered task-relevant for a given downstream task. The problem of false negatives could potentially be solved by selecting negatives in a non-random way, potentially taking a semi-supervised approach. We hope that this work will inspire interesting future research that take a holistic approach, considering the interplay between dataset and method.

7. Conclusions

In this paper, we have evaluated contrastive self-supervised learning on histopathology data. Effective contrastive self-supervised learning with respect to a particular downstream task

requires two criteria to be fulfilled, namely that the shared information between the positive views is high, and that the false negative rate is low. Our study shows that both these criteria are challenging to fulfill for histopathology applications, due to the characteristics of the datasets. Furthermore, we have shown that the explicit and implicit heuristics used for ImageNet does not necessarily apply in the domain of histopathology. We conclude that SSL for histopathology cannot be considered and used under the same assumptions as for natural images, and that in-depth understanding of the data is essential for training self-supervised models for histopathology applications.

Acknowledgments

Ethical Standards

The work follows appropriate ethical standards in conducting research and writing the manuscript, following all applicable laws and regulations regarding treatment of animals or human subjects.

Conflicts of Interest

We declare that we do not have any conflicts of interest.

References

Sanjeev Arora, Hrishikesh Khandeparkar, Mikhail Khodak, Orestis Plevrakis, and Nikunj Saunshi. A theoretical analysis of contrastive unsupervised representation learning. In *Proceedings of the 36th International Conference on Machine Learning (ICML 2019)*, pages 5628–5637, 2019.

Shekoofeh Azizi, Basil Mustafa, Fiona Ryan, Zachary Beaver, Jan Freyberg, Jonathan Deaton, Aaron Loh, Alan Karthikesalingam, Simon Kornblith, Ting Chen, Vivek Natarajan, and Mohammad Norouzi. Big Self-Supervised Models Advance Medical Image Classification. *arXiv:2101.05224 [cs, eess]*, 2021.

Philip Bachman, R. Devon Hjelm, and William Buchwalter. Learning representations by maximizing mutual information across views. In *Advances in Neural Information Processing Systems (NeurIPS 2019)*, pages 15509–15519, 2019.

Abeba Birhane and Vinay Uday Prabhu. Large Image Datasets: A Pyrrhic Win for Computer Vision? In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1537–1547, 2021.

Alexander Buslaev, Vladimir I. Iglovikov, Eugene Khvedchenya, Alex Parinov, Mikhail Druzhinin, and Alexandr A. Kalinin. Albumentations: Fast and Flexible Image Augmentations. *Information*, 11(2):125, 2020.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

Krishna Chaitanya, Ertunc Erdil, Neerav Karani, and Ender Konukoglu. Contrastive learning of global and local features for medical image segmentation with limited annotations. In *Advances in Neural Information Processing Systems*, volume 33, pages 12546–12558. Curran Associates, Inc., 2020.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, pages 1597–1607, 2020a.

Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved Baselines with Momentum Contrastive Learning. *CoRR*, abs/2003.04297, 2020b.

Ching-Yao Chuang, Joshua Robinson, Yen-Chen Lin, Antonio Torralba, and Stefanie Jegelka. Debiased contrastive learning. In *Advances in Neural Information Processing Systems (NeurIPS 2020)*, 2020.

Ozan Ciga, Tony Xu, and Anne Louise Martel. Self supervised contrastive learning for digital histopathology. *Machine Learning with Applications*, 7:100198, 2022. ISSN 2666-8270.

Elijah Cole, Xuan Yang, Kimberly Wilber, Oisin Mac Aodha, and Serge Belongie. When Does Contrastive Visual Representation Learning Work? *CoRR*, abs/2105.05837, 2021.

Yin Cui, Yang Song, Chen Sun, Andrew Howard, and Serge J. Belongie. Large scale fine-grained categorization and domain-specific transfer learning. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 4109–4118, 2018.

Olivier Dehaene, Axel Camara, Olivier Moindrot, Axel de Lavergne, and Pierre Courtiol. Self-Supervision Closes the Gap Between Weak and Strong Supervision in Histology. *CoRR*, abs/2012.03583, 2020.

J. Deng, W. Dong, R. Socher, L. Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

Alexey Dosovitskiy, Jost Tobias Springenberg, Martin Riedmiller, and Thomas Brox. Discriminative Unsupervised Feature Learning with Convolutional Neural Networks. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014.

William Falcon and Kyunghyun Cho. A framework for contrastive self-supervised learning and designing a new approach. *CoRR*, abs/2009.00104, 2020.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut Learning in Deep Neural Networks. *Nature Machine Intelligence*, 2(11):665–673, 2020. ISSN 2522-5839.

Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. In *6th International Conference on Learning Representations (ICLR 2018)*, 2018.

Jacob Gildenblat and Eldad Klaiman. Self-Supervised Similarity Learning for Digital Pathology. *CoRR*, abs/1905.08139, 2020.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Ávila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent - A new approach to self-supervised learning. In *Advances in Neural Information Processing Systems (NeurIPS 2020)*, 2020.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum Contrast for Unsupervised Visual Representation Learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9726–9735, 2020.

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. *arXiv:2008.05571 [cs, eess]*, 2020.

Gustav Larsson, Michael Maire, and Gregory Shakhnarovich. Colorization as a Proxy Task for Visual Understanding. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 840–849, 2017.

Hao Li, Pratik Chaudhari, Hao Yang, Michael Lam, Avinash Ravichandran, Rahul Bhotika, and Stefano Soatto. Rethinking the hyperparameters for fine-tuning. In *8th International Conference on Learning Representations (ICLR 2020)*, 2020.

Hongwei Li, Fei-Fei Xue, Krishna Chaitanya, Shengda Luo, Ivan Ezhov, Benedikt Wiestler, Jianguo Zhang, and Bjoern Menze. Imbalance-Aware Self-supervised Learning for 3D Radiomic Representations. In Marleen de Bruijne, Philippe C. Cattin, Stéphane Cotin, Nicolas Padoy, Stefanie Speidel, Yefeng Zheng, and Caroline Essert, editors, *Medical*

Image Computing and Computer Assisted Intervention – MICCAI 2021, pages 36–46. Springer International Publishing, 2021a. ISBN 978-3-030-87196-3.

Jiajun Li, Tiancheng Lin, and Yi Xu. SSLP: Spatial Guided Self-supervised Learning on Pathological Images. In Marleen de Bruijne, Philippe C. Cattin, Stéphane Cotin, Nicolas Padoy, Stefanie Speidel, Yefeng Zheng, and Caroline Essert, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, Lecture Notes in Computer Science, pages 3–12, Cham, 2021b. Springer International Publishing. ISBN 978-3-030-87196-3.

Karin Lindman, F. Rose Jerónimo, Martin Lindvall, and Caroline Bivik Stadler. Skin data from the Visual Sweden project DROID. 2019.

Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen A. W. M. van der Laak, Bram van Ginneken, and Clara I. Sánchez. A Survey on Deep Learning in Medical Image Analysis. *Medical Image Analysis*, 42:60–88, 2017. ISSN 13618415.

Geert Litjens, Peter Bandi, Babak Ehteshami Bejnordi, Oscar Geessink, Maschenka Balkenhol, Peter Bult, Altuna Halilovic, Meyke Hermsen, Rob van de Loo, Rob Vogels, Quirine F. Manson, Nikolas Stathonikos, Alexi Baidoshvili, Paul van Diest, Carla Wauters, Marcory van Dijk, and Jeroen van der Laak. 1399 H&E-stained sentinel lymph node sections of breast cancer patients: The CAMELYON dataset. *GigaScience*, 7(6), 2018.

Jingyu Liu, Gangming Zhao, Yu Fei, Ming Zhang, Yizhou Wang, and Yizhou Yu. Align, Attend and Locate: Chest X-Ray Diagnosis via Contrast Induced Attention Network With Limited Supervision. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10631–10640. IEEE, 2019. ISBN 978-1-72814-803-8.

Ming Y. Lu, Richard J. Chen, Jingwen Wang, Debora Dillon, and Faisal Mahmood. Semi-Supervised Histology Classification using Deep Multiple Instance Learning and Contrastive Predictive Coding. *CoRR*, abs/1910.10825, 2019.

Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. UMAP: Uniform Manifold Approximation and Projection. *J. Open Source Softw.*, 3(29):861, 2018.

Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI*, pages 69–84, 2016.

Senthil Purushwalkam and Abhinav Gupta. Demystifying contrastive self-supervised learning: Invariances, augmentations and dataset biases. In *Advances in Neural Information Processing Systems (NeurIPS 2020)*, 2020.

Maithra Raghu, Chiyuan Zhang, Jon Kleinberg, and Samy Bengio. Transfusion: Understanding Transfer Learning with Applications to Medical Imaging. *CoRR*, abs/1902.07208, 2019.

Joshua David Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. Contrastive learning with hard negative samples. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*, 2021.

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.

Chetan L. Srinidhi, Seung Wook Kim, Fu-Der Chen, and Anne L. Martel. Self-supervised driven consistency training for annotation efficient histopathology image analysis. *Medical Image Analysis*, 75:102256, 2022. ISSN 1361-8415.

Karin Stacke, Claes Lundström, Jonas Unger, and Gabriel Eilertsen. Evaluation of Contrastive Predictive Coding for Histopathology Applications. In *Proceedings of the Machine Learning for Health NeurIPS Workshop*, pages 328–340. PMLR, 2020.

Caroline Bivik Stadler, Martin Lindvall, Claes Lundström, Anna Bodén, Karin Lindman, Jeronimo Rose, Darren Treanor, Johan Blomma, Karin Stacke, Nicolas Pinchaud, Martin Hedlund, Filip Landgren, Mischa Woisetschläger, and Daniel Forsberg. Proactive Construction of an Annotated Imaging Database for Artificial Intelligence Training. *Journal of Digital Imaging*, 34(1):105–115, 2021. ISSN 1618-727X.

Alex Tamkin, Mike Wu, and Noah D. Goodman. Viewmaker networks: Learning views for unsupervised representation learning. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*, 2021.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *European conference on computer vision (ECCV 2020)*, pages 776–794. Springer, 2020a.

Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning? In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020b.

Antonio Torralba and Alexei A. Efros. Unbiased look at dataset bias. In *The 24th IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2011)*, pages 1521–1528, 2011.

Tuan Truong, Sadegh Mohammadi, and Matthias Lenga. How Transferable Are Self-supervised Features in Medical Image Classification Tasks? *arXiv:2108.10048 [cs]*, 2021.

Yao-Hung Hubert Tsai, Yue Wu, Ruslan Salakhutdinov, and Louis-Philippe Morency. Self-supervised learning from a multi-view perspective. In *9th International Conference on Learning Representations (ICLR 2021)*, 2021.

Michael Tschannen, Josip Djolonga, Paul K. Rubenstein, Sylvain Gelly, and Mario Lucic. On mutual information maximization for representation learning. In *8th International Conference on Learning Representations (ICLR 2020)*, 2020.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation Learning with Contrastive Predictive Coding. *CoRR*, abs/1807.03748, 2019.

Bastiaan S. Veeling, Jasper Linmans, Jim Winkens, Taco Cohen, and Max Welling. Rotation equivariant cnns for digital pathology. In *Medical Image Computing and Computer Assisted Intervention - MICCAI 2018 - 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part II*, pages 210–218, 2018.

Feng Wang and Huaping Liu. Understanding the Behaviour of Contrastive Loss. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2495–2504. IEEE, 2021. ISBN 978-1-66544-509-2.

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 *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI 2021)*, pages 186–195. Springer, 2021.

Mike Wu, Chengxu Zhuang, Milan Mosse, Daniel Yamins, and Noah Goodman. On Mutual Information in Contrastive Learning for Visual Representations. *CoRR*, abs/2005.13149, 2020.

Zhirong Wu, Yuanjun Xiong, Stella Yu, and Dahua Lin. Unsupervised Feature Learning via Non-Parametric Instance-level Discrimination. *CoRR*, abs/1805.01978, 2018.

Yutong Xie, Jianpeng Zhang, Zehui Liao, Yong Xia, and Chunhua Shen. PGL: prior-guided local self-supervised learning for 3D medical image segmentation. *arXiv preprint arXiv:2011.12640*, 2020.

Ke Yan, Jinzheng Cai, Dakai Jin, Shun Miao, Adam P Harrison, Dazhou Guo, Youbao Tang, Jing Xiao, Jingjing Lu, and Le Lu. Self-supervised learning of pixel-wise anatomical embeddings in radiological images. *arXiv preprint arXiv:2012.02383*, 2020.

Kaiyu Yang, Jacqueline Yau, Li Fei-Fei, Jia Deng, and Olga Russakovsky. A Study of Face Obfuscation in ImageNet. *CoRR*, abs/2103.06191, 2021a.

Pengshuai Yang, Zhiwei Hong, Xiaoxu Yin, Chengzhan Zhu, and Rui Jiang. Self-supervised Visual Representation Learning for Histopathological Images. In Marleen de Bruijne, Philippe C. Cattin, Stéphane Cotin, Nicolas Padoy, Stefanie Speidel, Yefeng Zheng, and Caroline Essert, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, Lecture Notes in Computer Science, pages 47–57, Cham, 2021b. Springer International Publishing. ISBN 978-3-030-87196-3.

Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 3320–3328. Curran Associates, Inc., 2014.

Chenyu You, Yuan Zhou, Ruihan Zhao, Lawrence Staib, and James S Duncan. Simcvd: Simple contrastive voxel-wise representation distillation for semi-supervised medical image segmentation. *arXiv preprint arXiv:2108.06227*, 2021.

Yang You, Igor Gitman, and Boris Ginsburg. Large Batch Training of Convolutional Networks. *arXiv:1708.03888 [cs]*, 2017.

Richard Zhang, Phillip Isola, and Alexei A. Efros. Colorful image colorization. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III*, pages 649–666, 2016.

Hong-Yu Zhou, Shuang Yu, Cheng Bian, Yifan Hu, Kai Ma, and Yefeng Zheng. Comparing to Learn: Surpassing ImageNet Pretraining on Radiographs by Comparing Image Representations. In Anne L. Martel, Purang Abolmaesumi, Danail Stoyanov, Diana Mateus, Maria A. Zuluaga, S. Kevin Zhou, Daniel Racoceanu, and Leo Joskowicz, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, pages 398–407. Springer International Publishing, 2020. ISBN 978-3-030-59710-8.

Appendix A. Datasets

In Table 4 the number of slides and patches are given for the different datasets. All patches were extracted at 0.5 microns per pixel resolution with size 256x256 pixels. The unsupervised dataset were sampled without overlap in a uniform grid. For breast, maximally 1000 samples per slide was randomly selected from slides in the training set, resulting in approximately 270k patches. Similarly for skin, the unsupervised training set consists of samples selected without using labels, resulting in approximately 270k patches. Potential sampling points were found by, in tissue regions, extracting samples from a random, uniform grid. From these candidates, 1000 samples per slide were extracted at random.

The supervised training, validation and test set for breast follows the patch coordinates of PatchCamelyon, but was resampled at the above mentioned resolution and size. For skin, the supervised datasets were constructed as follows. All slides were sampled in a uniform grid with 50% overlap, resulting in total of 1.3 million candidate patches. From these candidates, a subset where selected for each dataset, with the follow criteria. For the training set, the large class imbalance was mitigated by selecting patches such that for each class, the total number of patches was `min`(75 000, `all available`). This resulted in approx. 320k patches. For validation, 700 patches from each class were randomly selected from the slides in the validation set. Similarly, 3700 patches from each class were randomly selected from the slides in the test set to formulate the test patches. The validation and test sets are therefore class balanced. There is no patient overlap between the supervised datasets.

Table 4: Number of whole-slide images (WSIs) and patches for breast and skin data.

Unsupervised		*Training*	*Validation*	*Test*
Breast	*# WSI*	270	N/A	N/A
	# patches	265048	N/A	N/A
Skin	*# WSI*	65	N/A	N/A
	# patches	271675	N/A	N/A
Supervised				
Breast	*# WSI*	216 / 100 / 50 / 20 / 10	54	129
	# patches	262144 / ∼120000 / ∼60000 / ∼25000 / ∼10000	32768	32768
Skin	*# WSI*	65 / 50 / 20 / 10	8	23
	# patches	317243 / ∼235000 / ∼95000 / ∼50000	3500	18500

Appendix B. Augmentations

The augmentations applied were done using Pytorch Transforms (`https://pytorch.org/vision/stable/transforms.html`) or with Albumentations (Buslaev et al., 2020). The im-

plementation of Gaussian blur was taken from: `https://github.com/facebookresearch/moco`. Examples are shown in Figure 8, and implementation details in Table 5.

Table 5: List of applied augmentations, with corresponding parameters and probability of application.

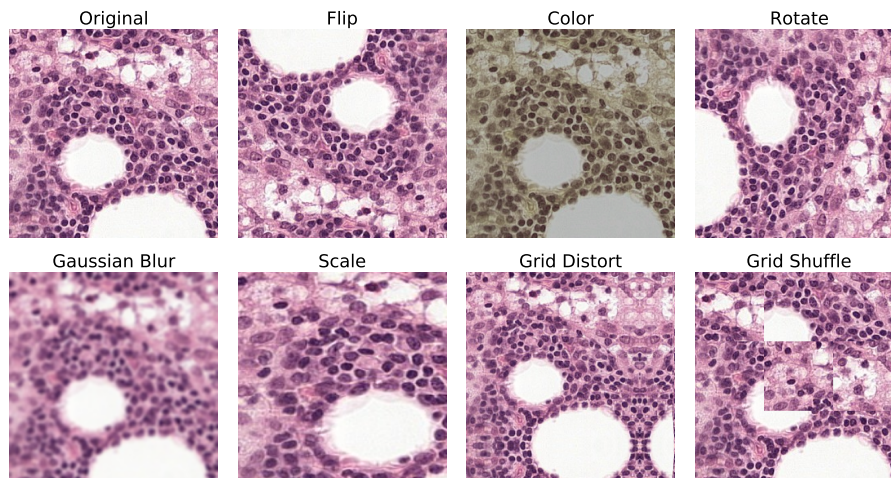|	Transformation	Params	Probability
Base	*Random Crop*	size: 224x224	1.0
	Flip	-	0.5
	Rotation (fixed 90 degrees)	-	0.5
	Color Jitter	brightness: 0.8	0.8
		contrast: 0.8	
		saturation: 0.8	
		hue: 0.2	
SimCLR original	*Scale*	scale: $\{0.2, 0.95\} - 1.0$	1.0
	Gaussian Blur	sigma: 0.1–2.0	$\{0, 0.5\}$
	Grid Distortion	num_steps: 9	$\{0, 0.5\}$
		distort_limit: 0.2	
		border_mode: 2	
	(Grid) Shuffle	grid: (3,3)	$\{0, 0.5\}$

Figure 8: Examples of augmentations

Appendix C. Results

Tables 6 and 7 show the results presented in Figure 4 in numerical form.

Table 6: Linear performance for patch-level PatchCamelyon, shown as accuracy (%) and patch-level AUC

	Model	Supervised size (#slides)				
		10	20	50	100	216
Accuracy (%)	*Supervised*	73.3 ± 4.8	77.8 ± 2.3	82.5 ± 0.9	89.3 ± 1.7	93.1 ± 0.2
	ImageNet Supervised	74.9 ± 4.9	80.5 ± 3.1	84.2 ± 0.5	85.6 ± 0.4	86.1 ± 0.1
	SimCLR Breast, Original	76.7 ± 3.9	77.8 ± 2.7	83.4 ± 2.6	86.5 ± 0.5	88.7 ± 0.3
	SimCLR Breast, Base + Scale	78.0 ± 5.0	80.2 ± 1.8	85.5 ± 1.8	87.4 ± 0.8	89.6 ± 0.1
	Multidata, Original	72.6 ± 2.9	73.6 ± 2.7	79.6 ± 2.4	82.3 ± 1.4	84.3 ± 0.1
	Multidata, Base + Scale	73.2 ± 3.2	73.4 ± 3.2	79.5 ± 1.6	81.8 ± 1.6	83.1 ± 0.1
AUC	*Supervised*	0.805 ± 0.055	0.855 ± 0.027	0.900 ± 0.010	0.960 ± 0.009	0.979 ± 0.001
	ImageNet Supervised	0.838 ± 0.046	0.887 ± 0.024	0.918 ± 0.004	0.928 ± 0.002	0.935 ± 0.000
	SimCLR Breast, Original	0.830 ± 0.072	0.865 ± 0.057	0.920 ± 0.032	0.947 ± 0.008	0.957 ± 0.001
	SimCLR Breast, Base + Scale	0.842 ± 0.074	0.884 ± 0.031	0.929 ± 0.019	0.944 ± 0.010	0.957 ± 0.000
	Multidata, Original	0.820 ± 0.041	0.837 ± 0.044	0.888 ± 0.021	0.910 ± 0.015	0.921 ± 0.001
	Multidata, Base + Scale	0.827 ± 0.047	0.839 ± 0.053	0.892 ± 0.015	0.909 ± 0.012	0.909 ± 0.001

Table 7: Linear performance (patch-level accuracy (%)), AIDA-LNSK

	Model	Supervised size (#slides)			
		10	20	50	65
Accuracy (%)	*Supervised*	61.4 ± 3.9	66.6 ± 1.8	69.6 ± 0.7	71.4 ± 0.3
	ImageNet Supervised	64.2 ± 3.9	66.6 ± 1.2	68.3 ± 0.4	69.0 ± 0.2
	SimCLR Skin, Original	59.9 ± 1.8	62.2 ± 1.4	65.8 ± 0.8	66.7 ± 0.6
	SimCLR Skin, Base + GridDist + Shuffle	65.2 ± 3.0	67.3 ± 1.7	71.1 ± 1.3	72.0 ± 0.4
	Multidata, Original	63.9 ± 5.1	67.4 ± 1.7	70.3 ± 0.7	71.2 ± 0.2
	Multidata, Base + GridDist + Shuffle	62.7 ± 4.9	65.9 ± 1.5	69.0 ± 0.5	69.5 ± 0.3

Table 8 shows the results presented in Figure 6 in numerical form.

Table 8: Linear performance (patch-level accuracy (%)) of SimCLR model trained with original augmentations on breast data for 50 epochs, evaluated on PatchCamelyon (50 slides subset).

Batch Size	Learning Rate			
	0.3	0.6	1.2	2.4
256	0.81 ± 0.02	0.81 ± 0.02	0.80 ± 0.02	-
512	0.81 ± 0.02	0.81 ± 0.02	0.81 ± 0.02	0.82 ± 0.01
1024	0.82 ± 0.01	0.81 ± 0.01	0.82 ± 0.01	0.80 ± 0.02
2048	0.81 ± 0.01	0.80 ± 0.01	0.81 ± 0.01	0.80 ± 0.01

In Table 9, results are shown comparing SimCLR model to BYOL. The models were trained on in-domain data, with either the original SimCLR augmentations or the best performing augmentations from Table 1. The models were trained for 200 epochs, with batch size 1024. The SimCLR model was trained with learning rate 1.2, but BYOL was found to be needed lower learning rate 0.2. The results are presented as the mean linear patch-wise accuracy of the supervised subset of 50 slides, evaluted over 5 runs.

Table 9: Comparison between SimCLR method and BYOL, trained for 200 epochs, evaluated on linear patch-level accuracy of the supervised subset of 50 slides.

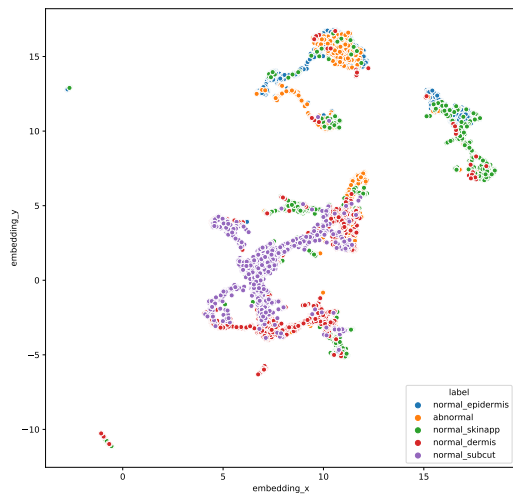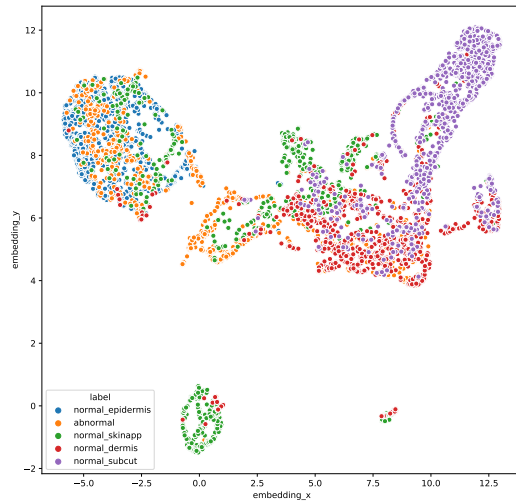| Dataset	Method	Augmentation	
		Original	Best.
Breast	*SimCLR*	0.83 ± 0.03	0.85 ± 0.02
	BYOL	0.79 ± 0.01	0.83 ± 0.00
Skin	*SimCLR*	0.66 ± 0.01	0.71 ± 0.01
	BYOL	0.48 ± 0.03	0.48 ± 0.03

Appendix D. Discussion
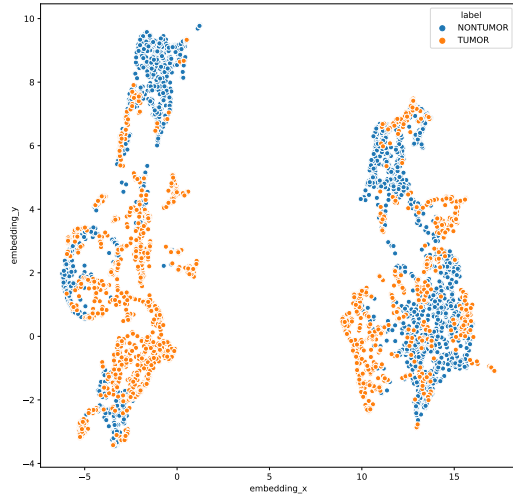
(a) SimCLR trained with Original augmentations on breast data.
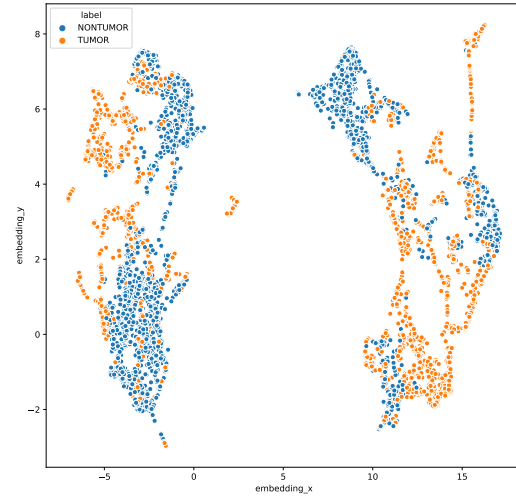
(b) SimCLR trained with Base + Scale augmentations on breast data.

(c) SimCLR trained with Original augmentations on skin data.

(d) SimCLR trained with Base + Grid shuffle + Distort on skin data.

Figure 9: UMAP (McInnes et al., 2018) visualizations of encoder embeddings of models trained on in-domain data for 200 epochs, evaluated on 5000 randomly selected samples from the test sets (class balanced). Top row: breast data, bottom row: skin data.
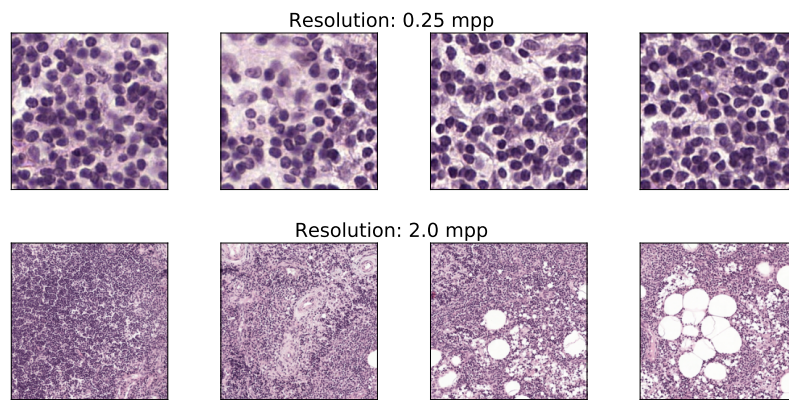
Figure 10: Examples neighboring of patches extracted at different resolutions, with constant patch size. The different fields of view and level of details visible impacts what features the contrastive model will learn.