Title
Label-efficient Representation Learning for Medical Image Analysis

Author
Yang, Jiawei

Publication Date
2023

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Label-efficient Representation Learning for Medical Image Analysis

A thesis submitted in partial satisfaction

of the requirements for the degree

Master of Science in Electrical and Computer Engineering

by

Jiawei Yang

2023

ABSTRACT OF THE THESIS

Label-efficient Representation Learning for Medical Image Analysis

by

Jiawei Yang
Master of Science in Electrical and Computer Engineering
University of California, Los Angeles, 2023
Professor Lei He, Chair

This thesis aims to partially tackle the inherent challenges of data-hungry deep learning methods for medical image analysis due to the scarcity of annotated training data in the medical domain. The focus is on investigating novel solutions within the realms of few-shot learning, multiple-instance learning, and self-supervised learning, specifically centering on histopathology images for coherence.

The first part of the research involves the use of contrastive learning (CL) and latent augmentation (LA) to enhance the efficiency and generalizability of few-shot learning in histology images. The study seeks to understand the conditions under which self-supervised models outperform supervised ones and explores the potential of self-supervised representations. For instance, it reveals that SSL models pre-trained on pathological images excel in few-shot classification settings compared to supervised models. This is because SSL models learn class-agnostic information, whereas supervised models, which focus on discriminative features, are sensitive to shifts in data distribution. Additionally, it demonstrates that LA, by introducing semantic variations in an unsupervised way, can significantly improve few-shot classification performance.

The second part presents ReMix, a novel framework for multiple-instance learning (MIL)-based whole-slide image (WSI) classification. ReMix addresses training efficiency and data

diversity challenges by substituting instances with instance prototypes (patch cluster centroids) and employing online, stochastic, and flexible latent space augmentations to enforce semantic-perturbation invariance. This technique has been shown to boost the performance and efficiency of both spatial-agnostic and spatial-aware MIL methods.

Finally, the study delves into self-supervised learning (SSL) for dense prediction tasks in pathology images. A new SSL framework, Concept Contrastive Learning (ConCL), is introduced, proven to outperform previous state-of-the-art SSL methods. The main objective of ConCL is to enhance detection and segmentation tasks in computational pathology, which are often heavily dependent on annotated data, hence challenging to execute efficiently and accurately. A roadmap is provided for pre-training a superior encoder for downstream dense prediction tasks. Furthermore, a simple, dependency-free concept-generating method is proposed that does not rely on external segmentation algorithms or saliency detection models.

In summary, this thesis broadens the understanding of deep learning applications in healthcare, demonstrating the power of data augmentation and representation learning in medical image analysis across various settings. It encourages further investigation into these challenges to enhance the speed and accuracy of diagnoses, improve treatment decisions, and reduce medical errors.

The thesis of Jiawei Yang is approved.

Lin Yang

Bolei Zhou

Lei He, Committee Chair

University of California, Los Angeles

2023

To my parents, Huaying and Changwu,

for their forever support and trust

To my beloved girlfriend, Yanxu Chen,

for her years' accompanying

Table of Contents

List of Figures

List of Tables

Preface

I wish to express my profound gratitude to my advisor, Professor Lei He. Over the years, his steadfast support, wisdom, and mentorship have not only contributed to my academic accomplishments but also to my personal growth. Professor He has been an unwavering pillar of guidance, always ready to provide crucial insights and clear directions even in the most challenging situations.

My heartfelt thanks go to my thesis committee members, Bolei Zhou and Lin Yang, for their invaluable support and insightful advice throughout my Master's study. I owe a debt of gratitude to Professor Zhou for his sage research guidance, especially when venturing into new areas of study. His acute intuition and incisive insights into research have been immensely helpful. Equally, I am grateful for having been taught by Professor Yang, whose two courses—Matrix Analysis and Stochastic Process—continue to influence my research. His enthusiasm and deep understanding of the fundamental mechanisms underlying various concepts have greatly impacted my academic development. I am truly honored to have had the opportunity to study and work under the guidance of such dedicated professors.

I also wish to acknowledge my fortunate opportunity to collaborate with my esteemed colleagues at Tencent AI Lab during my internships: Dr. Jianhua Yao, Dr. Hanbo Chen, and Dr. Jiangpeng Yan. Their indispensable guidance in histopathological image analysis has been fundamental to my success. I also extend my sincere appreciation to my friends at UCLA: Yuan Liang, Weinan Song, Dezhan Tu, Zhujun Fang, Jiawei Zhang, Zhuofu Tao, Yuwen Jia, Yingqiu Liu, and Sicheng Mo. Their constructive discussions on research and companionship in daily life have been a constant source of support and inspiration.

Finally, I cannot express enough gratitude to my parents, Huaying and Changwu, and my beloved girlfriend, Yanxu Chen. Their unwavering support and unconditional love throughout the years have been my bedrock. Words seem insufficient to convey the depth of my appreciation and love for them.

CHAPTER 1

Introduction

Medical imaging plays a pivotal role in medicine, serving as a primary data source for diagnoses. Physicians heavily rely on these images to aid in the diagnostic process, such as identifying and grading different diseases. Inspecting medical images is an essential step in modern medicine but requires non-trivial expertise and time budgets.

Over the years, deep learning (DL) has emerged as a versatile tool for various domains, healthcare being one of its most significant benefactors [47, 88, 48, 72, 41, 59, 97]. DL based applications are revolutionizing medical image analysis, bringing a new dawn of automation and precision in disease detection, diagnosis, treatment planning, and patient care. This thesis is an effort to contribute to this ongoing revolution by overcoming the key challenges that currently limit the potential of AI in the medical domain.

The principle challenge addressed here is the data-hungry nature of deep learning models. These models demand a vast amount of *annotated* training data to achieve accurate and reliable performance. However, acquiring such large-scale, annotated datasets within the medical domain presents a significant obstacle, given the tedious, time-consuming, and costly nature of medical data annotation. Simultaneously, there exists a treasure trove of unlabeled medical image data, which raises the question of how to efficiently harness this underutilized resource. Effectively exploiting these unlabeled datasets not only compensates for the shortage of annotated data, but it also opens up possibilities for improving the robustness and generalization capabilities of our models. Thus, the primary objective of this thesis is to develop effective and efficient methods for medical image analysis that account for these scenarios. These methods need to effectively utilize limited labeled data and leverage the vast amounts of available unlabeled data. Our aim is to generate solutions that can balance the need for precision and the practical realities of data scarcity in the medical imaging domain.

For the sake of coherence, the research within this thesis is focused on histopathology images, a crucial area of medical image analysis with high clinical value. We seek to address the challenges outlined earlier from two key perspectives: data augmentation and self-supervised learning.

Firstly, *data augmentation* is commonly recognized for its effectiveness in diversifying training samples, thereby assisting the model in avoiding overfitting, while also enhancing the generalizability of deep learning models. Traditionally, data augmentation techniques have been primarily performed in the input RGB space, utilizing either image transformation functions or generative adversarial networks (GANs). Nevertheless, in this thesis, we demonstrate the possibility and the benefits of augmenting data samples directly from the latent space, given that the representations therein are sufficiently powerful. We prove that this form of data augmentation is robust and generalizes well to various settings, including few-shot patch classification and whole-slide image classification. In addition, it is much more efficient compared to augmentation methods that happen in the input RGB space.

Secondly, we advocate for the importance of *representations*. High-quality representations could significantly enhance the performance of downstream tasks. In an attempt to better harness the vast amount of unlabeled data in the medical domain, we explore the impact of different self-supervised learning methods. These methods do not require manually annotated labels and show great promise for various downstream tasks. More specifically, we study two issues: the conditions under which self-supervised learning outperforms supervised learning in medical images, and the strategies for designing superior pre-training methods for tasks beyond classification, such as detection and segmentation. In addressing these issues, we propose different new techniques such as Concept Contrastive Learning (ConCL).

We believe the research presented in this thesis not only contributes to the broader understanding of deep learning applications in healthcare but also provides practical tools and methodologies that can be used to address real-world medical image analysis problems.

1.1 Thesis Outline

This thesis is divided into five chapters, each highlighting different aspects of our research on enhancing the label efficiency and generalizability of deep learning models for medical image analysis, particularly histopathology images. The chapters are as follows:

Chapter 1 introduces the fundamental challenges of data-hungry deep learning models and provides an overview of the specific focus of the thesis, including data augmentation and self-supervised learning. It sets the stage for the research and establishes the context for the ensuing chapters.

Chapter 2 presents our research on a combination of data augmentation and self-supervised learning for few-shot learning in histology images [115]. We elaborate on our approach to incorporate contrastive learning (CL) with latent augmentation (LA) to build an efficient few-shot system. This chapter details our experimental findings, with an emphasis on the generalizability and performance improvements of CL-based models compared to the traditional supervised learning models. It also provides our empirical understanding of when and why CL-based models generalize better than supervised models.

Chapter 3 delves into the challenges posed by whole-slide images (WSIs) for deep multiple instance learning (MIL) and presents our solution, ReMix [116]. WSIs are usually large, up to 10000x10000 pixels, yet of little numbers, making them hard to be processed by DL methods and prone to overfitting. We detail how our proposed ReMix method enhances training efficiency by reducing the number of instances in WSI bags, and ensures data diversity by incorporating bag-level latent augmentations. This chapter also presents the results of applying ReMix to different MIL methods, showing its generality and effectiveness.

Chapter 4 focuses on developing a new self-supervised (SSL) pre-training method for detection and segmentation tasks in computational pathology [114]. We introduce a new SSL framework, Concept Contrastive Learning (ConCL), and present our comprehensive experiments comparing ConCL to previous state-of-the-art SSL methods. This chapter outlines the road map toward a better dense prediction pre-training method and explores the components contributing to its success for pathology images. It ends with our proposed simple, dependency-free, and self-bootstrapping concept-generating method.

Chapter 5, the final chapter, wraps up the thesis by summarizing the findings from our research. It also discusses potential areas for further investigation, the implications of our work, and its potential impact on medical image analysis and healthcare.

CHAPTER 2

Towards Better Few-shot Histopathology Image Classification

2.1 Introduction

Histological images play a crucial role in providing phenotypical and diagnostic information for disease assessment and prognosis [91]. However, building computer-aided histological image classification systems is expensive due to the scarcity of well-annotated data. Additionally, histological images exhibit diverse characteristics, including variations in acquisition protocols, body sites, and tissue types. These significant domain shifts and variations pose challenges in training data-hungry models. Therefore, the key to developing robust diagnosis systems lies in training models with limited annotated samples.

In this chapter, we focus on addressing these challenges through *few-shot learning* (FSL). While FSL has shown success in natural images, its application in histological image analysis remains largely unexplored. To facilitate the study of FSL and generalized FSL (GFSL) in histology images, we set up three cross-domain tasks that involve near-, middle-, and out-domain shifts from base class to novel class. Additionally, we investigate the impact of homogeneous and heterogeneous shot selection, where few-shot samples come from the same whole slide image (WSI) or different ones.

To enable label-efficient learning and improve generalizability, we propose a few-shot system that incorporates *contrastive learning* (CL) with *latent augmentation* (LA). Our approach leverages CL to learn a meaningful encoder during pre-training, while LA transfers semantic variations in latent space from "unlabeled" base datasets. By fully exploiting the base dataset through learned model weights and captured latent variations, our method enables effective few-shot learning.

Interestingly, we observe a larger generalization gap between state-of-the-art CL models and supervised models in histological images compared to natural images. Previous studies on CL primarily focus on "iconic" natural images, where a dominant object occupies the image center. However, histological images contain multiple small objects (e.g., cells, nucleus) and various textures (e.g., muscle, mucus) densely distributed. Thus, they present a unique and relatively unexplored challenge. We aim to fill this gap by studying CL for non-iconic, multi-object, and multi-texture histological images. Furthermore, we provide empirical explanations for the observed generalization gap between CL models and supervised ones in this context.

To summarize, our chapter's key findings and contributions are as follows:

- We explore FSL in histological data, focusing on domain-specific problems.
- We propose a simple label-efficient method for few-shot learning that incorporates contrastive learning and latent augmentation. Through extensive experiments, we demonstrate consistent gains and improve generalizability.
- In contrast to findings in iconic natural images, we show that CL-learned models outperform supervised counterparts by a large margin in histology images. We provide empirical explanations for this observation, contributing to a better understanding of model generalization in the context of representation learning and histology image analysis.

A large portion of this chapter has been published in [115].

2.2 Related Work

Few-shot learning (FSL). FSL has been explored from various perspectives, such as metric-based and optimization-based approaches [33, 77]. This study follows a "pre-training and fine-tuning" methodology in the metric-based domain, where previous research typically learned a shared metric space using standard fully-supervised pre-training [93, 16, 101]. We propose the integration of self-supervised pre-training to enable more efficient label use and demonstrate that it can achieve stronger generalization than supervised pre-training.

FSL in medical images. FSL in medical images is in its early stages, especially in the case of histology images. Mahajan et al. [67] examined FSL methods for skin disease classification, while Chen et al. [17] addressed COVID-19 CT image classification using contrastive pre-training and prototypical network fine-tuning. In terms of histology images, Medela et al. [68] used a triplet loss [80] to pre-train an encoder, followed by a fine-tuned SVM classifier for few-shot domain adaptation. Concurrent to our work, Shakeri et al. [82] also proposed a benchmark for few-shot classification of histological images. Our work explores similar but distinct settings, with broader investigations conducted, such as the GFSL task and hetero-/homo-geneous few-shot selection.

Self-supervised learning. Self-supervised learning aims to develop useful representations without reliance on true labels. Recent leading variants can be classified into contrastive-based learning [20, 14, 42], cluster-based learning [10, 11], and expectation prediction based learning [38, 19]. Most of these studies have focused on pre-training on ImageNet-like images, with recent interest shifting towards images containing multi-objects and multi-textures [15]. We consider histology images as an ideal subject for such study, and demonstrate that contrastive learning can cluster structural "part-whole" information and maintain "global-local consistency", thus enabling better generalization for such data than supervised counterparts.

Representation variation augmentation. The concept of exploiting feature variations has a long history [45, 78]. Recent variants have further refined this idea. For instance, [40] and [81] use a generator to create "hallucinated" novel features from the variation of base samples. This technique is later extended to not rely on base samples [101]. Several other studies [103, 120, 62, 118, 102] have utilized class or intra-class variances to augment

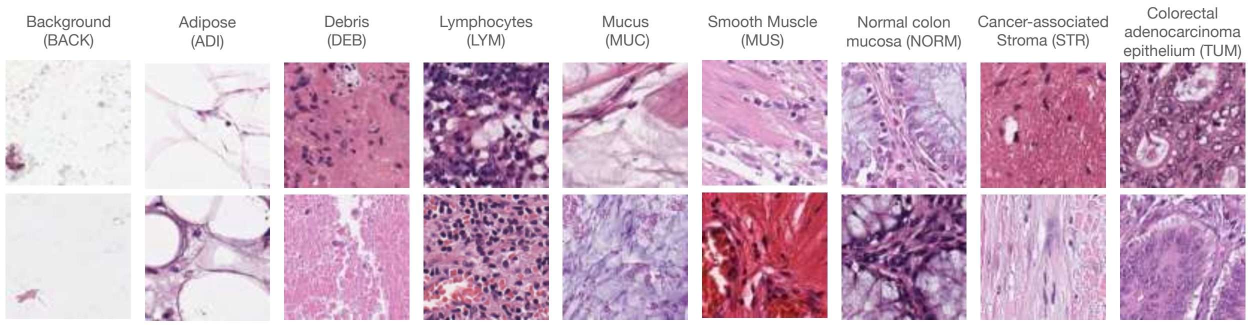


Figure 2.1: **Example images from NCT** [51]**.** Each column contains two samples from the same class (column name).

data for classification, segmentation, and "long-tail" problems. This study follows the line of these works, but instead of relying on label information, we obtain and transfer variations, allowing our method to scale gracefully to other label-hungry problems.

2.3 Preliminaries and Problem Formulation

Whole-Slide Image (WSI). Whole-slide images (WSIs) are digital scans of histology tissue slides obtained through biopsy or surgery. Due to their micron-sized pixels and centimeter-sized slides, WSIs are typically gigapixel in size and are divided into numerous small "patches" for computational analysis. These patches serve as the basic units for patch-level classification. As WSIs can exhibit variations in tissue context and staining quality, the extracted patches retain the styles of their source WSIs, leading to inter-WSI domain shifts. Moreover, unlike iconic natural images that primarily feature a dominant object in their centers, histological patches contain multiple small objects and texture-like tissues. This distinction makes the classification process different from traditional recognition systems that focus on dominant objects.

Few-shot Learning (FSL). Few-shot learning aims to train models using a large "base" dataset and then generalize to unseen classes with limited labeled data. Formally, the base dataset is defined as $\mathcal{D}_{base} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N_{base}} \subset \mathcal{X}_{base} \times \mathcal{Y}_{base}$, where $\mathcal{X}_{base}$ is the sample set, and $\mathcal{Y}_{base}$ is the corresponding label space. The novel dataset $\mathcal{D}_{novel} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N_{novel}}$ has a disjoint label space, i.e., $\mathcal{Y}_{base} \cap \mathcal{Y}_{novel} = \emptyset$, where $\mathcal{Y}_{novel}$ represents the novel label space. A few-shot learner is trained on $\mathcal{D}base$ and evaluated on a series of meta-tasks sampled from $\mathcal{D}_{novel}$. Each meta-task is defined as $\mathcal{T} = \{(\mathcal{S}_i, \mathcal{Q}_i)\}_{i=1}^{I}$, where $\mathcal{S} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{NK} \sim \mathcal{D}_{novel}$ is a small training set, referred to as the support set, and $\mathcal{Q} = \{\mathbf{x}_i\}_{i=1}^{NQ} \sim \mathcal{X}_{novel}$ is a small test set, known as the query set, with I denoting the number of tasks. This formulation represents an N-way K-shot (Q-query) task, where N classes are sampled from $\mathcal{Y}_{novel}$, each with K labeled samples for training and Q unlabeled samples for testing. Typically, K is smaller than Q, for example, $K = 1$ or 5, and $Q = 15$. The evaluation stage is often referred to as the meta-testing stage.

Generalized Few-shot Learning (GFSL). In contrast to FSL, GFSL samples meta-tasks from a joint dataset $\mathcal{D}_{joint} = \mathcal{D}_{base} \cup \mathcal{D}_{novel}$, with a joint label space $\mathcal{Y}_{joint} = \mathcal{Y}_{base} \cup \mathcal{Y}_{novel}$. In GFSL, both the support and query sets contain samples from both seen base classes.

2.4 Methods

Consider a few-shot classifier $f = f_\theta \circ f_\phi$, where f_ϕ is an embedding function, also known as a feature extractor. It maps a high-dimensional input image $\mathbf{x} \in \mathbb{R}^{3HW}$ into a low-dimensional latent space $\mathbb{R}^d$. The classifier f_θ is trained on the support set $\mathcal{S}$ and predicts results for the query set $\mathcal{Q}$. The parameters ϕ and θ correspond to f_ϕ and f_θ, respectively. Our method consists of two phases: a) pre-training f_ϕ on base datasets and b) training f_θ on support sets with latent augmentation during the meta-testing stage. Figure 2.2 provides an overview of our methods. We elaborate on these phases in the following subsections.

2.4.1 Pre-training

Current paradigms in FSL for training f_ϕ lie in two folds: i) meta-training, also known as *episodic* training, where base datasets are divided into various episodic N-way K-shot meta-tasks that simulate meta-learning; and ii) standard training, which involves fully supervised classification pre-training without splitting the data. The former one emphasizes the idea of meta-learning for fast adaption [79, 33], while the latter one attributes the success of FSL to feature reuse [74] or good representations [16, 93]. In this work, we follow the standard training approach and believe that better-learned encoders lead to stronger generalizability.

Fully-supervised pre-training (FSP). We perform joint training of a feature extractor f_ϕ and a proxy classifier f_ψ using the standard cross-entropy loss on a base dataset. After pre-training, only f_ϕ is retained and fixed for downstream tasks. We refer to the embedding functions learned through FSP as f_ϕ^{FSP}.

Contrastive-learning pre-training (CLP). Self-supervised learning methods alleviate the need for data annotation. In this work, we focus on contrastive learning, specifically MoCo-v3 [20], which currently achieves state-of-the-art performance. MoCo-v3 consists of three components: a feature extractor (backbone) f_ϕ, a projection head f_g, and a prediction head f_q. Given an unlabeled base training dataset $\mathcal{D}^u_{base} = \{\mathbf{x}_i\}_{i=1}^{N_{base}}$, the model learns to minimize the contrastive loss function with respect to the unlabeled batch data:

$$\phi^*, g^*, q^* = \arg\min_{\phi,g,q} \mathbb{E}_{\mathbf{x},\mathbf{x}' \overset{t}{\sim} \mathcal{D}^u_{\text{base}}} \left[\mathcal{L}_{\text{CLP}} \left(f_q \circ f_g \circ f_\phi(\mathbf{x}), f_{\tilde{g}} \circ f_{\tilde{\phi}}(\mathbf{x}'); \phi, g, q\right)\right], \tag{2.1}$$

where $\mathcal{L}_{CLP}$ represents the contrastive loss function. $\mathbf{x}$ and $\mathbf{x}'$ are two views of the same image obtained by applying random data augmentation t. $\tilde{\phi}$ and $\tilde{g}$ denote the momentum updated copies of ϕ and g, respectively. In short, contrastive learning aims to maximize the similarity between positive pairs (two augmented views of the same image) while minimizing the similarity between negative pairs (two different images). After CLP, the auxiliary heads f_g and f_q are removed, while f_ϕ is retained and fixed, denoted as f_ϕ^{CLP}.

Figure 2.2: **Overview.** With pre-trained feature extractor (a), N-way K-shot classifiers are learnt (b) based on LA (d) to classify WSI patches (c). Given a novel representation $\mathbf{z}$, LA generates its new features from the most likely variation in the base dictionary, so few-shot novel samples can be proliferated in a reasonable way, and the decision boundary could therefore be improved.

2.4.2 Latent Augmentation

The pre-trained feature extractor f_ϕ only transfers parts of available knowledge in base datasets by reusing the learned weights. The more transferable knowledge is inherent in *data representations*. It is reasonable to assume that base classes and novel classes share similar modes of variations [101] since they are all histology-related. Such inductive biases allow us to transfer variations from seen tissues or styles to unseen ones. Here we propose to transfer the representation variations in a simple *unsupervised* way. Below, we first introduce *latent augmentation* and then discuss our motivations and intuitions about it.

Base dictionary and Latent augmentation (LA). We aim to optimally leverage training data, by both reutilizing pre-trained model weights f_ϕ and enabling the transfer of potential semantic shifts in clustered representations. With an unlabeled base dataset, K-Means is performed on the representations $\mathbf{z} = f_\phi(\mathbf{x})$ to get C clusters (Figure 2.2-(a), red arrows). We construct a *base dictionary*, $\mathcal{B} = \{(\mathbf{c}_i, \mathbf{\Sigma}_i)\}_{i=1}^{C}$, where $\mathbf{c}_i$ and $\mathbf{\Sigma}_i$ represent the i-th cluster prototype and its intra-cluster covariance matrix respectively. In essence, $\mathcal{B}$ encapsulates how f_ϕ envisions base dataset samples would diverge in the latent space for each cluster, using a multivariate Gaussian $\mathcal{N}(\mathbf{c}_i, \mathbf{\Sigma}_i)$. During the meta-testing stage, with the base dictionary $\mathcal{B}$, LA queries the most likely variations from $\mathcal{B}$ using original representations $\mathbf{z}$, leading to additive augmentation $\tilde{\mathbf{z}} = \mathbf{z} + \boldsymbol{\delta}$ (Figure 2.2-(b,d)). Here, $\boldsymbol{\delta}$ is sampled as $\boldsymbol{\delta} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_{i^*})$, where i^* corresponds to the maximum cosine similarity between $\mathbf{z}$ and $\mathbf{c}_i$. The classifier $f\theta$ is trained on both original $\mathbf{z}$ and augmented $\tilde{\mathbf{z}}$ representations (Figure 2.2-(c)).

2.4.2.1 Intuitions and motivations on latent augmentation

Why Variation Transfer Works. LA aims at transferring the knowledge of variations. Such knowledge brings *semantic* diversity from base classes to novel classes. For instance, cancerous cells, derivatives of normal ones, can be simulated through LA given limited cancerous samples, employing variations captured in base dictionaries. This mirrors a pathologist's knowledge expansion from familiar to unfamiliar phenotypes. From an under-representative learning perspective [120], emulating latent variations brings underrepresented distributions nearer to regular ones. In low-data learning scenarios, few-sample distributions are uncalibrated [118]; thus, utilizing base class distributions can potentially rectify novel class calibrations. Moreover, LA serves as a consistency regularization technique, enforcing classifier predictions to remain consistent across varied perturbations, beneficial in low-data circumstances [2, 6, 90]. We later demonstrate in §2.5.3 that LA surpasses data augmentation (DA) significantly, fulfilling DA's role.

Why linear additive augmentation is meaningful. Thoroughly trained deep networks are theorized to excel at linearizing deep features [4, 94], motivating *linear* inter/extrapolation of features, namely, additive generation of new features. Recent studies [24] validate this by investigating universal label-preserving additive augmentations in latent space across varied data modalities, endorsing the efficacy of simple linear transformations.

Why base dictionary construction is warranted in both FSP and CLP. FSP employs the classification task as a surrogate to train effective encoders f_ϕ. During optimization, features are impelled to amplify their dot-product similarity with class weights in f_ψ, thereby constituting a significant metric space. In the context of CLP, contrastive loss—a form of metric-based loss—nudges alike features closer and disparate representations apart, also yielding an informative metric space. Thus, feature distance in the representation space is meaningful for both FSP and CLP, warranting unsupervised clustering to form a base dictionary.

2.5 Experiments

2.5.1 Setup

Datasets. Accounting for tissue variations across body sites, we employ three public histology datasets from different sites to create tasks with diverse domain shifts: NCT-CRC-HE-100K (NCT) from the colon site [51], LC25000 (LC-25K) from the lung and colon [8], and PAIP19 (PAIP) from the liver [52]. NCT comprises 9 classes with a total of 100k distinct patches of size 224×224. LC-25K includes 5 classes, with 5,000 patches in each, sized 768×768. PAIP contains 50 WSIs, each of size about 45k×45k, with 3 annotated mask classes. For LC-25K, patches are resized to 224×224. For PAIP, foreground tissues are cropped into 75k patches of size 224×224 and labeled by majority voting. Novel and base classes from different organs are considered **out-domain**, and those from the same organ are **near-domain** if from the same source; otherwise, they are **middle-domain** due to imaging protocol differences.

Task i) Near-Domain Task (GFSL Study). NCT is split randomly into a training set (80k images) and a test set (20k images) by 80%/20%. The training set undergoes a leave-one-class-out process to create 9 base datasets, and the test set is utilized as $\mathcal{D}_{joint}$ for evaluation, generating 9 sub-tasks, each with a novel class omitted from pre-training datasets.

Task ii) Mixture-Domain Task (FSL Study). NCT's entire training set (80k images) serves as $\mathcal{D}_{base}$, and LC-25K is used as $\mathcal{D}_{novel}$. Two of the five classes in LC-25K are colon-related (middle-domain novel classes), while the remaining three are lung-related (out-domain novel classes).

Task iii) Out-Domain Task (FSL Study). As in the mixture-domain task, NCT's training set is used as $\mathcal{D}_{base}$ and PAIP as $\mathcal{D}_{novel}$. Considering the liver tissues from PAIP differ from colon tissues in NCT, we view them as out-domain novel classes. To examine *heterogeneous* and *homogeneous* shot selection, we use WSI ID to split PAIP into a support WSI set (15 WSIs with 22.5k images) and a query WSI set (35 WSIs with 52.5k images). Evaluation draws support and query samples from their respective WSI sets, with the *heterogeneous* strategy selecting few-shot samples from various support WSIs, and the *homogeneous* strategy selecting from a single randomly chosen support WSI.

Evaluation. Unless specified, the near-domain task evaluates methods over 1000×9 (9 sub-tasks) random meta-tasks; the mixture- and out-domain tasks over 1000 randomly sampled meta-tasks. All meta-tasks use 15 samples per class as the query set. We report the average F1-score and a 95% confidence interval. Given the unequal numbers of base and novel classes, we adopt the convention from GZSL [106] and GFSL [85] to report their average harmonic mean.

Implementations. I. Pre-training. We employ ResNet-18 as the embedding function f_ϕ and use l_2-normalized features for clustering and downstream meta-tasks as in previous FSL studies [93, 16]. **II. Latent Augmentation.** For reproducibility, we apply faiss [50], a clustering library, to execute K-means with a fixed seed. The base dictionary contains 16 prototypes ($C = 16$), discussed in the ablation section §2.5.3). Each sample is augmented 100 times (including the original one) by LA in each meta-task.

Compared methods. Recent studies [16, 93], including a concurrent work on histology image [82], indicate that standard pre-training yields result comparable to complex episodic training. Therefore, we compare methods using standard pre-training:

1. *NearestCentroid*: This method calculates class centroids from support sets and assigns query samples to the nearest centroids, as demonstrated in [100], [89], and [21];
2. *LinearClassifier*: This method trains a new fully-connected layer using different loss functions [16, 57] with respect to support samples or employs direct linear models from scikit-learn [71], such as LogisticRegression [118, 93].

For ease of implementation and consistency, we employ NearestCentroid, along with two

l_2-regularized linear classifiers — LogisticRegression and RidgeClassifier, all from the scikit-learn APIs [71].

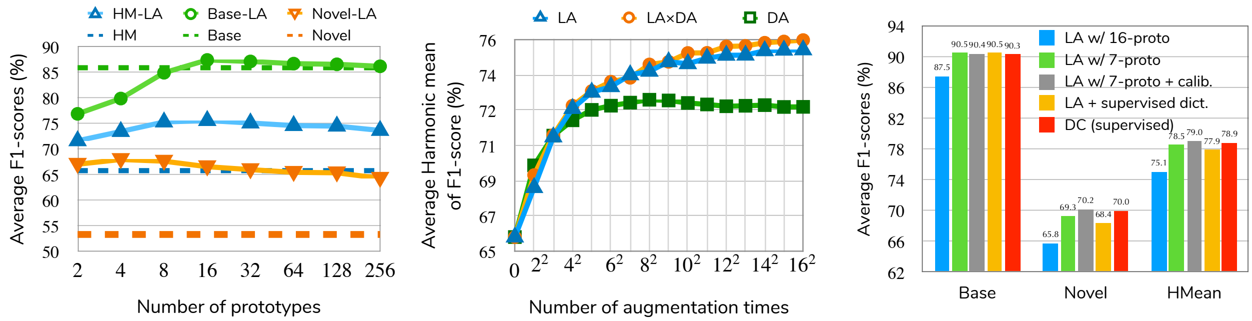
2.5.2 Main Results

Fully-supervised f_ϕ^{FSP} vs. Self-supervised f_ϕ^{CLP}. Table 2.1 reveals that CLP outperforms FSP in generalizing to novel classes significantly. Comparing the best vanilla entries (w/o. LA) using both pre-training methods, CLP exhibits an average improvement in HarmMean of 4%, 5%, and 8% in 1-/5-/10-shot settings for the near-domain task, and 10%, 19%, 16% for the mixture-domain task. Also, CLP representations benefit more from increased shot numbers than FSP's in both tasks, e.g., +17% vs. +11% and +12% vs. +10% when 1-shot escalates to 5-shot for linear classifiers in near- and mixture-domain tasks, respectively. Despite FSP's superior performance in base classes under full supervision, CLP exhibits better generalization to novel classes. Moreover, Table 2.3 confirms CLP's superiority over FSP in the out-domain task with a larger domain shift. This variance between FSP and CLP in histology images somewhat contradicts observations in natural images where they exhibit similar generalizability. We explore and discuss this in §2.5.4.

Latent augmentation yields consistent improvement. Regardless of pre-training methods, LA consistently outperforms baseline linear classifiers, attesting to its effectiveness. With base dictionaries, a limited number of few-shot samples can expand reasonably through the transfer of latent variations. This improvement persists from near-domain to mixture-domain tasks (Table 2.1), though the enhancement becomes less pronounced in the out-domain task (Table 2.3). This is expected, as the three classes defined in PAIP (non-tumor, viable-tumor, and other) are extremely coarse-grained and may encompass several fine-grained classes. As a result, few samples may not sufficiently represent their intricate semantics. This observation doesn't undermine the effectiveness of latent augmentation; instead, it reassures its validity.

Table 2.1: **Main results in near-/mixture-domain tasks.** In near-domain task, the "Base"/"Novel" columns report average F1-scores of the base/novel classes; the "HarmMean" columns report their average harmonic mean. In mixture-domain task, the same metrics are reported w.r.t. middle-domain classes and out-domain classes. "±" numbers denote 95% confidence interval across multiple runs. "LA" denotes latent augmentation. The bold numbers denote the best while the underscored numbers denote the second best.

	1-shot			5-shot			10-shot
9-way-K-shot	Near-domain task						
Methods	Base	Novel	HarmMean	Base	Novel	HarmMean	HarmMean
Fully-supervised pre-training (FSP)							
NearestCentroid	77.38±0.96	**43.80±1.12**	**54.84±1.03**	88.64±0.41	57.67±0.80	68.36±0.53	71.00±0.46
LogisticRegression	75.14±1.03	37.80±1.17	48.84±1.09	88.45±0.40	48.76±0.93	59.99±0.55	66.39±0.45
RidgeClassifier	75.89±1.02	37.55±1.18	48.75±1.09	88.44±0.40	45.73±0.97	56.96±0.57	60.33±0.48
LogisticRegression + LA (ours)	**78.88±0.94**	43.42±1.14	54.83±1.02	**90.85±0.36**	**63.54±0.74**	**73.63±0.48**	**78.14±0.39**
RidgeClassifier + LA (ours)	76.19±1.03	40.71±1.16	51.95±1.07	88.86±0.41	53.90±0.90	64.87±0.55	66.96±0.46
Contrastive-learning pre-training (CLP)							
NearestCentroid	71.45±0.95	51.95±1.03	58.81±0.98	83.11±0.52	65.36±0.80	72.51±0.62	75.18±0.54
LogisticRegression	70.83±1.01	48.76±1.12	56.13±1.06	84.04±0.50	62.69±0.87	70.89±0.62	76.83±0.51
RidgeClassifier	71.24±0.99	49.18±1.12	56.56±1.05	85.89±0.46	66.12±0.83	73.73±0.58	79.45±0.45
LogisticRegression + LA (ours)	72.11±0.95	53.15±1.08	59.82±1.01	**86.43±0.46**	76.68±0.61	80.67±0.51	85.48±0.40
RidgeClassifier + LA (ours)	**72.60±0.99**	**54.50±1.11**	**60.89±1.04**	86.18±0.47	**78.00±0.60**	**81.28±0.51**	**86.17±0.40**
5-way-K-shot	Mixture-domain task						
Methods	Middle	Out	HarmMean	Middle	Out	HarmMean	HarmMean
Fully-supervised pre-training (FSP)							
NearestCentroid	45.65±1.27	**54.94±1.22**	49.87±1.24	49.01±1.05	61.28±0.78	54.56±0.90	55.75±0.84
LogisticRegression	40.07±1.35	48.00±1.44	43.68±1.39	49.42±1.02	54.18±1.04	51.69±1.03	56.12±0.93
RidgeClassifier	41.46±1.36	48.74±1.43	44.81±1.39	55.28±0.98	56.12±1.05	55.70±1.01	60.77±0.88
LogisticRegression + LA (ours)	46.98±1.33	53.34±1.30	**49.95±1.31**	65.51±0.81	**62.64±0.87**	**64.04±0.84**	**67.60±0.73**
RidgeClassifier + LA (ours)	**47.70±1.38**	52.13±1.35	49.82±1.36	**67.45±0.80**	60.97±0.95	64.04±0.86	67.23±0.74
Contrastive-learning pre-training (CLP)							
NearestCentroid	71.42±1.14	52.01±1.05	60.19±1.09	84.50±0.49	65.31±0.71	73.68±0.58	76.30±0.49
LogisticRegression	72.16±1.06	51.14±0.97	59.86±1.01	83.91±0.49	61.98±0.71	71.29±0.58	74.89±0.48
RidgeClassifier	**72.57±1.04**	51.13±0.96	59.99±1.00	85.22±0.43	62.47±0.72	72.09±0.54	75.84±0.46
LogisticRegression + LA (ours)	71.77±1.09	52.73±1.03	60.79±1.06	87.51±0.39	72.92±0.65	79.55±0.48	84.95±0.41
RidgeClassifier + LA (ours)	71.86±1.08	**52.92±1.04**	**60.95±1.06**	**88.55±0.38**	**74.04±0.65**	**80.64±0.48**	**86.32±0.39**



(a) Effect of number of prototypes (b) Effect of number of augmentation times (c) Effect of using labels and calibration

Figure 2.3: **Ablations on latent augmentation.** (a) The effect of varying the number of prototypes. Dashed lines correspond to baselines for the solid lines of matching colors. (b) The effect of the number of augmentation times. The harmonic mean is plotted. "LA$\times$DA" denotes that T latent augmentations are applied after T traditional data augmentations (resulting in T^2 total augmentations). (c) The effect of using labels and calibration. "DC" refers to Distribution Calibration. "calib." refers to calibration.

2.5.3 Ablations

To assess the impact of design choices, we perform ablation studies by excluding two cancer-related classes, specifically cancer-associated stroma (STR) and colorectal adenocarcinoma epithelium (TUM), from NCT, treating them as novel classes, and using the remaining classes as base classes. Unless stated otherwise, all ablations are conducted on CLP models using RidgeClassifier for 300 meta-tasks in a 5-shot setting.

Number of prototypes in base dictionary. Figure 2.3 (a) shows how performance varies with the number of prototypes C. We observe the similar tendency between base class and novel class, where their harmonic means peak at $C = 16$; we subsequently choose $C = 16$ for all experiments. Besides, the performance of base classes and novel classes shows opposite trends from $C = 4$ to $C = 16$. The trade-off exists here that as the granularity of clusters increases ($C\uparrow$), the intra-cluster variance decreases, which results in better grouping accuracy but brings less semantic variation. The novel classes benefit from larger variation while the base classes benefit from more accurately estimated variation since they have been

Table 2.2: **Ablations on covariance type.** See text for more details.

Cov Type	Base	Novel	HMean
None	85.85±0.78	53.27±1.63	65.74±1.06
Tied	79.35±1.08	**65.32±1.21**	71.65±1.14
Diag	<u>85.91±0.88</u>	62.66±1.42	<u>72.46±1.08</u>
Spherical	85.78±0.87	62.00±1.39	71.97±1.07
Full (default)	**87.51±0.80**	<u>65.79±1.36</u>	**75.11±1.01**

Table 2.3: **Results in out-domain tasks.** Average F1-scores from 1000 meta-tasks are reported.

RidgeClassifier	Homogeneous			Heterogeneous		
3-way K-shot	FSP	CLP	CLP+LA	FSP	CLP	CLP+LA
$K=1$	36.90	<u>42.56</u>	**43.14**	/	/	/
$K=5$	39.00	48.91	<u>49.83</u>	43.35	52.25	**53.67**
$K=10$	40.26	50.57	<u>51.62</u>	45.91	55.96	**58.35**
$K=50$	41.53	51.76	<u>53.71</u>	50.54	61.88	**65.38**
$K=100$	41.23	52.74	<u>54.25</u>	52.45	64.03	**67.56**

exposed in training. Nevertheless, LA demonstrates its robustness by consistent improvement over baselines (solid vs. dashed lines of same color in Fig. 2.3-(a)).

DA vs. LA, and number of augmentation times. Here we compare LA with data augmentation (DA), and their combination. Figure 2.3-(b) shows that LA outperforms DA by a large margin. The boost brought by DA saturates easily and keeps dropping thereafter, while LA keeps improving with all tested cases. Besides, DA can marginally improve LA (LA×DA v.s. LA). We conclude that LA has already covered the role played by DA in an implicit way since the most of gains are brought by LA. It is worth emphasizing the computation budget involved in LA (addition in $\mathbb{R}^d$ space) is significantly lower than DA (image augmentation in $\mathbb{R}^{3HW}$ space and encoder forwards). Therefore, we run all experiments only with LA.

Utilizing Label Information. LA constructs the base dictionary without utilizing *any* label information, including the number of classes. However, when label information is available, similar methods like Distribution Calibration (DC) [118] can be employed. Figure 2.3-(c) presents comparisons between using labels and calibration. When supervised, both "DC" and "LA+supervised dict." deliver competitive performance. Interestingly, when provided with the number of base classes, "LA w/ 7-proto" achieves superior results compared to using a 16-prototype dictionary, and it performs comparably to the supervised DC approach. The performance can be further improved with calibration. These results suggest that with LA, simply knowing the number of base classes can be sufficient to achieve results comparable to those obtained when all example labels are known.

Covariance Types. We also explore other types of covariances that LA can use. Specifically, we consider: 1) "Tied", where all clusters share a covariance matrix estimated from the entire base dataset, 2) "Diag", where each cluster has its own diagonal covariance matrix, *i.e.*, diagonal elements are a variance vector and non-diagonal elements are zeros, 3) "Spherical", where each cluster has its own single variance scalar shared by all feature dimensions. The results shown in Table 2.2 demonstrate that LA improves performance with all types of covariances. This emphasizes the importance of diversifying few-shot samples with variation. Using a full covariance estimation provides the best performance.

Heterogeneous vs. Homogeneous Patch Selection. We examine the heterogeneous and homogeneous patch selection strategies defined in the out-domain task (§2.5.1). Table 2.3 presents the results. Two key observations can be made: i) Heterogeneous selection consistently provides higher baseline performance compared to homogeneous selection; and ii) LA contributes more significantly to improvements in heterogeneous selection. This indicates that heterogeneous patches offer reliable and diverse "anchor" samples compared to homogeneous patches, which can thus benefit more from leveraging the base dictionary.

2.5.4 More discussion

Disparity between f_ϕ^{CLP} and f_ϕ^{FSP} Influences the Choice of Base Learner. From Table 2.1, we note that i) the strongest baselines for CLP and FSP can vary, and ii) the simple NearestCentroid model can sometimes surpass the performance of vanilla l_2-regularized linear classifiers for FSP. Here, we briefly provide some insights on these observations.

The representations generated by CLP can have different distributions compared to those

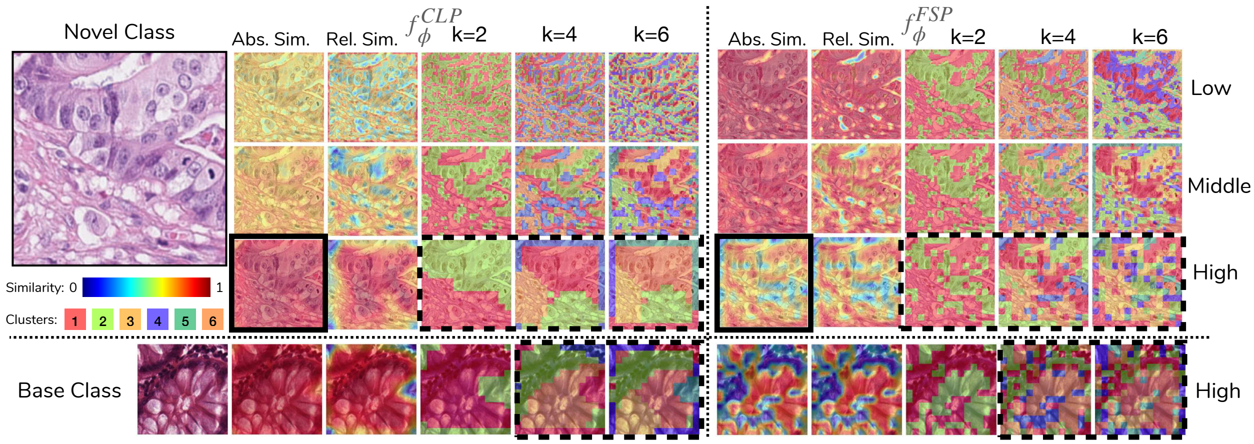


Figure 2.4: **Visualization of samples learned by CLP and FSP.** "Abs./Rel. Sim." columns show the absolute/relative cosine similarity between the global feature and the local ones. Relative similarity is the min-max normalized absolute similarity. k indicate the cluster numbers used by K-Means. "Low", "middle" and "high" denote using features from stage-3, 4, and 5 from a ResNet. See text in §2.5.4 for discussion.

of FSP, as has also been noted in [42]. Given a limited number of training samples, different classifiers can form their own biases when building decision boundaries, which subsequently leads to differing degrees of generalizability.

Furthermore, no regularization techniques are employed during FSP [14], for instance, weight decay [55], DropBlock [57, 93], or "distill" regularization [93]. Although the linear classifiers incorporate an l_2 penalty, they may still overfit in such a representation space when only a limited number of samples are available. Consequently, the simplest NearestCentroid model, which possesses the least complexity, can yield better results than these overfitted linear models.

Why do CLP Models Generalize Better than FSP Ones in Histology Images? In an attempt to understand why such a significant generalization gap exists, we followed the methodology from [15] to examine how features cluster in space. Specifically, we visualized the cosine similarity between a feature map (a set of local representations) and its global average (global representation). We also performed K-Means on the feature maps from different layers (i.e., stages 3, 4, and 5 of ResNet) with varying numbers of clusters.

From Figure 2.4, we observed that the FSP model maintained a high degree of global-local similarity in the lower and middle levels, while the CLP model retained this high similarity at a higher level (solid boxes). Furthermore, the CLP model extracted low-level features related to edges and subsequently aggregated adjacent similar structures (dashed boxes). In contrast, the FSP model was able to differentiate nuclei at lower and middle levels but failed to encode structure-related features in deeper layers.

These findings deviate from those found in ImageNet-like images [15], where FSP and CLP displayed no difference across layers[1]

As observed from further visualizations of base class samples (see bottom of Figures 2.4), the disparity between FSP and CLP is not limited to previously unseen classes, but also present in seen classes.

In the bottom row of Figure 2.4, it's observed that FSP focuses primarily on the most discriminative parts, leaving the remaining "redundant parts" disordered (indicated by the dashed box). However, when a new class is introduced, the discriminative parts are likely to change. FSP's inability to encode all relevant information could be responsible for its struggle to generalize to new classes.

Meanwhile, CLP captures most tissue-structure-related features, which are potentially useful for recognizing novel classes, possibly leading to better generalizability. However, it is interesting to note that FSP and CLP models exhibit similar behavior in ImageNet dataset under the same visualization process (a comparison can be found on the website).

So, why does this disparity exist? The ImageNet dataset is more diverse with 1000 classes and approximately 1.28M images, compared to histology datasets. FSP models in ImageNet need to recognize the discriminative parts of all 1000 classes. In such a case, redundant information in one class might aid in the recognition of another class. Hence, FSP may eventually encode most of the available information useful for new classes related

[1] See [15] or `https://contrastive-learning.github.io/intriguing/` for a comparison.

to ImageNet classes.

However, histology datasets usually lack a diverse range of annotated classes that would aid the development of a comprehensive FSP model. An intriguing question for future work could be whether CLP always outperforms FSP in terms of generalization when pre-training on a base dataset with a limited number of annotated classes, and whether the generalization gap would increase as label diversity decreases.

Yet, the visualization results and the significant generalization gap demonstrated in our work remain empirical observations. Our discussion attempts to unravel possible reasons behind them. We hope our work will contribute to the further development of representation learning, histology image analysis, and beyond.

2.6 Conclusion

In this work, we have conducted an initial investigation into the problem of few-shot learning for histology images. We've integrated contrastive learning and latent augmentation to fully harness training data in an unsupervised manner. This approach allows our method to elegantly scale to other large problems requiring abundant labels. Importantly, our study demonstrates that the generalization gap between state-of-the-art contrastive learning pre-training methods and supervised pre-training in histology images is larger than in ImageNet experiments. We analyze the possible reasons behind this and provide our empirical understanding.

CHAPTER 3

Towards Label and Computation Efficient Training for Multiple Instance Learning of Whole Slide Image Classification

3.1 Introduction

Whole-slide pathological images (WSIs) offer critical insights for disease diagnosis and assessment, yet their analysis demands substantial expertise and time [91]. Deep learning (DL) has significantly contributed to enhancing the efficiency of WSI diagnostic systems [47, 88, 48, 72, 41, 59, 97]. However, the successful application of DL depends on massive datasets and diverse training samples, necessitating efficient pipelines for large datasets and diversification techniques such as data augmentations. WSIs present unique challenges in these respects due to their massive size and lack of diversity.

WSIs, with up to $100k \times 100k$ pixels, are difficult to process with DL models [12]. Despite efforts to address them in an end-to-end manner at the cost of 300 GB or more memory [12], a more feasible solution is to divide each WSI into equal-sized "patches" or "tiles" and sort for weakly supervised multiple instance learning (MIL) methods [110, 63, 83, 22]. Within the context of MIL, a WSI with extracted patches is considered a *bag* with multiple *instances*. We refer to the number of instances of one bag as its *bag size*. The bag size usually varies strikingly in practice; *e.g.*, the Camelyon16 [3] dataset has an average bag size of $8k$ at $20\times$ magnification (a commonly used magnification), with the largest bag size surpassing $50k$. The varying bag size would lead to an unbalanced input/output (I/O) stream and make the parallelization hard since bags of different sizes cannot be directly composed into a batch. Overall, the conventional MIL-based WSI classification pipeline is memory-expensive (large bags), I/O unstable (varying bag sizes), and computation-inefficient (small batch size). These problems can hinder current MIL methods from scaling to giant datasets.

Moreover, while WSIs may contain a large number of training patches, the data is often repetitive and lacks diversity. Enhancing data diversity is crucial since DL models perform better with more diverse labeled data. Current augmentation methods are inefficient for WSIs, given that augmenting a single WSI requires tens of thousands of transformations or new patches, leading to longer training periods [87].

In response, we introduce `ReMix`, a general, efficient, and effective MIL-based WSI classification framework. `ReMix` reduces the bag sizes significantly by using clustered instance prototypes to represent a WSI. Then, it applies a novel data augmentation method, "Mix," which introduces online, stochastic, and flexible latent space augmentations. This method combines different bags by appending, replacing, interpolating instance prototypes, or transferring semantic variations among different bags, thereby enforcing the model to learn perturbation-invariant class-related features.

The proposed `ReMix` framework, despite its simplicity, is highly effective and can be integrated with various state-of-the-art MIL methods for WSI classification to enhance their performance. Empirical evaluations on two public and one in-house dataset reveal that `ReMix` consistently improves generalization performance and reduces the training cost.

- We propose a general, simple yet flexible, and effective method to improve the training efficiency of the MIL framework for WSI classification.
- We propose a novel and efficient latent augmentation method for MIL-based WSI classification, an area yet unexplored.
- We significantly enhance the performance of existing state-of-the-art MIL methods, reducing the costs considerably.

A large portion of this chapter has been published in [116].

3.2 Related Work

3.2.1 MIL in WSI analysis

Multiple instance learning (MIL) is a viable approach to address the weak supervision issue inherent in WSI classification. Due to the large size of WSIs, most studies opt for two-stage learning, training a *patch encoder* to map tissue patches (tiles) to feature vectors, and then a *MIL learner* aggregates all feature embeddings using various mechanisms such as max-pooling and attention-based pooling [49, 58, 83]. These aggregated representations, or bag representations, are used for final predictions. There are several strategies for patch encoder learning. The first, SimpleMIL [23], treats all instances in a WSI as sharing the bag-level label and trains a classification model accordingly. Recent methods have also found self-supervised learning effective for pre-training patch encoders [58, 115, 112].

3.2.2 Clustering in WSI analysis

End-to-end training of patch encoder and MIL learner is enabled by [107], proposing to randomly split WSI patches into k groups and select representative patches from each for training. Centroids are recomputed and samples reassigned at each epoch's beginning. [84] follows a similar process but utilizes K-Means clustering. However, clustering in every epoch for every slide, as done in [107] and [84], can be exceedingly time-consuming. Various sophisticated sampling strategies have been developed to alleviate this [107]. [119] clusters over extracted features based on an ImageNet-pre-trained encoder to define phenotype groups and sample patches for MIL training. In contrast, our work performs clustering once post pre-training and uses cluster centroid vectors as input, and further leverages cluster covariance in the "Mix" step to better capture a cluster's distribution.

3.2.3 Data augmentation

Data augmentation is crucial for deep learning when training samples are scarce, a common situation in medical imaging. While widely studied for natural images [7, 6, 90, 28, 29, 30, 122, 121], and medical images [34, 86, 31, 111, 123, 117, 115], most approaches augment samples in the input space, posing an efficiency challenge for gigapixel WSIs. The work of [115] is closely related to ours, applying latent space augmentation for few-shot patch classification for WSIs. However, it is limited to instance-level patches and lacks bag-level slide augmentations. We introduce more latent space augmentations applicable to WSIs. To our knowledge, our work is among the first to explore bag-level augmentations for WSI analysis.

3.3 Method

In this section, we first introduce the preliminary knowledge of MIL in Section 3.3.1, then elaborate on the detailed steps of `ReMix` in Section 3.3.2, introduce a straightforward extension of `ReMix` applied to spatial-aware MIL methods in Section 3.3.2.4, and finally discuss some intuitions on the effectiveness of `ReMix` in Section 3.3.3.

3.3.1 Preliminary: MIL Formulation

Multiple instance learning (MIL) aims to address the weakly supervised classification problem. Under the MIL setting, a dataset that has N bags is formulated as $\mathcal{D} = \{(B_i, y_i)\}_{i=1}^{N}$, where $B_i = \{x_j\}_{j=1}^{N_i}$ denotes the i-th bag that has N_i instances, and y_i is the bag label. In WSI classification, each WSI corresponds to a bag, and all patches extracted from it are regarded as its instances. The average bag size varies across datasets and patch magnifications. For example, the average bag size of the Camelyon16 dataset [3] under $20\times$ magnification is about $8k$, while the largest bag size is around $50k$.

This work focuses on the *spatial-agnostic* MIL methods that do not rely on the spatial relationship between instances to make predictions. A general *spatial-agnostic* embedding-

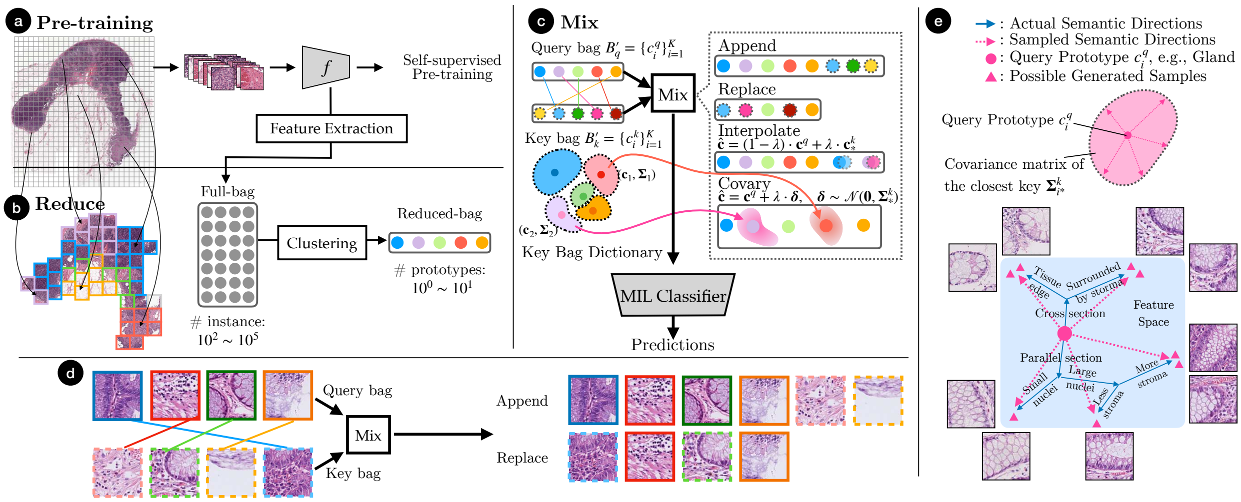


Figure 3.1: `ReMix`**'s overview.** (a) Patch encoder pre-training. (b) Reduce the number of instances by substituting them with prototypes (right); several patches can abstract a large-size whole slide image (left). (c) Mix the bags by appending, replacing, interpolating prototypes, or transferring intra-cluster covariance from other WSIs. (d) A visual illustration of append-augmentation and replace-augmentation. (e) A visual illustration of covary-augmentation.

based MIL classification process can be expressed as

$$\hat{y}_i = g\left(P(f(x_1), ..., f(x_{N_i}))\right), \tag{3.1}$$

where $f(\cdot)$ is a *patch encoder*, $g(\cdot)$ is a *MIL learner* that aggregates information and makes final predictions, and P denotes a permutation operator. The notation of P is only used to mark the permutation-invariance property of a spatial-agnostic MIL classifier.

We also introduce a straightforward extension of `ReMix` to *spatial-aware* MIL methods in Section 3.3.2.4 and show its effectiveness in Section 3.4.5.

3.3.2 ReMix

3.3.2.1 Overview

Figure 3.1 illustrates the `ReMix` approach. Initially, (a) we train a patch encoder using self-supervised contrastive learning. Following that, (b) we assemble reduced bags using cluster prototypes and gather their covariance matrices. Finally, (c) we utilize the "mix" augmentation shown in (d, e) for MIL training.

3.3.2.2 Patch encoder pre-training

The weak supervision nature of WSI classification challenges the training of patch encoders due to a lack of adequate patch-level labels. Conventional end-to-end training methods that utilize all patches are often costly [9, 58], inefficient, and sometimes unfeasible. Therefore, we adhere to the common two-stage training scheme, where a patch encoder $f(\cdot)$ is initially trained, and subsequently, a MIL learner $g(\cdot)$ is trained on the extracted features. Typically, a pre-trained encoder is used [58, 119], such as an ImageNet-supervised pre-trained encoder. Several works [47, 9, 41, 13] follow SimpleMIL [23] to train a patch encoder based on noisy labels, where the bag labels are assigned to all instances within the bags. Patch classification is then conducted using these pseudo labels. Despite its popularity, a recent study shows that its success is significantly associated with the proportion of label-related patches [58]. However, as we will demonstrate later in the experiments, this type of pre-training eventually fails in one of our studied datasets, suggesting its use cases are limited.

Self-supervised learning methods such as SimCLR [14] and MoCo [42] produce effective representations by maximizing the similarity between two different augmented views of the same patch, and minimizing it between views from different patches. Recent studies have recognized the superiority of self-supervised pre-training on large-scale and imbalanced WSI patches over other methods [58, 26, 115]. Additionally, self-supervised pre-training that doesn't depend on class label information is preferable for label-hungry WSI problems. We follow [58] to use a state-of-the-art self-supervised learning method – SimCLR [14] for patch encoder pre-training. It is important to note that the choice of patch-encoder is relatively orthogonal to our `ReMix` framework and downstream MIL learner's training. For the sake of completeness, we briefly discuss available pre-training methods here; however, their training budgets are not considered in this work.

3.3.2.3 Reduce

Conventionally, all patches extracted from a WSI are assembled as a bag for downstream MIL classification [58, 47]. However, the bag size fluctuates from bag to bag in a range from $N_i = 500$ to $N_i = 50,000$, depending on whether it is from the needle biopsies or large/small excisions. On the one hand, the large bags could lead to high I/O costs and high memory consumption during training. On the other hand, the strikingly varying bag sizes could make the I/O stream unstable and the training inefficient.

To tackle them, `ReMix` reduces the bag size via clustering. Stemmed from the nature of WSIs that a large portion of tissue patches could be repetitive and redundant, we propose substituting instances with instance prototypes. Specifically, for each bag, we perform K-Means clustering on patches' representations to obtain K clusters and use their prototypes (centroids) to represent the bag:

$$B_i' = \{\mathbf{c}_k\}_{k=1}^{K}, \text{where } \mathbf{c}_k = \frac{1}{N_k}\sum_{i=1}^{N_k} f(x_i) \tag{3.2}$$

B' is referred to as *reduced-bag* and $\mathbf{c}_k$ corresponds to the k-th prototype. A WSI thumbnail in Figure 3.1-(b) depicts how several patches (reduced-bag) can provide sufficient information of the entire WSI (full-bag) for certain downstream tasks. The reduced-bag (the leftmost of (b)) contains less than 1% number of patches compared to the full-bag (the leftmost of (a)). Informally and visually, we can see they contain almost the same information since all the representative patches are preserved. The reduced-bag can be seen as a denoised abstraction of the full-bag.

To further exploit WSI information, inspired by [115], we construct a *bag dictionary* as $\Phi_i = \{(\mathbf{c}_k, \mathbf{\Sigma}_k)\}_{k=1}^{K}$ for each bag, where $\mathbf{\Sigma}_k$ corresponds to the intra-cluster covariance matrix of the k-th cluster:

$$\mathbf{\Sigma}_k = \frac{1}{N_k - 1}\sum_{i=1}^{N_k} \left(f(x_i) - \mathbf{c}_k\right)\left(f(x_i) - \mathbf{c}_k\right)^T \tag{3.3}$$

A bag dictionary captures how its instances distribute at a high level by modeling a multivariate Gaussian distribution $\mathcal{N}(\mathbf{c}_k, \mathbf{\Sigma}_k)$. Besides, the covariance can manifest the semantic directions inherent in each cluster, *e.g.*, how features vary in that cluster. Therefore, adding semantic translation vectors sampled from the covariance matrix could produce meaningful features. Circles with dashed boundaries in Figure 3.1-(c, e) illustrate the covariance of different clusters.

3.3.2.4 Mix

Data diversity, the second challenge we outlined in the introduction, is a major hurdle for deep learning models. These models tend to overfit when there are limited labeled training samples. Data augmentation can provide additional artificial data [24, 35, 87, 90] and enhance data diversity. Simple solutions such as applying image processing functions, for instance, cropping, flipping, or color jittering, are beneficial for typical-sized image recognition but can be highly inefficient for WSIs, given their large sizes and resolution. More advanced augmentations involve training distinct GANs for different classes to generate new training samples. However, training GANs demands non-trivial resources and hasn't been well-established for WSIs due to issues of tissue imbalance, weak supervision, and more. Neither of these solutions can be directly applied to WSI classification without careful consideration.

Rather than conducting augmentation in the input RGB space, `ReMix` applies efficient latent space augmentation by mixing bags. We propose a "mix" augmentation, as depicted in Figure 3.1-(c). After the "reduce" step, we consider the bag as now being composed of fundamental semantic prototypes; some are class-related, and others are complementary "contexts" that don't impact final decisions. Bags from the same class are likely to share similar fundamental semantic prototypes. As a result, a prototype in one bag could find a similar "cousin" prototype in another bag. The "cousin prototype" is the most similar prototype from another bag. Guided by this, we mix only the "cousins". In doing so, the risk of losing the original class identity after augmentation can be significantly reduced. Moreover, beyond cluster prototypes, the intra-cluster covariance also provides rich information - it reflects the semantic directions in each cluster. Figure 3.1-(e) demonstrates that translating a given prototype by expressive semantic directions can generate meaningful samples. It's important to note that while we can only illustrate the augmentation using simple patch instances, the clusters in practice can contain more complex semantic information.

Specifically, we design four different "mix" augmentations: append, replace, interpolate and covary. When a bag is fed into a MIL classifier, we randomly sample another bag from the same class and "mix" them. Without loss of generality, we define the former bag as a query bag $B'_q = \{\mathbf{c}_i^q\}_{i=1}^K$, and the latter bag as a key bag $B'_k = \{\mathbf{c}_i^k\}_{i=1}^K$. Their instances $\mathbf{c}^q$ and $\mathbf{c}^k$ are called query prototypes and key prototypes. For each query prototype $\mathbf{c}_i^q$, we find its closest key prototype $\mathbf{c}_{i^*}^k$, and then augment the query bag with an applying probability of p by one of the following four augmentations:

- **Append:** append the closest key prototype $\mathbf{c}_{i^*}^k$ to query bag B'_q: $B'_q = \{\mathbf{c}_1^q, ..., \mathbf{c}_i^q, ..., \mathbf{c}_{i^*}^k\}$.
- **Replace:** replace the query prototype $\mathbf{c}_i^q$ with its closest key prototype $\mathbf{c}_{i^*}^k$: $B'_q = \{\mathbf{c}_1^q, ..., \mathbf{c}_{i^*}^k, ...\}$.
- **Interpolate:** append an interpolated feature

$$\hat{\mathbf{c}}_i = (1 - \lambda) \cdot \mathbf{c}_i^q + \lambda \cdot \mathbf{c}_{i^*}^k \tag{3.4}$$

  to the query bag B_q, where λ is a strength hyper-parameter: $B'_q = \{\mathbf{c}_1^q, ..., \mathbf{c}_i^q, ..., \hat{\mathbf{c}}_i\}$.
- **Covary:** generate a new feature from the key covariance matrix by

$$\hat{\mathbf{c}}_i = \mathbf{c}_i^q + \lambda \cdot \boldsymbol{\delta}, \quad \boldsymbol{\delta} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{i^*}^k) \tag{3.5}$$

  and append it to the bag B_q, where λ is a strength hyper-parameter and $\boldsymbol{\Sigma}_{i^*}^k$ is the covariance matrix corresponding to the closest key prototype $\mathbf{c}_{i^*}^k$: $B'_q = \{\mathbf{c}_1^q, ..., \mathbf{c}_i^q, ..., \hat{\mathbf{c}}_i\}$.

In addition to four individual augmentations, we propose to combine them sequentially as a "joint" augmentation.

- **Joint:** Apply "append", "replace", "interpolate", and "covary" with independent probability p.

Figure 3.1-(d,e) illustrate how "append", "replace", and "covary" augmentation would behave visually. It is important to sample another bag from the same class and mix the query prototype with the most similar key prototype since it helps preserve critical class-related information and reduces the risk of losing the original class identity. The above procedures are applied in the reduced-bag and via simple operations such as appending or numerical addition, which are highly efficient.

In `ReMix`, we primarily aim to apply it to spatial-agnostic MIL models. However, it can be easily extended to spatial-aware MIL methods as well. Here's a simple extension approach, although there could be more sophisticated ones.

To ensure minimum modification and not disrupt the design principles of spatial-aware MIL methods, we make two changes to `ReMix`. First, we use the full-bag representation directly for training. For the "interpolate" and "covary" mix augmentations, we use the original full bags as query bags and the reduced-bags as key bags. This means that the representations of patches in full-query-bags are combined with the prototypes' representations in reduced-key-bags. This approach reduces the time complexity of computing the pairwise similarity among instances from $\mathcal{O}(N^2)$ to $\mathcal{O}(N)$, where N is the full-bag size. This is beneficial in practice, as N can range from hundreds to tens of thousands. The second change is to replace original patches with generated features rather than appending them. This ensures that the spatial information is not altered.

This process is a variant of `ReMix`, as it necessitates both the "reduce" (building reduced-key-bags) and "mix" (interpolating instance features) steps.

3.3.3 Intuitions on `ReMix`'s Effectiveness

3.3.3.1 Implicit data re-balance behavior

Tissue imbalance is a typical property of WSIs. Most similar patches almost convey the same information about a WSI but could dominate in numbers over other distinct minority patches. Using the representative prototypes for bag representation can be seen as an implicit data re-balance mechanism that bridges the gap between the majority and minority numbers. It alleviates the tissue imbalance issue to some extent. Besides, using the mean embedding of a group of similar patches could obtain a more accurate and less noisy tissue representation.

3.3.3.2 Efficient semantic consistency regularization

Consistency regularization underlies many successful works, such as semi-supervised learning [7, 6, 90]. Usually, consistency regularization enforces models' predictions to be invariant under different data augmentations [90]. Instead of augmenting instances using image processing functions in the input RGB space, `ReMix` augments the bags by bringing various semantic changes in the latent space. Guided by bag labels and prototypes similarity, such changes are class-identity-preserving. The bag instance combination is no longer static and unaltered but diverse and dynamic, *i.e.*, different new bags can be fed into the MIL classifier every time. "Mix" can be seen as an efficient semantic consistency regularization method that enforces semantic-perturbation-invariant and is computational friendly.

3.3.3.3 Why clustering and additive latent augmentation work

When learned properly, the deep representation space is shown to be highly linearized [4, 94]. Consequently, the distance metrics could demonstrate the similarity between patches, making clustering meaningful. Moreover, in such a space, linear transformation, *e.g.*, interpolating features or adding semantic translation vector $\boldsymbol{\delta}$, is likely to provide plausible representations [24]. The mixed bag representations can serve as hard examples that help models generalize better [56, 126, 104].

3.3.4 Datasets and Metrics

3.3.4.1 UniToPatho

UniToPatho [5] is a public dataset comprising 9536 hematoxylin and eosin (H&E) stained large-size images extracted from 292 WSIs. The slides are scanned at 20× magnification

(0.4415 μm/px). There are six classes in this dataset, *i.e.*, Normal tissue (NORM), Hyperplastic Polyp (HP), Tubular Adenoma with High-Grade dysplasia (TA.HG), and Low-Grade dysplasia (TA.LG), and Tubulo-Villous Adenoma with High-Grade dysplasia (TVA.HG), and Low-Grade dysplasia (TVA.LG). We use the official split of 204/88 slides for training/testing. This dataset provides large-size images extracted at $\sigma = 800$ (1812×1812 pixels patches) and $\sigma = 7000$ ($15,855 \times 15,855$ pixels patches), where σ denotes the physical pixel size in μm. We refer readers to [5] for more details.

Formally, we refer to the two variants of the UniToPatho dataset as **UniToPatho800** and **UniToPatho7000**. For patch processing, we divide the provided large images to 224×224 pixels. Patches with average saturation lower than 30 are considered as background and dropped. Under this setting, for UniToPatho800, the average bag size is about $1.6k$, with the largest bag size surpassing $20k$. UniToPatho7000's average bag size is about $4.9k$, and the largest bag size is $59k$. Overall, UniToPatho7000 has larger bags and more noise. Classification tasks in UniToPatho800 to UniToPatho7000 provide a smooth increment in recognition difficulty.

3.3.4.2 Camelyon16

Camelyon16 [3] is a publicly available dataset consisting of 400 H&E stained slides from breast cancer screening. It contains two classes, *i.e.*, normal and tumor. For this dataset, we directly use the pre-computed features provided by DSMIL [58] without further processing. Each feature vector is fused by features from $20\times$ and $5\times$ magnifications. We refer readers to [58] for more details. There are 271/129 slides in the training/testing set. The average bag size is about $8k$, with the largest bag size surpassing $50k$.

3.3.4.3 Colon10

Colon10 is an in-house dataset comprising 100 H&E WSIs of colon polyps obtained from 100 patients. A collaborating hospital provides the data. It has 10 classes, *i.e.*, Hyperplasia/normal, Adenoma, Villous adenocarcinoma, Tubulovillous adenoma, High-Grade dysplasia, Adenocarcinoma, Carcinoma in situ, Intramucosal carcinoma, Mucinous adenocarcinomas, and Signet ring cell carcinoma. WSIs are acquired under 40× magnitude with 0.23 μm per pixel. We downsample the images to 20× magnitude for analysis. A sliding window of size 224 × 224 without overlap is adopted to crop foreground patches. Patches with average saturation lower than 30 are considered as background and dropped. There are 100 slides in total and 10 slides for each class. We randomly divide them into 70/30 slides to build training/test sets. The average bag size is around 12.2k, while the largest bag size is 41k.

3.3.4.4 NCT-CRC

To study the robustness of `ReMix` to the choice of patch encoder, we also pre-train patch encoders on the NCT-CRC-HE-100K dataset [51], referred to as the NCT dataset. It contains 100,000 non-overlapping patches extracted from H&E-stained colorectal cancer and normal tissues. All images are of size 224 × 224 at 0.5 μm per pixel (20× magnification). This dataset has 9 classes, and the class distribution is roughly balanced. We randomly choose 80% of NCT to be the pre-training dataset.

3.3.4.5 Metrics

We report class-wise averaged precision, recall, accuracy, and their average. To alleviate the issue of randomness, we run all experiments 10 times and report the mean performance.

3.3.5 Implementation Details

3.3.5.1 Patch encoder

We follow SimCLR [14] to pre-train ResNet-18 encoders on the UniToPatho800 dataset [5], the in-house Colon10 dataset, and the NCT dataset [51] respectively. For the Camelyon16 dataset [3], we use the pre-computed features provided by [58]. We use the codebase of

OpenSelfSup[1] [27] for pre-training. The following elaborates on each component in pre-training.

1. **Architecture:** we use ResNet-18 as the backbone, a two-layer non-linear projection head [14] for contrast, and a temperature parameter of 0.1.

2. **Normalization:** we use ImageNet normalization parameters, *i.e.*, mean=(0.485, 0.456, 0.406) and std=(0.229, 0.224, 0.225).

3. **Augmentation for pre-training:** we use the default augmentation settings in the repository, *i.e.*, RandomResizedCrop to 224×224, RandomHorizontalFlip, ColorJitter in the ranges of brightness 0.8, contrast 0.8, saturation 0.8, and hue 0.2 with a probability of 0.8, RandomGrayscale at a probability of 0.2, GaussianBlur with $\sigma_{min} = 0.1$ and $\sigma_{max} = 2.0$ at a probability of 0.5.

4. **Optimizer:** we use the LARS optimizer with an initial learning rate of 0.6, a weight decay of 1e-6, and a momentum of 0.9.

5. **Schedule:** we use a CosineAnnealing learning rate scheduler with a 10-epoch warm-up.

6. **Training:** we pre-train the encoder for 200 epochs and use the last model for downstream tasks. The batch size for training is 512.

3.3.5.2 MIL models

To demonstrate that `ReMix` can be MIL model-agnostic, we use two previous state-of-the-art deep MIL models, ABMIL [49] and DSMIL [58], for our experiments. ABMIL and DSMIL are attention-based MIL methods that compute the attention-weighted sum of instance features as the bag representation. They differ in the way of attention computing. ABMIL [49] predicts the attention score of each patch using a multi-layer perceptron (MLP) without explicit patch relation modeling. DSMIL [58] is a dual-stream method that comprises an

[1] The codebase's name has changed from OpenSelfSup to mmselfsup.

Table 3.1: **Main results.** "Pre", "Rec", "Acc", and "Avg" denote precision, recall, accuracy, and their average, respectively. Bold and underlined numbers are the first and second best entries among the row sections. All results are averaged over 10 independent runs. Numbers are shown in percentage (%). "no aug." means no augmentation. The "best improvement "Δ" reports the best gain of `ReMix` from the corresponding methods trained on full-bags.

	UniToPatho800				Unitopatho7000				Camelyon16				Colon10			
Methods\Metrics	Pre	Rec	Acc	Avg	Pre	Rec	Acc	Avg	Pre	Rec	Acc	Avg	Pre	Rec	Acc	Avg
TransMIL [83]	58.75	56.14	68.52	61.14	54.18	42.78	60.54	52.50	88.27	85.98	87.91	87.39	63.28	62.67	62.67	62.87
CLAM [66]	60.19	71.98	63.29	65.16	54.18	50.84	61.22	55.39	91.63	93.22	92.48	92.44	60.58	60.67	60.67	60.64
ABMIL [49]	56.18	58.50	60.11	58.26	48.69	58.55	56.62	54.62	92.47	92.79	93.02	92.76	71.42	63.00	63.00	65.81
DSMIL [58]	72.92	79.41	76.36	76.23	59.90	63.41	63.38	62.23	94.37	93.39	94.11	93.96	58.33	59.33	59.33	59.00
ReMix-ABMIL (no aug.)	69.93	72.85	68.75	70.51	57.66	58.89	61.35	59.30	93.97	93.15	93.95	93.69	74.19	68.33	68.33	70.28
ReMix-ABMIL (append)	71.81	74.54	69.09	71.81	64.77	61.81	58.24	61.61	94.59	93.38	94.34	94.10	77.23	**74.33**	**74.33**	75.30
ReMix-ABMIL (replace)	70.16	74.34	68.75	71.08	63.82	61.40	55.27	60.17	94.60	93.52	94.42	94.18	74.72	73.67	73.67	74.02
ReMix-ABMIL (interpolate)	71.55	75.54	70.23	72.44	64.67	63.32	62.16	63.38	94.65	93.49	94.42	94.19	77.77	73.00	73.00	74.59
ReMix-ABMIL (covary)	**72.32**	**76.71**	**71.02**	**73.35**	64.56	58.16	59.05	60.59	**94.75**	**93.55**	**94.49**	**94.26**	75.76	70.33	70.33	72.14
ReMix-ABMIL (joint)	72.13	76.00	70.91	73.01	**64.90**	**64.38**	**62.70**	**63.99**	94.69	93.45	94.42	94.18	**78.45**	**74.33**	**74.33**	**75.70**
Best Improvement Δ	+16.14	+18.21	+10.91	+15.09	+16.20	+5.82	+6.08	+9.37	+2.28	+0.76	+1.47	+1.50	+7.03	+11.33	+11.33	+9.89
ReMix-DSMIL (no aug.)	76.14	79.26	77.95	77.78	61.74	65.17	66.89	64.60	95.68	93.44	94.80	94.64	66.18	66.67	66.67	66.51
ReMix-DSMIL (append)	77.91	80.56	**81.02**	79.83	64.56	65.97	64.19	64.90	96.39	**94.10**	**95.43**	**95.31**	70.44	70.52	70.52	70.49
ReMix-DSMIL (replace)	76.60	79.30	78.64	78.18	63.62	66.76	59.32	63.24	95.33	93.44	94.65	94.47	67.10	67.00	67.00	67.03
ReMix-DSMIL (interpolate)	76.99	80.26	80.00	79.08	64.80	67.28	66.08	66.05	96.39	93.96	95.35	95.23	68.40	68.18	68.18	68.25
ReMix-DSMIL (covary)	77.72	80.52	80.46	79.57	64.88	**68.73**	**67.43**	**67.01**	**96.51**	93.88	95.35	95.25	71.20	70.33	70.33	70.62
ReMix-DSMIL (joint)	**78.20**	**80.94**	80.68	**79.94**	**66.21**	66.91	66.35	66.49	96.18	93.97	95.27	95.14	**72.44**	**70.82**	**70.82**	**71.36**
Best Improvement Δ	+5.28	+1.53	+4.66	+3.71	+6.31	+5.33	+4.06	+4.79	+2.14	+0.71	+1.32	+1.35	+14.11	+10.95	+10.95	+12.36

instance branch and a bag branch. The instance branch identifies the highest scored instance while the bag branch measures the similarity between other patches and the highest scored instance and then utilizes the similarity scores to compute attention.

We use DSMIL's codebase for MIL models' implementation and training. Unless otherwise specified, all MIL models are optimized for 50 epochs by the Adam optimizer [53] with an initial learning rate of 2e-4 and a cosine annealing learning rate schedule [65]. The mini-batch size is 1 (bag) for a fair comparison, despite that `ReMix` can easily scale it up since the reduced bags have the same number of instances and thus can be composed into a batch for parallel computing.

For comparison, we further adopt the official codes of TransMIL [83] and CLAM [66] and train them for 50 epochs for a fair comparison. Other settings, *e.g.*, learning rate and optimizer remain the same as their original releases.

3.3.5.3 Hyper-parameters

There are three hyper-parameters in `ReMix`, *i.e.*, the number of prototypes K, the augmentation probability p, and the strength λ. To study the effects of different hyper-parameters, we first sweep K in $\{1, 2, 4, 8, 16, 32\}$ to find the optimal K for each method and dataset. For simplicity and bag diversity, we set $p = 0.5$ in our main experiments for 4 individual augmentations, $p = 0.1$ for the "joint" augmentation, and uniformly sample λ from $(0, 1)$ in all experiments. For three datasets we study, both MIL methods share the optimal K values: $K = 1$ for UniToPatho800 dataset, $K = 4$ for UniToPatho7000 dataset, $K = 8$ for Camelyon16 dataset, and $K = 16$ for Colon10 dataset. We provide the empirical studies for each hyper-parameter and the design choices in Section 3.4.2, *e.g.*, studying the robustness of the choice of augmentation probability p, the choice of the number of prototypes K, and more.

3.4 Experiments

3.4.1 Main Results

3.4.1.1 Metrics comparison

Table 3.1 shows the main results for four datasets. Regardless of the difference in baselines (DSMIL and ABMIL), the results demonstrate `ReMix`'s superiority and robustness. Even without "mix" augmentations (no aug.), `ReMix` can improve DSMIL and ABMIL by only the "reduce" step in all datasets, *e.g.*, +13.75% and +3.22% precision for ABMIL and DSMIL, respectively, in the UniToPatho800 dataset, +1.50%/+1.31% precision for them in the Camelyon16 dataset, and +2.77%/+7.85% for them in the Colon10 dataset. Overall,

Table 3.2: **Comparison of training budgets.** Numbers are estimated from 50-epoch training on the same machine with an 8GB Tesla T4-8C virtual GPU. "Original / ReMix" rows show the multiplier between the original's and ReMix version's budgets.

	Average Seconds / Epoch		Memory Peak		FLOPs	
Methods \Datasets	UniToPatho800	Camelyon16	UniToPatho800	Camelyon16	UniToPatho800	Camelyon16
ABMIL	18.41″	235.72″	55.63 MB	332.12 MB	840.51M	4.20G
ReMix-ABMIL	0.84″	1.10″	6.45 MB	8.76 MB	531.46K	4.20M
Original / ReMix	21.93×	214.29×	8.61×	37.91×	1581.51×	999.76×
DSMIL	19.20″	255.14″	66.58 MB	364.72 MB	1.06G	5.25G
ReMix-DSMIL	0.85″	1.12″	6.46 MB	8.76 MB	1.49M	5.38M
Original / ReMix	22.57×	227.80×	10.31×	41.63×	713.38×	975.49×

† All the data are stored in a distributed storage platform, which might exacerbate the I/O problem for large bags.

ABMIL benefits more from `ReMix` than DSMIL. DSMIL computes *self-attention* to explicitly consider the similarity between different instances inside a bag, while ABMIL directly predicts attention scores using an MLP for all instances without such explicit inter-instance relation computing. For this reason, we conjure that ABMIL's attentions are more likely to overfit than DSMIL's, and thus, the denoised reduced-bags can benefit it more. Representative prototypes can ease the recognition process and alleviate the overfitting problem. These results suggest that `ReMix` can reduce data noise in the bag representations to some extent, improving performance.

Applying "Mix" augmentation further improves the performance of reduced-bags (no aug.) by a considerable margin, *e.g.*, +2.27% and +3.07% accuracy for ReMix-ABMIL and ReMix-DSMIL, respectively, in the UniToPatho800 dataset, and +3.58% and +15.30% precision for them in the Colon10 dataset. The proposed four latent augmentations perform similarly well across different datasets and MIL methods, indicating their robustness. Especially, "covary" augmentation achieves top-tier performance in most datasets, confirming our motivation that transferring others' covariance in the latent space could provide reliable and diversified variations for semantic augmentation. Using full-bags can be seen as a particular case of augmenting the prototypes with their own covariances. However, such bags are static and unaltered, as discussed in Section 3.3.3. In contrast, with `ReMix`, the reduced and augmented bags can be more diverse and dynamic. Such augmentations are helpful for low-data applications like WSI classification. "Joint" augmentation integrates advantages from different latent augmentations and is the most robust augmentation. For example, it achieves top 2 performance in six of eight settings (2 MIL methods $\times$ 4 datasets).

Among the studied datasets and tasks, classification in Camelyon16 is the easiest since it is a binary classification problem with many samples for each class. In contrast, UniToPatho and Colon10 datasets are 6-class and 10-class classification problems, respectively, and they have fewer samples for each class than the Camelyon16 dataset. From the "best improvement" rows in Table 3.1, it is clear that *ReMix* improves more for datasets that have fewer training samples and more classes. This indicates `ReMix`'s good property for "small" data and "hard" problems.

Overall, solid gains observed in Table 3.1 have confirmed the effectiveness of the proposed `ReMix` framework. We next demonstrate its efficiency.

3.4.1.2 Training budgets

We compare the training budgets, *i.e.*, the average training time per epoch, the peak memory consumed during training, and the estimated floating-point operations per second (FLOPs) during one iteration in Table 3.2. Our `ReMix` framework outperforms other entries in all training budgets. It costs nearly 20$\times$ less training time but obtains better results for both MIL methods in the UniToPatho800 (*e.g.*, +10.91% accuracy). `ReMix` framework uses fewer FLOPs to finish one iteration, *e.g.*, 5.25G FLOPs v.s. 5.38M FLOPs. Moreover, it takes `ReMix` a much shorter training time to achieve better results than the original ones in the Camelyon16 dataset, whose average bag size is about 5$\times$ as big as UniToPatho800's. It can be expected that the training efficiency gains would enlarge as the bag size and the number of WSIs in the dataset increase. With more data collected in the real world, we argue that the training efficiency should be as important as the classification performance when scaling up to large datasets. Therefore, we emphasize the superiority of `ReMix` in being an efficient

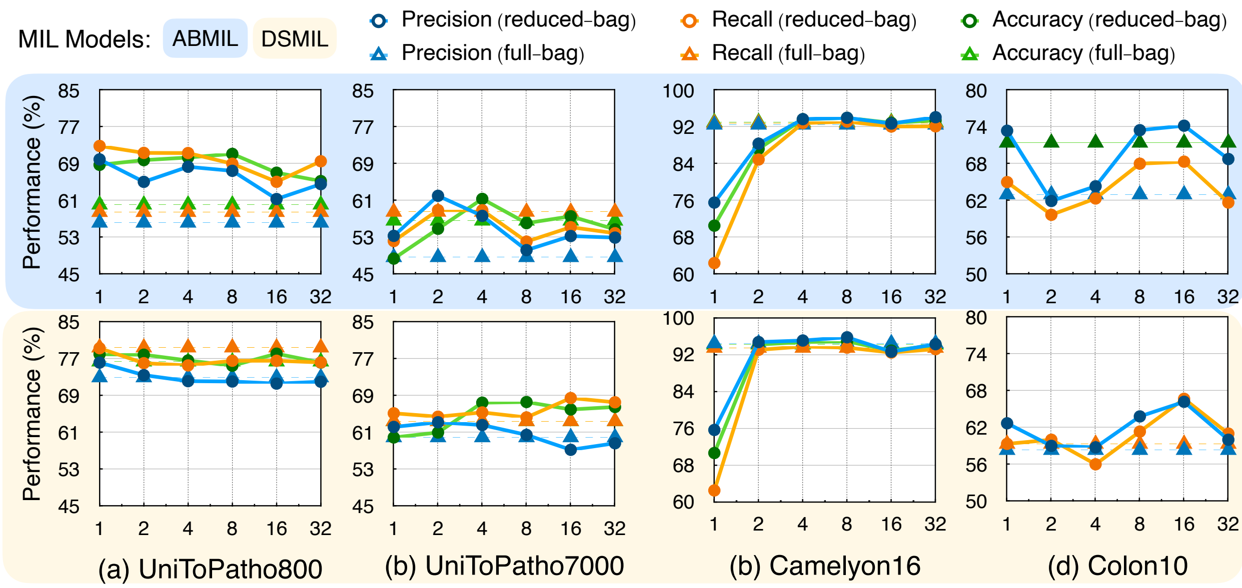


Figure 3.2: **Empirical study on the number of prototypes.** Horizontal axes denote the number of prototypes in the reduced-bags. Baselines are trained on the full-bags. The results are an average of 10 runs. Blue and orange blocks denote the mil models, ABMIL and DSMIL, respectively.

framework.

Table 3.3: **Empirical study on augmentation probabilities.** The displayed metrics are the average of precision, recall and accuracy. Best performance of each row is in bold. All results are averaged over 10 runs. Numbers are shown in percentage (%).

	UniToPatho7000 dataset											
Method	ABMIL						DSMIL					
Aug.\Prob.	0.1	0.3	0.5	0.7	0.9	$\mathbb{E}$(aug|p)	0.1	0.3	0.5	0.7	0.9	$\mathbb{E}$(aug|p)
baseline (full-bag)	Not Applicable					54.62	Not Applicable					62.23
ReMix (append)	57.16	60.75	**61.61**	59.70	58.40	59.52	**65.95**	65.80	64.90	63.20	62.10	64.39
ReMix (replace)	**60.38**	60.02	60.17	56.54	56.67	58.76	**65.75**	64.86	63.24	60.45	62.09	63.28
ReMix (interpolate)	59.45	64.02	63.38	**64.38**	62.16	**62.68**	**67.48**	66.24	66.05	66.39	67.29	**66.69**
ReMix (covary)	58.76	58.61	60.59	60.72	**61.11**	<u>59.96</u>	66.21	66.86	67.01	**67.09**	65.80	<u>66.59</u>
$\mathbb{E}$(p|aug)	58.94	60.85	**61.44**	60.33	59.58	60.23	**66.35**	65.94	65.30	64.28	64.32	65.24

3.4.2 Ablation study

In this section, we investigate the robustness of `ReMix` to different components and discuss its design choices.

3.4.2.1 Ablation on the number of prototypes K

Figure 3.2 shows the performance of using different numbers of prototypes for bag representations. In the UniToPatho800 dataset (Fig. 3.2-(a)), both ABMIL and DSMIL achieve the best results with only one prototype. Besides, ABMIL is quite robust to the choice of K in that it consistently outperforms the full-bag representations (baseline) with the reduced-bag representations. This can be expected since the UniToPatho800 dataset contains small patches that are mainly concentrated on tissues of interest. When it comes to the UniToPatho7000 dataset, more prototypes are needed for optimal performance ($K = 4$) as the bag size enlarges and the mixture of different types of tissues inside the bag is inevitable. In the Camelyon16 dataset (Fig. 3.2-(b)), `ReMiX` performs similarly well when $K \geq 4$, with $K = 8$ being the best. Camelyon16 has a severe issue of tissue imbalance that the lesion area of tumorous tissues accounts for only 10% to 30% of all tissue areas. More prototypes are needed for bag representation to preserve the minority information. Nevertheless, training on the reduced-bags ($1 \sim 100$ instances/bag) is still significantly cheaper than full-bags ($10^3 \sim 10^5$ instances/bag). Similarly, `ReMix` performs the best with 16 prototypes for both models in the Colon10 dataset (Fig. 3.2-(c)). Both MIL methods share similar curves across three datasets, showing the generality of the "reduce" step. This study confirms our motivation that several representative prototypes could provide sufficient information of the full-bag for specific downstream WSI classification tasks.

3.4.2.2 Ablation on the augmentation probability p

Table 3.3 compares `ReMix` when applying different augmentations under different probabilities p in the UniToPatho7000 dataset. We estimate the expected performance given an

Table 3.4: **Empirical study on training epochs.** The reported numbers shown in percentage (%) are the average of precision, recall and accuracy. All results are the mean of 10 trials with their standard deviations denoted by±.

	UniToPatho7000 dataset			
	ABMIL		DSMIL	
Epoch	Full-bag	Reduced-bag	Full-bag	Reduced-bag
50	54.62±3.91	59.30±3.17	62.23±2.06	64.60±1.64
100	55.60±7.61	64.14±2.26	59.86±2.12	**65.57**±2.02
200	55.41±5.73	**64.58**±1.84	60.17±1.32	65.21±3.50

augmentation method with varying probability ($\mathbb{E}(\text{aug}|p)$) and the expected performance given a fixed probability with varying augmentation methods ($\mathbb{E}(p|\text{aug})$). In our main experiments, we naively choose $p = 0.5$ to demonstrate the effectiveness of `ReMix`. Beyond the naive selection of p, better performance can be achieved with a properly tuned probability. In practice, one can use a validation set for parameters tuning. The expected performance of different augmentations ($\mathbb{E}(\text{aug}|p)$) shows that our `ReMix` can improve baselines in expectation (*e.g.*, +8.06% and +4.46% for "interpolate" in ABMIL and DSMIL, respectively). These results indicate the robustness of `ReMix` to the choice of augmentation probability.

3.4.2.3 Ablation on training epochs

Training the MIL classifiers for 50 epochs on full-bags might put them at a disadvantage compared to training on reduced-bags since full-bags have much more instances and therefore need longer training. To test it, we compare the candidates with longer training in Table 3.4. When trained on full-bags, only ABMIL gains from longer training, and the performance of DSMIL even drops considerably. When trained on reduced-bags, both MIL methods start to benefit from longer training, showing the better potential of reduced-bags. Overall, all the tested cases support the superiority of `ReMix` regardless of the number of training epochs.

3.4.2.4 Append or replace with generated features?

In Section 3.3.2.4, we append the newly generated features to the query bag. An alternative is to replace the original feature with the generated features. Table 3.5 reports the comparison results. Appending newly generated features is slightly better than replacing the original feature with the generated features. This is anticipated, as replacing the original feature with generated features introduces perturbation twice. One perturbation is the newly generated prototype, which is not as accurate as other fundamental semantic prototypes (cluster center). The other is the replacement operation. Therefore, the noise may accumulate. In contrast, appending the newly generated features preserves the original fundamental semantic prototypes. Nevertheless, replacing with generated features also works well.

Table 3.5: **Empirical study on appending or replacing with generated features.** The displayed metrics are the average of precision, recall, and accuracy. All results are the average over 10 trials with their standard deviations denoted by±. Numbers are shown in percentage (%).

	UniToPatho7000 dataset			
Method	ABMIL		DSMIL	
Augs.\mode	Append	Replace	Append	Replace
ReMix (interpolate)	**63.38**±3.21	61.10±5.23	**66.05**±1.55	65.41±1.11
ReMix (covary)	**60.59**±3.72	59.90±3.15	**67.07**±1.87	65.51±1.37

3.4.3 ReMix is robust to pre-trained encoders

Our results thus far are based on self-supervised pre-trained encoders, which are known to provide good representations. We next demonstrate that both the "reduce" and "mix" steps can generalize to other pre-trained encoders.

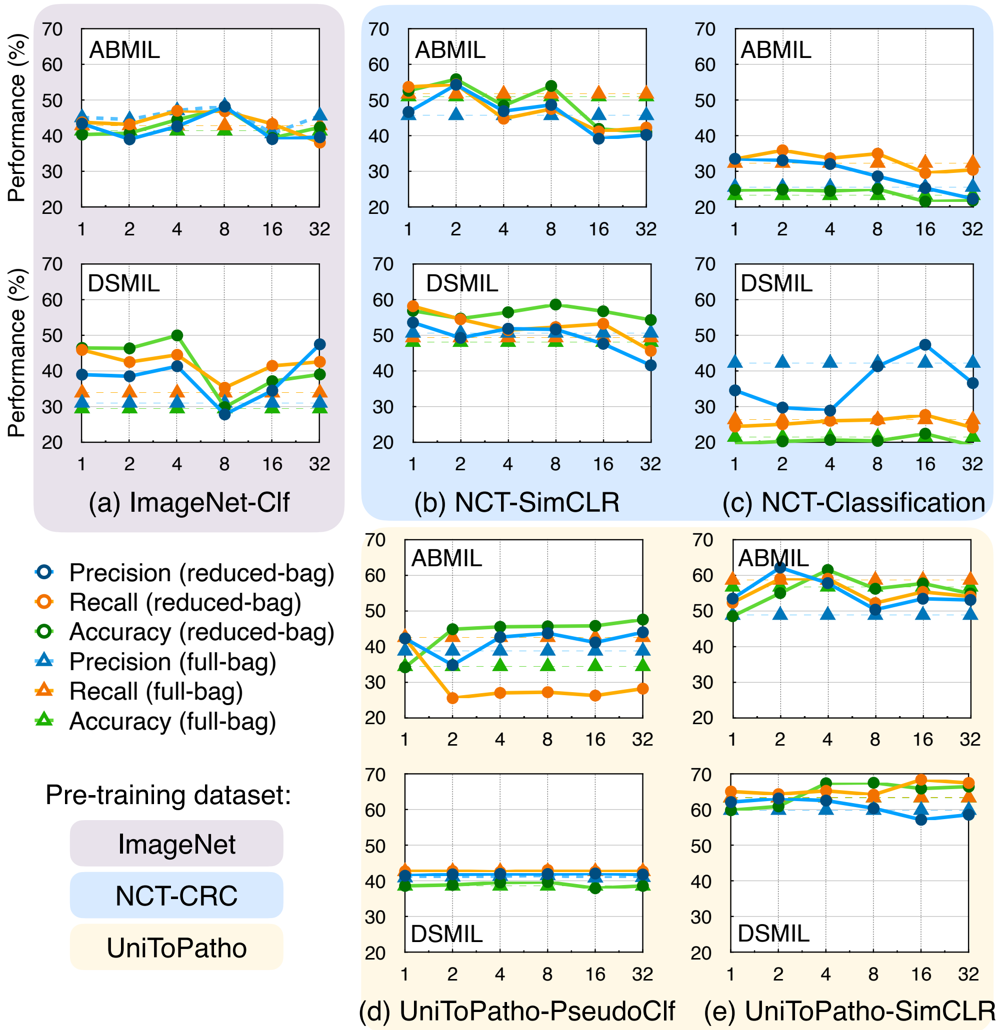


Figure 3.3: **Empirical study on the impact of encoder to ReMix and prototype selection in the UniToPatho7000 dataset.** Horizontal axes denote the number of prototypes in the reduced-bags. Results are the average of 10 runs.

3.4.3.1 "Reduce" is robust to pre-trained encoders

Figure 3.3 shows how "reduce" performs with different pre-trained encoders in the UniToPatho7000 dataset. Our previous observation that reduced-bags are competitive or even outperform full-bags can also generalize to other pre-trained encoders. Notably, the SimCLR encoder pre-trained on the UniToPatho dataset performs the best in general (Fig. 3.3-(e)), followed by the SimCLR encoder pre-trained on the NCT dataset (Fig. 3.3-(b)). This emphasizes the importance of pre-training on the target datasets themselves and indicates the superiority of self-supervised pre-training. There are other two interesting observations. First, although NCT is a colorectal tissue dataset which should be similar to UniToPatho, a classification-oriented encoder pre-trained on it does not transfer well to the UniToPatho dataset (Fig. 3.3-(c)) and even falls behind the ImageNet-supervised pre-trained encoder (Fig. 3.3-(a)); in fact, it is the worst encoder among others. Second, despite the popularity and success of SimpleMIL [23], pre-training in previous works [47, 9, 41, 13], it does not necessarily perform better than ImageNet-supervised pre-trained encoder. These two observations might challenge some common beliefs and encourage people to rethink the pre-training methods to use.

3.4.3.2 "Mix" is robust to pre-trained encoders

In addition to the results presented in Table 3.1, which are based on a self-supervised encoder pre-trained on the UniToPatho dataset, Table 3.6 further shows how `ReMix` improves other pre-trained encoders in the UniToPatho7000 dataset. For simplicity, we only study ImageNet classification pre-trained and NCT SimCLR pre-trained encoders. The boosted performance indicates the generality of the "Mix" augmentation, which means `ReMix` does not pose a strict requirement for the patch encoder and can be applied to existing encoders without re-training.

3.4.4 ReMix improves abnormality recognition

We visualize the attention scores predicted by ABMIL in the Colon10 dataset in Figure 3.4. The attention of the original ABMIL method only focuses on parts of abnormal tissues, while with `ReMix`, more complete coverage of abnormal tissue is observed. Though seeing only several instances per bag during training, our `ReMix` behaves decently and improves the original model in highlighting relevant patches. This implies the advantage of using representative reduced-bags over uncurated full-bags. This experiment also demonstrates the potential of `ReMix` in generating pseudo-instance-level labels, which might help semi-

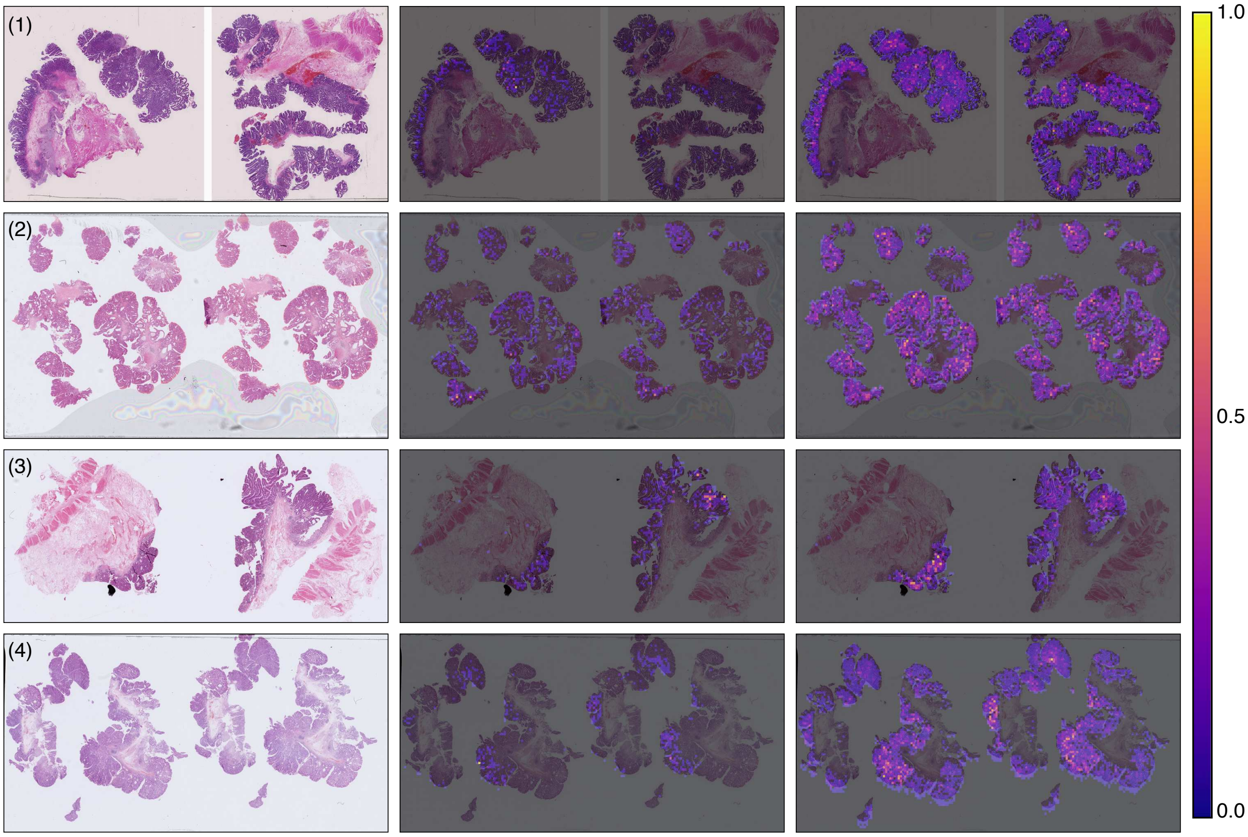


Figure 3.4: **Visualization of attention maps.** (a) Original WSIs. (b) Attention maps of ABMIL trained on full-bags. (c) Attention maps of ABMIL trained with ReMix-joint. The classes of (1) and (4) are high-Grade dysplasia and mucinous adenocarcinomas, respectively. The class of (2) and (3) is carcinoma in situ.

supervised learning, semantic segmentation, and other problems.

3.4.5 Spatial-aware MIL methods also benefit from `ReMix`

Our `ReMix` is presumed to be applied to spatial-agnostic MIL models, but it can have a straightforward extension to spatial-aware MIL methods, as introduced in Section 3.3.2.4. We demonstrate this on TransMIL [83], a recent state-of-the-art spatial-aware MIL method. Table 3.7 reports the results using the same probability of 0.5 as previous experiments. Both augmentations can improve TransMIL in all three metrics. This study supports the use of

Table 3.6: Comparison of ReMix applied on different pre-trained patch encoders. We report the "Average" metric here, *i.e.*, the average of precision, recall, and accuracy. Results are averaged over 10 trails with their standard deviations denoted by±. Numbers in parentheses denote the improvements from corresponding full-bag representations. We use $K = 4$ for "reduce" for both encoders here.

	UniToPatho Dataset	
Pre-trained encoder	Methods	Average (%)
ImageNet-Clf	DSMIL	31.34±6.05
	+ReMix (no aug.)	45.16±6.10 (+13.82)
	+ReMix (joint)	50.82±3.73 (+19.48)
NCT-SimCLR	DSMIL	49.17±5.48
	+ReMix (no aug.)	53.15±2.73 (+3.88)
	+ReMix (joint)	55.44±2.79 (+6.27)

`ReMix` also for spatial-aware MIL methods. In addition to the current naive extension, we believe more improvement would emerge if the "reduce" step could be more properly integrated with spatial-aware MIL methods, which we leave for future work.

3.5 Limitations and Future Works

Despite `ReMix`'s empirical success demonstrated in this work, some limitations still exist. First, `ReMix` relies on K-Means clustering to obtain fundamental semantic prototypes. However, the K-Means clustering algorithm has underlying assumptions about the data for its success, *e.g., i.i.d.* samples and isotropic feature distribution, which are not always satisfied for WSI tasks. In addition, tiny regions of interest might be overlooked during the clustering step, which could contribute to the failure of `ReMix`. Second, there is an underlying requirement for the number of instances to estimate the covariance matrix for a cluster.

Table 3.7: "Mix" augmentation improves spatial-aware MIL method. The "Average" column reports the average of precision, recall, and accuracy. Results are averaged over 10 independent runs with their standard deviations shown after±.

	Camelyon16 Dataset			
Methods\Metrics	Precision (%)	Recall (%)	Accuracy (%)	Average (%)
TransMIL [83]	88.27±1.40	85.98±1.60	87.91±1.17	87.39±1.25
+ReMix (interpolate)	90.20±2.00	**88.61**±1.43	**89.95**±1.48	**89.55**±1.37
+ReMix (covary)	**90.92**±2.08	87.49±2.30	89.66±1.83	89.07±1.86
	Colon10 Dataset			
TransMIL [83]	63.28±4.74	62.67±3.79	62.67±3.79	62.87±4.06
+ReMix (interpolate)	**65.28**±3.33	64.67±2,81	64.67±2,81	**64.87**±2.87
+ReMix (covary)	63.78±3.78	**65.34**±2.33	**65.34**±2.33	64.82±2.74

An insufficient number of patches might yield inaccurate cluster prototypes and ill-defined covariance matrices, possibly degenerating final performance. Using dynamic numbers of prototypes for different WSIs could be a way to address it.

The success of `ReMix` has been supported for WSI classification tasks for image modality in this work but could go beyond. We also expect its application to survival prediction and other WSI analysis tasks where data diversity is the major issue. `ReMix` also has potentials in multi-modality learning problems, *e.g.*, images with tabular data. Interpolating features or transferring semantics via covariance matrices are also feasible for tabular data representations. More intriguing and interesting methods might be mined from the joint use of `ReMix` for different modalities of data.

3.6 Conclusion

This paper presents `ReMix`, a general and efficient framework for MIL-based WSI classification. For spatial-agnostic MIL models, `ReMix` reduces the number of instances in WSI bags by substituting instances with instance prototypes. Subsequently, `ReMix` enhances data diversity by mixing the bags using various latent space augmentation techniques. Furthermore, for spatial-aware MIL models, `ReMix` can provide performance improvement by simply employing the "Mix" augmentation.

Overall, `ReMix` enhances the performance of previous state-of-the-art MIL classification methods, often with less computational resources, demonstrating its effectiveness and efficiency. To the best of our knowledge, the combined use of reduce" and mix" has not been previously studied in slide-level WSI analysis. We anticipate that the "Mix" augmentation method proposed in this work will inspire further research in this domain, where data augmentation is crucial yet underexplored.

CHAPTER 4

Bootstrapping yourself: Concept Contrastive Learning for Better Dense Representations

4.1 Introduction

Computational pathology is rapidly advancing due to deep learning (DL) applications on whole slide images (WSIs) [92]. The use of pre-trained model weights is a common practice to mitigate the annotation load, with self-supervised learning (SSL) methods, free of annotations, gaining recent interest [42, 14, 38]. SSL methods, initiated by contrastive learning [39, 105, 14, 42, 18], have largely focused on image-level representations, leaving a gap for dense prediction tasks such as object detection and instance segmentation, leading to detection-friendly pre-training methods [99, 113, 64, 76, 46, 95, 108, 109]. However, similar studies in the pathology image domain remain scarce. This research aims to address this by applying SSL to dense prediction tasks in pathology images.

We introduce the **Con**cept **C**ontrastive **L**earning (ConCL) framework, contrasting local semantic regions instead of image-level representations [105, 14, 42]. ConCL is an abstraction of dense contrasting frameworks encompassing related works. We first benchmark current leading SSL methods and DenseCL [99], revealing a performance gap that indicates the advantage of dense (grid-level) contrasting over image-level contrasting. Inspired by these differences and pathology images' characteristics, we enhance ConCL through several explorations, focusing on dense prediction pre-training success factors and optimal concepts for pathology images. The results suggest that a randomly initialized model can group meaningful concepts and aid dense pre-training. The final ConCL framework outperforms various state-of-the-art SSL methods across different conditions.

The contributions of this work are as follows:

- We present one of the first systematic studies of SSL methods for dense prediction tasks in pathology images, narrowing the gap between studies in natural and pathology images.
- We introduce ConCL, an SSL framework for dense pre-training, and show that a randomly initialized model can learn semantic concepts, improving itself while achieving competitive results.
- We demonstrate the importance of *dense* pre-training in pathology images and provide observations that could contribute to other applications in pathology image analysis or beyond.

We hope this work could provide useful data points and encourage the community to conduct ConCL pre-training for problems of interest.

A large portion of this chapter has been published in [114].

4.2 Related work

Contrastive learning. Deep learning's success owes much to the use of vast amounts of data. When limited data is available, transferring knowledge from pre-trained models is an alternative [36, 43]. SSL methods, which learn from label-free pretext tasks such as colorization [124, 125] and denoising [96], have gained attention. Instance discrimination [39, 105, 42, 18, 14], a pretext task in contrastive learning [42, 18, 14, 70, 105, 11], optimizes similarity between positive pairs while minimizing it between negative pairs. Later methods, like SwAV[11] and PCL [60], combined contrasting with clustering.

Dense prediction pre-training. Good image-level representations do not guarantee better performance in dense prediction tasks. Hence, recent studies focused on dense prediction pre-training [99, 113, 76, 95, 108, 109, 64, 46]. For example, DenseCL [99] applies contrastive loss at pixel-level, while *Self*-EMD [64] performs non-contrastive dense predicting. However, the efficiency of external mask generators used in these works is untested in pathology images, motivating our proposed concept mask generator.

SSL in pathology images. SSL methods in pathology images remain under-studied. Some domain-specific self-supervised pretext tasks have been proposed [54], and SimCLR [14] has been studied for various tasks in pathology images [25]. Nevertheless, studies on detection/segmentation-friendly SSL methods in pathology images are scarce. Our work addresses this gap, proposing a roadmap toward better dense prediction performance in pathology images.

4.3 Method

4.3.1 Preliminary: Instance Contrastive Learning

MoCo[42] abstracts the instance discrimination task as a dictionary look-up problem. Specifically, for each encoded query q, there is a set of encoded keys $\{k_0, k_1, k_2, ...\}$ in a dictionary. The instance discrimination task is to pull closer q and its matched positive key k_+ in the dictionary while spreading q away from all other negative keys k_-. When using the dot-product as similarity measurement, a form of contrastive loss function based on InfoNCE[70] becomes:

$$L_q = -\log \frac{\exp(q \cdot k_+/\tau)}{\exp(q \cdot k_+/\tau) + \sum_{k_-} \exp(q \cdot k_-/\tau)} \tag{4.1}$$

where τ is a temperature hyper-parameter [105]. Queries q and keys k are computed by a query encoder and a key encoder, respectively [42, 18]. Formally, $q = h(\texttt{GAP}(f_5(x_q)))$, where h is a MLP projection head as per [14]; $\texttt{GAP}(\cdot)$ denotes global-average-pooling, and $f_5(x)$ represents the outputs from the stage-5 of a ResNet [44]. Keys k are computed similarly using the key encoder. In MoCo [42], the negative keys are stored in a queue to avoid using large batches [14].

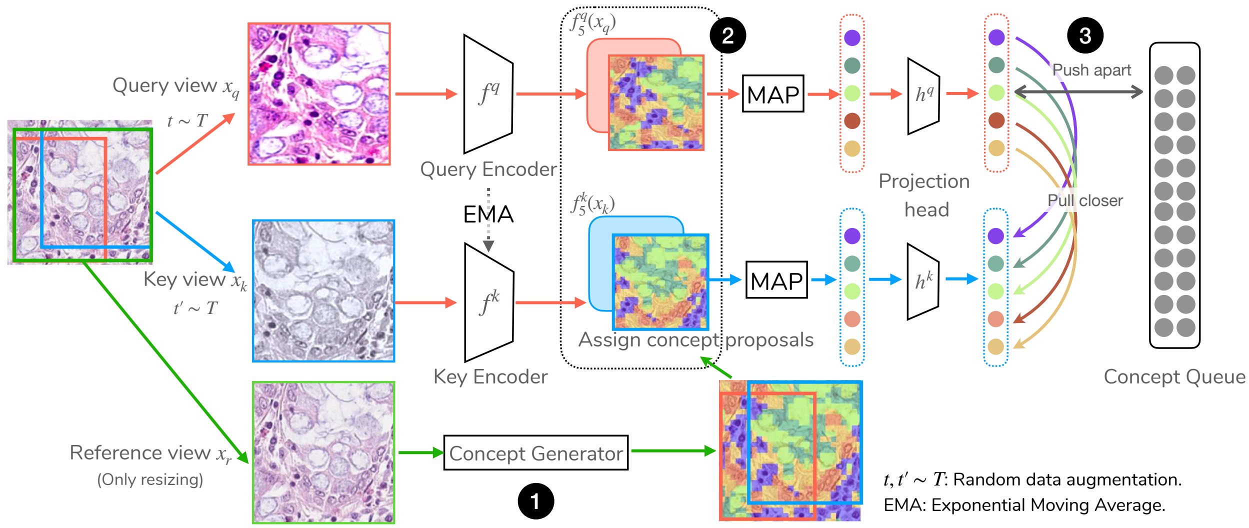


Figure 4.1: **ConCL overview.** ConCL has three steps: (1) Given a query view x_q and a key view x_k, their union region is cropped as a reference view x_r. ConCL obtains concept proposals by processing x_r with a "concept generator." (2) For the shared concepts, ConCL computes their representations via masked average pooling (MAP). (3) ConCL optimizes concept contrastive loss (Eq. (4.2)), and enqueues the concept prototypes from the key encoder to the concept queue.

4.3.2 Concept Contrastive Learning

Instance contrastive methods [14, 42, 105] do well in discriminating among image-level instances, but dense prediction tasks usually require discriminating among local details, *e.g.*, object instances or object parts. We abstract such local details, or say, fine-grained semantics as "concepts." A concept does not necessarily represent an object. Instead, any sub-region in an image could be a concept since it contains certain different semantics. From the perspective of dense prediction, it is desirable to build concept-sensitive representations. For example, one WSI patch usually contains multiple small objects, *e.g.*, nucleus, glands, and multiple texture-like tissues, *e.g.*, mucus [92, 51]. To successfully detect and segment objects in such images, models need to learn more information from local details. To this end, we propose a simple but effective framework — ***Con****cept* ***C****ontrastive* ***L****earning* (ConCL). Figure 4.1 shows its overview, which we elaborate on below.

Concept discrimination. We first define a pretext task named concept discrimination. Similar to instance discrimination [105, 39], concept discrimination requires a model to discriminate among the representations of the same but augmented concepts and the representations of different concepts. We formulate concept discrimination by extending the instance-level queries and keys to concept-level. Specifically, given an encoded query concept q^c and a set of encoded key concepts $\{k_0^c, k_1^c, k_2^c, ...\}$, we derive concept contrastive loss as:

$$L_c = -\log \frac{\exp(q^c \cdot k_+^c/\tau)}{\exp(q^c \cdot k_+^c/\tau) + \sum_{k_-^c} \exp(q^c \cdot k_-^c/\tau)} \tag{4.2}$$

where τ is the same temperature parameter and k_-^c are keys in the concept queue — the queue to store concept representations. This objective brings representations of different views of the same concept closer and spreads representations of views from different concepts apart.

Concept mask proposal. We use masks to annotate fine-grained concepts explicitly. Assume a mask generator is given, as diagramed at the bottom of Figure 4.1; we first pass a reference view x_r, defined as the circumscribed rectangle crop of the union of two views, into the mask generator to obtain a set of concept masks — $\mathcal{M}_r = \{m_i\}_{i=1}^K$, where K is the number of concepts. Since the reference view contains both the query view and the key view, their concept masks $\mathcal{M}_q$ and $\mathcal{M}_k$ are immediately obtained if we restore them in the reference view. Then, we derive concept representations in both views by masked average pooling (MAP) with resized concept masks. Specifically, we compute $q^c = h\left(\mathtt{MAP}\left(f_5(x_q), m_c\right)\right)$ and k^c similarly, where $\mathtt{MAP}\left(z, m\right) = \sum_{ij} m_{ij} \cdot z_{ij} / \sum_{ij} m_{ij}$, and $z \in \mathbb{R}^{CHW}$ denotes feature maps, $m \in \{0, 1\}^{HW}$ is a binary indicator for each concept. Here, only the shared concepts in both views are considered, *i.e.*, $m_c \in \mathcal{M}_q \cap \mathcal{M}_k$.

Our analysis hereafter focuses on 1) What makes the success of dense prediction pretraining? 2) What kind of concepts are good *for pathology images*? Different answers to these two questions reveal the characteristics of pathology images and the disparity between natural and pathology images, as we explore in Section 4.4. Below, we first introduce the benchmark pipeline and setups.

4.3.3 Benchmark Pipeline

Despite the extensive benchmarks in natural images for dense tasks, to our knowledge, such studies are unfortunately *absent* in current works for pathology. Note that studying SSL methods in pathology images is still at an early stage. Most current works focus on employing image-level SSL methods for classification tasks. Orthogonal to theirs, we investigate a wider range of SSL methods for object detection and instance segmentation tasks, which are of high clinical value. We hope our work could provide useful data points for future work.

We briefly introduce the datasets here:

- *Pre-training dataset.* We use NCT-CRC-HE-100K[51] dataset, referred to as NCT, for pre-training. It contains 100,000 non-overlapping patches extracted from hematoxylin and eosin (H&E) stained colorectal cancer and normal tissues. All images are of size 224×224 at 0.5 MPP (20$\times$ magnification). We randomly choose 80% of NCT to be the pre-training dataset.

- *Transferring dataset.* We use two public datasets, the gland segmentation in pathology images challenge (GlaS) dataset [88] and the colorectal adenocarcinoma gland (CRAG) dataset [37], and follow their official train/test splits for evaluation. GlaS [88] collects images of 775$\times$522 from H&E stained slides with object-instance-level annotation; the images include both malignant and benign glands. CRAG [37] collects 213 H&E stained images taken from 38 WSIs with a pixel resolution of 0.55μm/pixel at 20$\times$ magnification. Images are mostly of size 1512$\times$1516 with object-instance-level annotation. We study the performance of object detection and instance segmentation.

Experimental setup. We pre-train all the methods on the NCT training set for 200 epochs. For ConCL pre-training, we warm up the model by optimizing instance contrastive loss (Eq. (4.1)) for the first 20 epochs and switch to concept contrastive loss (Eq. (4.2)). Then, we use the pre-trained backbones to initialize the detectors, fine-tune them on the training sets of transferring datasets, and test them in the corresponding test sets. Unless otherwise specified, we run all the transferring experiments 5 times and report the averaged performance.

4.4 Towards Better Concepts: a Roadmap

In this section, we first benchmark some popular state-of-the-art SSL methods for dense pathology tasks. Then, we start with DenseCL [99] and derive better concepts along the way, directed by the questions raised in the previous section.

4.4.1 Benchmarking SSL methods for Dense Pathology Tasks

Benchmark results. Table 4.1 (baselines and prior SSL arts) shows the transferring performance for GlaS dataset (left columns) and CRAG dataset (right columns), respectively. We report results using 200-epoch pre-trained models and a 1× fine-tuning schedule. On the GlaS dataset [88], we observe that the gap between training from randomly initialized models and training from supervised pre-trained models is relatively smaller compared to those in the natural image domain [19, 18, 38, 14]. Nonetheless, state-of-the-art SSL methods all exceed supervised pre-training, meeting the same expectation as in natural images. Yet, on the CRAG dataset [37], most of the pre-trained models, including both the self-supervised ones and the supervised one, fail to achieve competitive performance compared to training from randomly initialized weights. The only exception is DenseCL [99], a dense contrasting method.

Among the image-level SSL methods, MoCo-v2 [18] performs the best in GlaS and the second-best in CRAG. Enhanced by dense contrasting, DenseCL [99] achieves the best results in both datasets. It should be emphasized that DenseCL [99] gets + 1.6 AP^{bb} for GlaS by using grid-level contrasting. This demonstrates the importance of designing dense pre-training frameworks when transferring to dense tasks since all the stragglers are only optimized for image-level representations. Thus, we here conclude *dense contrasting matters.*

Category	Methods	GlaS				CRAG			
		Detect		Segment		Detect		Segment	
		AP^{bb}	AP^{bb}_{75}	AP^{mk}	AP^{mk}_{75}	AP^{bb}	AP^{bb}_{75}	AP^{mk}	AP^{mk}_{75}
Baselines	Rand. Init.	49.8	57.3	52.1	60.7	51.1	57.0	50.6	57.3
	Supervised	50.2	56.9	53.2	62.1	49.2	55.2	49.4	55.0
Sec. 4.4.1 Prior SSL arts	SimCLR[14]	50.7	56.9	53.6	62.7	49.2	54.8	49.1	54.7
	BYOL[38]	50.9	57.7	53.9	62.6	49.9	55.8	49.3	55.3
	PCL-v2† [60]	49.4	55.9	51.9	61.0	51.0	56.6	50.5	56.7
	MoCo-v1[42]	50.0	56.2	52.1	59.9	47.2	51.1	47.5	52.0
	MoCo-v2[18]	52.3	60.0	55.3	65.0	50.0	55.7	50.3	56.8
	DenseCL[99]	53.9	62.0	56.5	66.2	52.3	58.2	52.2	59.8
Our differently instantiated ConCLs:									
Sec. 4.4.2 Grid concepts	(1) g-ConCL(s=3)	54.9	64.1	57.1	66.3	55.4	62.3	54.4	62.0
	(2) g-ConCL(s=5)	55.4	65.2	57.4	67.2	55.5	62.7	54.6	62.2
	(3) g-ConCL(s=7)	54.9	63.8	57.0	66.5	55.3	62.5	54.7	62.6
Sec. 4.4.3 Natural-image priors concepts	(4) fh-ConCL(s=50)	55.8	65.6	58.3	68.8	54.8	60.7	54.1	60.7
	(5) fh-ConCL(s=500)	56.2	65.9	57.7	67.9	54.7	61.9	53.8	60.5
	(6) bas-ConCL	56.1	66.1	58.1	68.1	54.2	61.1	53.4	60.8
Sec. 4.4.4 Bootstrapped concepts	(7) b-ConCL(f_4)	**56.8**	**66.2**	**58.7**	**68.9**	55.1	62.2	54.1	61.4
	(8) b-ConCL(f_5)	56.1	65.6	57.8	67.7	**56.5**	**63.3**	**55.3**	**62.9**

Table 4.1: **Main results of object detection and instance segmentation.** AP^{bb}: bounding box mAP, AP^{mk}: mask mAP.

4.4.2 Correspondence matters

From the previous section, we find dense contrasting is favored in both natural and pathology images, where DenseCL [99] all achieves top performance. The next question is: *can we improve the dense contrasting framework?* To answer it, we first summarize the overall pipeline of DenseCL [99]. DenseCL computes the dense representations of two views without global average pooling, *i.e.*, $f_5(x_q), f_5(x_k)$, and passes them to a dense projection head to

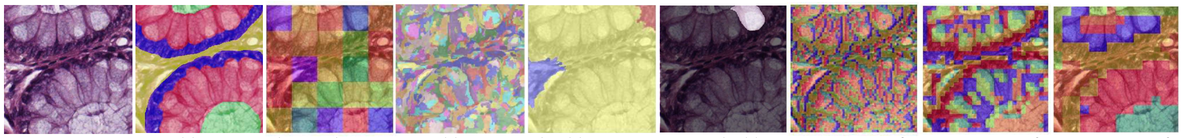


Figure 4.2: **Concept descriptors.** (a) Tissue concept illustration. (b) Grid concepts (s: grid number). (c-d) FH concepts (s: scale). (e) Binary saliency concepts, obtained from BASNet [73]. (f-h) Clustering concepts (f_i: ResNet output stage). The image is resized to 448×448 for better visualization.

obtain final grid features of size $\mathbb{R}^{128\times7\times7}$. Then it sets the most similar (measured by cosine similarity) grids in two views as positive pairs. As such, the correspondence of positive pairs is learned. However, the reliability of learned correspondence remains questionable and would affect the quality of learned representations.

To address that, we instantiate DenseCL [99] in ConCL by regarding the grid-prior as a form of concept, as shown in Figure 4.2-(b). We denote this ConCL instance as g-ConCL. Compared with DenseCL [99] (learned matching), ConCL naturally restores the positive correspondence from a reference view (precise matching Fig. 4.1-x_r). Table 4.1-(1-3) compares the original DenseCL [99] and ConCL-instantiated g-ConCL. The results indicate that g-ConCL with precise correspondence can boost DenseCL [99] by a large margin. Even with the simplest form of concepts, g-ConCL already has topped entries above it in Table 4.1. We believe other dense pre-training methods that learn the matching between grids, *e.g.*, *Self*-EMD [64], should perform similarly to DenseCL [99], and g-ConCL could outperform them. Thus, we conclude that *correspondence matters*.

4.4.3 Natural Image Priors in Pathology Images

ConCL is a general framework for using masks as supervision to discriminate concepts. Some previous works in natural image [128, 46, 127, 95, 98] also combines masks with contrastive learning, where the masks are provided by ground truth annotation [128, 98, 46], or supervised/unsupervised pseudo-mask generation [46, 127, 95]. The mask generators can be graph-based (*e.g.*, Felzenszwalb-Huttenlocher algorithm [32]), MCG [1], or other saliency detection models [73, 69] trained on designated natural image datasets. However, those methods werer originally proposed for nature images, and their success for pathology images remains unknown.

Here we instantiate ConCL by using Felzenszwalb-Huttenlocher (FH) algorithm [32] and BASNet [73] as concept generators, dubbed as fh-ConCL and bas-ConCL, respectively. FH [32] is a conventional graph-based segmentation algorithm that relies on local neighborhoods, while BASNet [73] is a deep neural network pre-trained on a curated saliency detection dataset, which only contains daily natural objects. We use these two as representatives to study if these natural image priors win twice in both natural and pathology images.

Specifically, we use the FH algorithm in the scikit-image package and set both "scale" and "size" hyper-parameters to s. We use the pre-trained BASNet provided by [73]. Figure 4.2-(c-e) shows some examples. Table 4.1 reports the results.

It is not surprising that the BASNet [73] cannot generate decent concept masks (Fig. 4.2-(e)) for pathology images since it is pre-trained on curated saliency detection datasets. What is surprising is that bas-ConCL does yield satisfactory results (Table 4.1-(6)). Similar observations are also found in fh-ConCLs (Table 4.1-(4,5)) that though the generated concept masks are coarse-grained, the resulted transferring performances are unexpectedly good. After inspecting more examples, we find that the generated masks maintain high coherence and integrity despite their coarse-grained nature. That said, each concept contains semantic-consistent objects or textures. For example, Figure 4.2-(d,e) can be seen as special cases of Figure 4.2-(a) that merge fine-grained semantics with coarse-grained ones. This property makes the major difference between fh-/bas-ConCLs and g-ConCLs, where the grid-concepts are less likely to have coherent semantics.

Thus, we here conclude that *coherence matters* and natural image priors also work in pathology images, though they mostly provide coarse-grained concepts.

4.4.4 Pathology Image Priors in Pathology Images

Can we obtain concept masks away with natural image priors? External dependency is not always wanted and sometimes may fail to provide the desired masks (*e.g.*, Fig. 4.2-(e)). We thus task ourselves to find a dependency-free concept proposal method. One of the key characteristics in pathology images is that they have rich low-level patterns and tissue structures. Can we use that prior instead?

Figure 4.2-(f-h) shows the clustering visualization from intermediate feature maps generated by a 10-epoch warmed-up MoCo-v2 [18]. Thanks to the rich structural patterns in pathology images, we find that simply clustering over the feature maps provided by a barely trained model can already generate meaningful structural concept proposals. We thus build upon this "free lunch" and use a "bootstrap your own *perception*" mechanism that is similar to the "bootstrap your own latent" mechanism in BYOL [38]. ConCL generates concept proposals from the momentum key encoder's perception while simultaneously improving and refining it via the online query encoder, leading to a "bootstrapping" behavior. Thus, we denote such ConCL as bootstrapped-ConCL (b-ConCL).

b-ConCL. The concept generator is now instantiated as a KMeans grouper. We first pass the reference view x_r to the key encoder to obtain a reference feature map from ResNet stage-i: $f_i(x_r) \in \mathbb{R}^{CHW}$. Then, we apply K-Means to group K underlying concepts. b-ConCL relies on neither external segmentation algorithms nor designated saliency detection models for natural images.

Our default setting is $K = 8$, and clustering from f_4 or f_5. We postpone the study of hyper-parameters, *i.e.*, the number of clusters in KMeans, and the clustering stage f_i to Section 4.5.2 and report the main results in Table 4.1-(7,8). We find b-ConCL tops other entries. Compared to MoCo-v2 [18], our direct baseline, b-ConCL outperforms it by +4.5 AP^{bb} and +3.1 AP^{mk}. Moreover, b-ConCL obtains more gains in terms of AP_{75} (+6.2 AP^{bb}_{75}, +3.7 AP^{mk}_{75}) compared to MoCo-v2 [18], which means it improves MoCo-v2 [18] by more accurate bounding box regression and instance mask prediction. This aligns with our motivation for ConCL since discriminating local concepts helps shape object borders.

Closing remarks. So far, we have included: i) dense contrasting matters; ii) correspondence matters; iii) coherence matters; iv) natural image priors, though they might only provide coarse-grained concepts, work in pathology images as well; and find v) a randomly initialized or barely trained convolutional neural network, thanks to the rich low-level patterns in pathology images and good network initialization, can generate good proposals that are *dense*, *fine-grained* and *coherent*, as shown in Figure 4.2. Though the coarse-grained concepts generated from natural image priors could also help tasks in our studied benchmarks, they might underperform when a fine-grained dense prediction task is given. We hope our closing remarks could be intriguing and guide future works in designing dense pre-training methods for pathology images and beyond.

4.5 More Experiments

In the previous section, we have explored how we can obtain concepts, what concepts are good, and find b-ConCL to be the best. We here conduct more experiments to study b-ConCL.

4.5.1 Robustness to Transferring Settings

Transferring with different detectors. Here we investigate the transferring performance with other detectors, *i.e.*, Mask-RCNN-C4 (C4) [75] and RetinaNet [61]. RetinaNet is a single-stage detector. It uses ResNet-FPN backbone features as Mask-RCNN-FPN but directly generates predictions without region proposal [75]. C4 detector adopts a similar two-stage fashion as Mask-RCNN but uses the outputs of the 4-th residual block as backbone features and re-targets the 5-th block to be the detection head instead of building a new one. These three representative detectors evaluate pre-trained models under different detector architectures. Results together with Mask-RCNN-FPN's are shown in Table 4.2.

Detector	Pretrain	GlaS Detection		CRAG Detection	
		AP^{bb}	AP^{bb}_{75}	AP^{bb}	AP^{bb}_{75}
MaskRCNN-C4	Rand. Init.	52.9	59.9	49.4	54.2
	Supervised	49.1(-3.8)	55.1(-4.8)	46.1(-3.3)	50.6(-2.3)
	MoCo-v2 [18]	53.6(+0.7)	61.8(+1.9)	48.3(-1.1)	52.6(-1.6)
	b-ConCL	55.8(+2.9)	63.6(+3.7)	49.8(+0.4)	54.3(+0.1)
MaskRCNN-FPN	Rand. Init.	49.8	57.3	51.1	57.0
	Supervised	50.2(+0.4)	56.9(-0.4)	49.2(-1.9)	55.2(-1.8)
	MoCo-v2 [18]	52.3(+2.5)	60.0(+2.7)	50.0(-1.1)	55.7(-1.3)
	b-ConCL	56.8(+7.0)	66.2(+8.9)	55.1(+4.0)	62.2(+5.2)
RetinaNet	Rand. Init.	46.4	51.0	45.2	47.6
	Supervised	44.7(-1.7)	48.4(-2.6)	43.1(-2.1)	44.8(-2.8)
	MoCo-v2 [18]	47.2(+0.8)	50.9(-0.1)	43.1(-2.1)	43.8(-3.8)
	b-ConCL	52.6(+6.2)	58.6(+7.6)	48.4(+3.2)	51.9(+4.3)

Table 4.2: **Detection performance using different detectors.** Results are averaged over 5 trials.

b-ConCL performs the best with all three detectors in both datasets. Notably, training from scratch (Rand. Init.) is one of the top competitors when the C4 detector is used. We conjecture that the pre-trained models are possibly overfitted to their pretext tasks in their 5-th blocks and thus are harder to be tuned than a randomly initialized 5-th block. In CRAG detection, only b-ConCL pre-trained models consistently outperform randomly initialized models. In addition, the most significant gap between MoCo-v2[18] and b-ConCL is found in the RetinaNet detector [61]. As also noted by [64], RetinaNet [61] is a single-stage detector, where the local representations from the backbone become more important than other two-stage detectors since results are directly predicted from them. b-ConCL is tasked to discriminate local concepts, and subsequently, the learned representations could be better than other pre-training methods here.

Transferring with different schedules. To investigate if b-ConCL's lead could persist with longer fine-tuning, we fine-tune Mask-RCNN-FPN with 0.5×, 1×, 2×, 3×, and 5× schedules. Table 4.3 shows the results. b-ConCL maintains its noticeable gains in longer schedules in both datasets, *e.g.*, b-ConCL achieves 56.2 mAP with a 0.5× schedule, which is better than MoCo-v2 [18] with a 5× schedule but costs 10 × less fine-tuning time. Similar observations are also found in CRAG, where the gap between b-ConCL and MoCo-v2 [18] becomes larger (see Δ row). Together, these results confirm b-ConCL's superiority across different fine-tuning schedules.

Method	GlaS dataset					CRAG dataset				
	Fine-tuning schedule					Fine-tuning schedule				
	0.5×	1×	2×	3×	5×	0.5×	1×	2×	3×	5×
Rand. Init.	49.1	49.8	51.4	51.8	52.7	50.2	51.1	51.9	52.4	52.8
Supervised	48.6	50.2	51.4	52.7	54.0	50.0	49.2	50.5	50.1	50.3
MoCo-v2[18]	51.4	52.3	53.7	54.2	55.7	50.2	50.0	50.2	50.8	51.8
b-ConCL	**56.2**	**56.8**	**57.7**	**58.3**	**59.0**	**54.8**	**55.1**	**55.4**	**55.6**	**56.0**
Δ	+4.8	+4.5	+4.0	+4.1	+3.3	+4.6	+5.1	+5.2	+4.8	+4.2

Table 4.3: **Detection performance under different fine-tuning schedules.** Results other than 1× schedule are averaged over 3 runs. Δ row shows b-ConCL's improvement over MoCo-v2. We report AP^{bb} here.

4.5.2 Ablation Study

In this section, we ablate the key factors in b-ConCL. Our default setting clusters $K = 8$ concepts from ResNet stage-4 ($f_4(\cdot)$). Since b-ConCL is built on MoCo-v2 [18], we use it as our direct baseline for comparisons.

Concept loss weight λ. We here study the generalized concept contrastive loss: $L = (1-\lambda)L_q + \lambda L_c$, where $\lambda \in [0, 1]$ is a concept loss weight parameter. It shows a natural way to combine concept contrastive loss with instance contrastive loss. We start by asking whether instance contrastive loss is indispensable during the training process of b-ConCL. We alter the concept loss weight λ, and Table 4.4a reports the results. We see a monotonically increasing performance as λ increases in both datasets, which emphasizes the importance of concept loss. When no warm-up is used (last row in Table 4.4a), only a slight performance drop is observed, meaning that warm-up is not the key component of b-ConCL. Warming-up with instance loss (Eq. (4.2)) is a special case of b-ConCL, where at the early training stage, each instance is regarded as a concept, and we then gradually increase the number of concepts as training goes on. Thus, the overall findings in this ablation support b-ConCL's advance over MoCo-v2 [18].

Number of concepts K. Here, we investigate how the number of concepts clustered during pre-training affects performance in downstream tasks. We report the results of different K in Table 4.4b. b-ConCL performs reasonably well when $K >= 4$, with most of performance peaking at $K = 8$. This demonstrates the robustness of b-ConCL to the choice of K. Note that the best performance for the GlaS dataset is higher than our default setting and outperforms all entries in Table 4.1, showing the potential room for b-ConCL.

Where to group $f_i(\cdot)$. b-ConCL groups concepts from a model's intermediate feature maps. Our default setting uses feature maps from stage-4 of a ResNet [44], denoted as $f_4(\cdot)$. We now ablate this choice in Table 4.4c. Clustering concepts from $f_4(\cdot)$ and $f_5(\cdot)$ works similarly well across two datasets. We choose $f_4(\cdot)$ as the default since it achieves top two performance in both datasets under both metrics. Besides, b-ConCL exceeds MoCo-v2 [18], whichever stage it groups concepts from. This again confirms the effectiveness and robustness of b-ConCL.

Larger model capacity. Table 4.4d shows the results of using a larger backbone, ResNet-50. b-ConCL maintains its leading position. For consistency to the previous ablation, a $1\times$ schedule is also used here, which could put ResNet-50 at a disadvantage since it has more parameters to tune in a relatively short schedule.

λ	GlaS AP^{bb}	GlaS AP^{bb}_{75}	CRAG AP^{bb}	CRAG AP^{bb}_{75}
0.0	52.3	60.0	50.0	55.7
0.1	53.6	61.1	50.5	55.9
0.3	53.6	61.8	51.7	57.1
0.5	53.6	61.8	51.3	57.0
0.7	55.2	64.1	53.1	59.9
0.9	56.0	65.1	53.6	59.6
1.0	**56.8**	**66.2**	**55.1**	**62.2**
1.0\w.	56.1	**66.2**	54.0	60.6

(a) **Concept loss weight**.

K	GlaS AP^{bb}	GlaS AP^{bb}_{75}	CRAG AP^{bb}	CRAG AP^{bb}_{75}
1	52.3	60.0	50.0	55.7
2	54.5	64.1	52.9	60.1
4	55.6	64.7	53.4	59.7
6	56.3	65.1	53.7	60.2
8	56.8	**66.2**	**55.1**	**62.2**
10	57.0	66.0	**55.1**	61.0
12	**57.4**	**66.2**	54.2	60.1
16	55.7	65.3	54.5	61.3

(b) **Number of concepts**.

K	GlaS AP^{bb}	GlaS AP^{bb}_{75}	CRAG AP^{bb}	CRAG AP^{bb}_{75}
None	52.3	60.0	50.0	55.7
$f_1(\cdot)$	55.0	65.1	53.3	60.0
$f_2(\cdot)$	55.0	64.7	53.7	60.4
$f_3(\cdot)$	<u>56.2</u>	**66.4**	53.0	59.6
$f_4(\cdot)$	**56.8**	<u>66.2</u>	<u>55.1</u>	<u>62.2</u>
$f_5(\cdot)$	56.1	65.6	**56.5**	**63.3**

(c) **Clustering stage.**

GlaS Detection

Pretrain	ResNet-18 AP^{bb}	ResNet-18 AP^{bb}_{75}	ResNet-50 AP^{bb}	ResNet-50 AP^{bb}_{75}
Rand.	49.8	57.3	49.9	56.1
Sup.	50.2	56.9	47.9	54.2
MoCo.v2	52.3	60.0	53.1	60.5
b-ConCL	**56.8**	**66.2**	**57.0**	**65.9**

(d) **Backbone capacities.**

Table 4.4: **Ablation Study.** We study the effect of different hyper-parameters to b-ConCL. Default settings are marked in gray and MoCo-v2 baselines are marked by gray. In (a), "\w." means no warm-up.

4.6 Conclusion and Broader Impact

In this work, we benchmark various SSL methods for dense tasks in pathology images and introduce the ConCL framework. We identify several key components essential for successful transfer to dense tasks: i) dense contrasting, ii) correspondence, iii) coherence, and more.

Ultimately, we developed a dependency-free concept generator that bootstraps from inherent data concepts, demonstrating robustness and competitiveness.

Although our initial results focus on pre-training and fine-tuning, ConCL's applicability extends to tasks such as few-shot detection or segmentation, and semi-supervised learning. Furthermore, ConCL could be beneficial for speech or tabular data analysis, where minimal prior knowledge can be employed. Fine-grained "concepts" can be extracted using contrastive learning and clustering in these data modalities.

CHAPTER 5

Conclusion and Dicussion

This final chapter consolidates the work presented in this thesis, providing an overarching review of the findings, their implications, and the potential future avenues for this research. The primary objective of this thesis was to enhance the label efficiency and generalizability of deep learning models in the field of medical image analysis, specifically in the context of histopathology images. Our endeavors in this regard have been centered around two key strategies: data augmentation and self-supervised learning.

In Chapter 2, we delved into the integration of contrastive learning (CL) with latent augmentation (LA) to devise an efficient few-shot learning system. The findings from our experimental analysis highlighted the benefits of CL, including superior generalizability compared to traditional supervised learning models. Our work also extended the understanding of how and why CL-based models demonstrate better generalization. This exploration provides a solid foundation for further research into few-shot learning in histology images and has potential implications for other label-hungry domains.

Our discussion in Chapter 3 centered on the challenge of handling large, high-resolution whole-slide images (WSIs) in the context of deep multiple instance learning (MIL). We presented our solution, ReMix, which effectively enhanced training efficiency through instance reduction and ensured data diversity through bag-level augmentations. The success of ReMix across various MIL methods underscores its versatility and effectiveness, opening up possibilities for its broader application in slide-level WSI analysis.

In Chapter 4, we introduced Concept Contrastive Learning (ConCL), a new self-supervised learning (SSL) framework, and demonstrated its superiority over previous state-of-the-art

SSL methods through extensive experimental analysis. We outlined the path toward a more effective dense prediction pre-training method in pathological images and highlighted a simple, dependency-free, self-bootstrapping concept-generating method. This work offers valuable insights into SSL's potential in the context of dense prediction tasks in pathology images, contributing to a better understanding of the role of pre-training in computational pathology.

The research presented in this thesis has made several contributions to the field of medical image analysis. The methods and findings reported herein can significantly impact healthcare, particularly in improving the efficiency and effectiveness of pathological diagnosis. Nonetheless, while we have made progress in enhancing label efficiency and model generalization, there remains considerable scope for further research. Future work could delve into refining and extending the methods presented in this thesis and exploring their applicability in other medical imaging domains. By continuing to challenge the limitations of current models and innovate, we can hope to further enhance the contribution of deep learning to medical image analysis and, by extension, healthcare outcomes.

Bibliography

[1] Pablo Arbeláez, Jordi Pont-Tuset, Jonathan T Barron, Ferran Marques, and Jitendra Malik. Multiscale combinatorial grouping. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 328–335, 2014.

[2] Philip Bachman, Ouais Alsharif, and Doina Precup. Learning with pseudo-ensembles. *Advances in neural information processing systems*, 27:3365–3373, 2014.

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

[4] Yoshua Bengio, Grégoire Mesnil, Yann Dauphin, and Salah Rifai. Better mixing via deep representations. In *International conference on machine learning*, pages 552–560. PMLR, 2013.

[5] Luca Bertero, Carlo Alberto Barbano, Daniele Perlo, Enzo Tartaglione, Paola Cassoni, Marco Grangetto, Attilio Fiandrotti, Alessandro Gambella, and Luca Cavallo. Unitopatho, 2021.

[6] David Berthelot, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Kihyuk Sohn, Han Zhang, and Colin Raffel. Remixmatch: Semi-supervised learning with distribution alignment and augmentation anchoring. *arXiv preprint arXiv:1911.09785*, 2019.

[7] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin Raffel. Mixmatch: A holistic approach to semi-supervised learning. *arXiv preprint arXiv:1905.02249*, 2019.

[8] Andrew A. Borkowski, Marilyn M. Bui, L. Brannon Thomas, Catherine P. Wilson, Lauren A. DeLand, and Stephen M. Mastorides. Lc25000 lung and colon histopathological image dataset. 2019.

[9] 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.

[10] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 132–149, 2018.

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

[12] Chi-Long Chen, Chi-Chung Chen, Wei-Hsiang Yu, Szu-Hua Chen, Yu-Chan Chang, Tai-I Hsu, Michael Hsiao, Chao-Yuan Yeh, and Cheng-Yu Chen. An annotation-free whole-slide training approach to pathological classification of lung cancer types using deep learning. *Nature communications*, 12(1):1–13, 2021.

[13] Hanbo Chen, Xiao Han, Xinjuan Fan, Xiaoying Lou, Hailing Liu, Junzhou Huang, and Jianhua Yao. Rectified cross-entropy and upper transition loss for weakly supervised whole slide image classifier. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 351–359. Springer, 2019.

[14] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[15] Ting Chen, Calvin Luo, and Lala Li. Intriguing properties of contrastive losses. *arXiv preprint arXiv:2011.02803*, 2020.

[16] Wei-Yu Chen, Yen-Cheng Liu, Zsolt Kira, Yu-Chiang Frank Wang, and Jia-Bin Huang. A closer look at few-shot classification. In *International Conference on Learning Representations*, 2019.

[17] Xiaocong Chen, Lina Yao, Tao Zhou, Jinming Dong, and Yu Zhang. Momentum contrastive learning for few-shot covid-19 diagnosis from chest ct images. *Pattern recognition*, 113:107826, 2021.

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

[19] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15750–15758, 2021.

[20] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. *arXiv preprint arXiv:2104.02057*, 2021.

[21] Yinbo Chen, Xiaolong Wang, Zhuang Liu, Huijuan Xu, and Trevor Darrell. A new meta-baseline for few-shot learning. *arXiv preprint arXiv:2003.04390*, 2020.

[22] Zhen Chen, Jun Zhang, Shuanlong Che, Junzhou Huang, Xiao Han, and Yixuan Yuan. Diagnose like a pathologist: Weakly-supervised pathologist-tree network for slide-level immunohistochemical scoring. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 47–54, 2021.

[23] Veronika Cheplygina, Lauge Sørensen, David MJ Tax, Marleen de Bruijne, and Marco Loog. Label stability in multiple instance learning. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 539–546. Springer, 2015.

[24] Tsz-Him Cheung and Dit-Yan Yeung. Modals: Modality-agnostic automated data augmentation in the latent space. In *International Conference on Learning Representations*, 2020.

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

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

[27] MMSelfSup Contributors. MMSelfSup: Openmmlab self-supervised learning toolbox and benchmark. https://github.com/open-mmlab/mmselfsup, 2021.

[28] Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation strategies from data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 113–123, 2019.

[29] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 702–703, 2020.

[30] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017.

[31] Zach Eaton-Rosen, Felix Bragman, Sebastien Ourselin, and M Jorge Cardoso. Improving data augmentation for medical image segmentation. 2018.

[32] Pedro F Felzenszwalb and Daniel P Huttenlocher. Efficient graph-based image segmentation. *International journal of computer vision*, 59(2):167–181, 2004.

[33] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning*, pages 1126–1135. PMLR, 2017.

[34] Maayan Frid-Adar, Idit Diamant, Eyal Klang, Michal Amitai, Jacob Goldberger, and Hayit Greenspan. Gan-based synthetic medical image augmentation for increased cnn performance in liver lesion classification. *Neurocomputing*, 321:321–331, 2018.

[35] Golnaz Ghiasi, Yin Cui, Aravind Srinivas, Rui Qian, Tsung-Yi Lin, Ekin D Cubuk, Quoc V Le, and Barret Zoph. Simple copy-paste is a strong data augmentation method for instance segmentation. *arXiv preprint arXiv:2012.07177*, 2020.

[36] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.

[37] Simon Graham, Hao Chen, Jevgenij Gamper, Qi Dou, Pheng-Ann Heng, David Snead, Yee Wah Tsang, and Nasir Rajpoot. Mild-net: Minimal information loss dilated network for gland instance segmentation in colon histology images. *Medical image analysis*, 52:199–211, 2019.

[38] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent: A new approach to self-supervised learning. *arXiv preprint arXiv:2006.07733*, 2020.

[39] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742. IEEE, 2006.

[40] Bharath Hariharan and Ross Girshick. Low-shot visual recognition by shrinking and hallucinating features. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3018–3027, 2017.

[41] Noriaki Hashimoto, Daisuke Fukushima, Ryoichi Koga, Yusuke Takagi, Kaho Ko, Kei Kohno, Masato Nakaguro, Shigeo Nakamura, Hidekata Hontani, and Ichiro Takeuchi. Multi-scale domain-adversarial multiple-instance cnn for cancer subtype classification with unannotated histopathological images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3852–3861, 2020.

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

[43] Kaiming He, Ross Girshick, and Piotr Dollár. Rethinking imagenet pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4918–4927, 2019.

[44] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[45] Katherine A Heller, Adam Sanborn, and Nick Chater. Hierarchical learning of dimensional biases in human categorization. In *NIPS*, pages 727–735. Citeseer, 2009.

[46] Olivier J Hénaff, Skanda Koppula, Jean-Baptiste Alayrac, Aaron van den Oord, Oriol Vinyals, and João Carreira. Efficient visual pretraining with contrastive detection. *arXiv preprint arXiv:2103.10957*, 2021.

[47] Le Hou, Dimitris Samaras, Tahsin M Kurc, Yi Gao, James E Davis, and Joel H Saltz. Patch-based convolutional neural network for whole slide tissue image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2424–2433, 2016.

[48] Yue Huang, Han Zheng, Chi Liu, Xinghao Ding, and Gustavo K Rohde. Epithelium-stroma classification via convolutional neural networks and unsupervised domain adaptation in histopathological images. *IEEE journal of biomedical and health informatics*, 21(6):1625–1632, 2017.

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

[50] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 2019.

[51] Jakob Nikolas Kather, Niels Halama, and Alexander Marx. 100,000 histological images of human colorectal cancer and healthy tissue, April 2018.

[52] Yoo Jung Kim, Hyungjoon Jang, Kyoungbun Lee, Seongkeun Park, Sung-Gyu Min, Choyeon Hong, Jeong Hwan Park, Kanggeun Lee, Jisoo Kim, Wonjae Hong, et al. Paip 2019: Liver cancer segmentation challenge. *Medical Image Analysis*, 67:101854, 2021.

[53] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

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

[55] Anders Krogh and John A Hertz. A simple weight decay can improve generalization. In *Advances in neural information processing systems*, pages 950–957, 1992.

[56] Michael Kuchnik and Virginia Smith. Efficient augmentation via data subsampling. *arXiv preprint arXiv:1810.05222*, 2018.

[57] Kwonjoon Lee, Subhransu Maji, Avinash Ravichandran, and Stefano Soatto. Meta-learning with differentiable convex optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10657–10665, 2019.

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

[59] Jiahui Li, Wen Chen, Xiaodi Huang, Shuang Yang, Zhiqiang Hu, Qi Duan, Dimitris N Metaxas, Hongsheng Li, and Shaoting Zhang. Hybrid supervision learning for pathology whole slide image classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 309–318. Springer, 2021.

[60] Junnan Li, Pan Zhou, Caiming Xiong, and Steven CH Hoi. Prototypical contrastive learning of unsupervised representations. *arXiv preprint arXiv:2005.04966*, 2020.

[61] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[62] Jialun Liu, Yifan Sun, Chuchu Han, Zhaopeng Dou, and Wenhui Li. Deep representation learning on long-tailed data: A learnable embedding augmentation perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2970–2979, 2020.

[63] Mingxia Liu, Jun Zhang, Ehsan Adeli, and Dinggang Shen. Landmark-based deep multi-instance learning for brain disease diagnosis. *Medical image analysis*, 43:157–168, 2018.

[64] Songtao Liu, Zeming Li, and Jian Sun. Self-emd: Self-supervised object detection without imagenet. *arXiv preprint arXiv:2011.13677*, 2020.

[65] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.

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

[67] Kushagra Mahajan, Monika Sharma, and Lovekesh Vig. Meta-dermdiagnosis: few-shot skin disease identification using meta-learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 730–731, 2020.

[68] Alfonso Medela, Artzai Picon, Cristina L Saratxaga, Oihana Belar, Virginia Cabezón, Riccardo Cicchi, Roberto Bilbao, and Ben Glover. Few shot learning in histopathological images: reducing the need of labeled data on biological datasets. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pages 1860–1864. IEEE, 2019.

[69] Duc Tam Nguyen, Maximilian Dax, Chaithanya Kumar Mummadi, Thi Phuong Nhung Ngo, Thi Hoai Phuong Nguyen, Zhongyu Lou, and Thomas Brox. Deepusps: Deep robust unsupervised saliency prediction with self-supervision. *arXiv preprint arXiv:1909.13055*, 2019.

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

[71] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[72] Tingying Peng, Melanie Boxberg, Wilko Weichert, Nassir Navab, and Carsten Marr. Multi-task learning of a deep k-nearest neighbour network for histopathological image classification and retrieval. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 676–684. Springer, 2019.

[73] Xuebin Qin, Zichen Zhang, Chenyang Huang, Chao Gao, Masood Dehghan, and Martin Jagersand. Basnet: Boundary-aware salient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7479–7489, 2019.

[74] Aniruddh Raghu, Maithra Raghu, Samy Bengio, and Oriol Vinyals. Rapid learning or feature reuse? towards understanding the effectiveness of maml. *arXiv preprint arXiv:1909.09157*, 2019.

[75] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28:91–99, 2015.

[76] Byungseok Roh, Wuhyun Shin, Ildoo Kim, and Sungwoong Kim. Spatially consistent representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1144–1153, 2021.

[77] Andrei A Rusu, Dushyant Rao, Jakub Sygnowski, Oriol Vinyals, Razvan Pascanu, Simon Osindero, and Raia Hadsell. Meta-learning with latent embedding optimization. *arXiv preprint arXiv:1807.05960*, 2018.

[78] Ruslan Salakhutdinov, Joshua Tenenbaum, and Antonio Torralba. One-shot learning with a hierarchical nonparametric bayesian model. In *Proceedings of ICML Workshop on Unsupervised and Transfer Learning*, pages 195–206. JMLR Workshop and Conference Proceedings, 2012.

[79] Jürgen Schmidhuber. *Evolutionary principles in self-referential learning, or on learning how to learn: the meta-meta-... hook*. PhD thesis, Technische Universität München, 1987.

[80] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.

[81] Eli Schwartz, Leonid Karlinsky, Joseph Shtok, Sivan Harary, Mattias Marder, Abhishek Kumar, Rogerio Feris, Raja Giryes, and Alex Bronstein. Delta-encoder: an effective sample synthesis method for few-shot object recognition. *Advances in Neural Information Processing Systems*, 31, 2018.

[82] Fereshteh Shakeri, Malik Boudiaf, Sina Mohammadi, Ivaxi Sheth, Mohammad Havaei, Ismail Ben Ayed, and Samira Ebrahimi Kahou. Fhist: A benchmark for few-shot classification of histological images. 2021.

[83] Zhuchen Shao, Hao Bian, Yang Chen, Yifeng Wang, Jian Zhang, Xiangyang Ji, et al. Transmil: Transformer based correlated multiple instance learning for whole slide image classification. *Advances in Neural Information Processing Systems*, 34:2136–2147, 2021.

[84] Yash Sharma, Aman Shrivastava, Lubaina Ehsan, Christopher A Moskaluk, Sana Syed, and Donald Brown. Cluster-to-conquer: A framework for end-to-end multi-instance learning for whole slide image classification. In *Medical Imaging with Deep Learning*, pages 682–698. PMLR, 2021.

[85] Xiahan Shi, Leonard Salewski, Martin Schiegg, and Max Welling. Relational generalized few-shot learning. In *31st British Machine Vision Conference 2020, BMVC 2020, Virtual Event, UK, September 7-10, 2020*. BMVA Press, 2020.

[86] Hoo-Chang Shin, Neil A Tenenholtz, Jameson K Rogers, Christopher G Schwarz, Matthew L Senjem, Jeffrey L Gunter, Katherine P Andriole, and Mark Michalski. Medical image synthesis for data augmentation and anonymization using generative adversarial networks. In *International workshop on simulation and synthesis in medical imaging*, pages 1–11. Springer, 2018.

[87] Connor Shorten and Taghi M Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1):1–48, 2019.

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

[89] Jake Snell, Kevin Swersky, and Richard S Zemel. Prototypical networks for few-shot learning. *arXiv preprint arXiv:1703.05175*, 2017.

[90] Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *arXiv preprint arXiv:2001.07685*, 2020.

[91] Chetan L Srinidhi, Ozan Ciga, and Anne L Martel. Deep neural network models for computational histopathology: A survey. *Medical Image Analysis*, page 101813, 2020.

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

[93] Yonglong Tian, Yue Wang, Dilip Krishnan, Joshua B Tenenbaum, and Phillip Isola. Rethinking few-shot image classification: a good embedding is all you need? In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, pages 266–282. Springer, 2020.

[94] Paul Upchurch, Jacob Gardner, Geoff Pleiss, Robert Pless, Noah Snavely, Kavita Bala, and Kilian Weinberger. Deep feature interpolation for image content changes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7064–7073, 2017.

[95] Wouter Van Gansbeke, Simon Vandenhende, Stamatios Georgoulis, and Luc Van Gool. Unsupervised semantic segmentation by contrasting object mask proposals. *arXiv preprint arXiv:2102.06191*, 2021.

[96] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pages 1096–1103, 2008.

[97] Sophia J Wagner, Christian Matek, Sayedali Shetab Boushehri, Melanie Boxberg, Lorenz Lamm, Ario Sadafi, Dominik JE Waibel, Carsten Marr, and Tingying Peng. Make deep learning algorithms in computational pathology more reproducible and reusable. *Nature Medicine*, pages 1–3, 2022.

[98] Wenguan Wang, Tianfei Zhou, Fisher Yu, Jifeng Dai, Ender Konukoglu, and Luc Van Gool. Exploring cross-image pixel contrast for semantic segmentation. *arXiv preprint arXiv:2101.11939*, 2021.

[99] Xinlong Wang, Rufeng Zhang, Chunhua Shen, Tao Kong, and Lei Li. Dense contrastive learning for self-supervised visual pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3024–3033, 2021.

[100] Yan Wang, Wei-Lun Chao, Kilian Q Weinberger, and Laurens van der Maaten. Simpleshot: Revisiting nearest-neighbor classification for few-shot learning. *arXiv preprint arXiv:1911.04623*, 2019.

[101] Yu-Xiong Wang, Ross Girshick, Martial Hebert, and Bharath Hariharan. Low-shot learning from imaginary data. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7278–7286, 2018.

[102] Yulin Wang, Gao Huang, Shiji Song, Xuran Pan, Yitong Xia, and Cheng Wu. Regularizing deep networks with semantic data augmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[103] Yulin Wang, Xuran Pan, Shiji Song, Hong Zhang, Gao Huang, and Cheng Wu. Implicit semantic data augmentation for deep networks. *Advances in Neural Information Processing Systems*, 32:12635–12644, 2019.

[104] Sen Wu, Hongyang Zhang, Gregory Valiant, and Christopher Ré. On the generalization effects of linear transformations in data augmentation. In *International Conference on Machine Learning*, pages 10410–10420. PMLR, 2020.

[105] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3733–3742, 2018.

[106] Yongqin Xian, Christoph H Lampert, Bernt Schiele, and Zeynep Akata. Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. *IEEE transactions on pattern analysis and machine intelligence*, 41(9):2251–2265, 2018.

[107] Chensu Xie, Hassan Muhammad, Chad M Vanderbilt, Raul Caso, Dig Vijay Kumar Yarlagadda, Gabriele Campanella, and Thomas J Fuchs. Beyond classification: Whole slide tissue histopathology analysis by end-to-end part learning. In *Medical Imaging with Deep Learning*, pages 843–856. PMLR, 2020.

[108] Enze Xie, Jian Ding, Wenhai Wang, Xiaohang Zhan, Hang Xu, Peize Sun, Zhenguo Li, and Ping Luo. Detco: Unsupervised contrastive learning for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8392–8401, 2021.

[109] Zhenda Xie, Yutong Lin, Zheng Zhang, Yue Cao, Stephen Lin, and Han Hu. Propagate yourself: Exploring pixel-level consistency for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16684–16693, 2021.

[110] Yan Xu, Tao Mo, Qiwei Feng, Peilin Zhong, Maode Lai, I Eric, and Chao Chang. Deep learning of feature representation with multiple instance learning for medical image analysis. In *2014 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 1626–1630. IEEE, 2014.

[111] Yuan Xue, Qianying Zhou, Jiarong Ye, L Rodney Long, Sameer Antani, Carl Cornwell, Zhiyun Xue, and Xiaolei Huang. Synthetic augmentation and feature-based filtering for improved cervical histopathology image classification. In *International conference on medical image computing and computer-assisted intervention*, pages 387–396. Springer, 2019.

[112] Jiangpeng Yan, Hanbo Chen, Xiu Li, and Jianhua Yao. Deep contrastive learning based tissue clustering for annotation-free histopathology image analysis. *Computerized Medical Imaging and Graphics*, 97:102053, 2022.

[113] Ceyuan Yang, Zhirong Wu, Bolei Zhou, and Stephen Lin. Instance localization for self-supervised detection pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3987–3996, 2021.

[114] 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 *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXI*, pages 523–539. Springer, 2022.

[115] Jiawei Yang, Hanbo Chen, Jiangpeng Yan, Xiaoyu Chen, and Jianhua Yao. Towards better understanding and better generalization of low-shot classification in histology images with contrastive learning. In *International Conference on Learning Representations*, 2022.

[116] Jiawei Yang, Hanbo Chen, Yu Zhao, Fan Yang, Yao Zhang, Lei He, and Jianhua Yao. Remix: A general and efficient framework for multiple instance learning based whole slide image classification. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part II*, pages 35–45. Springer, 2022.

[117] Jiawei Yang, Yao Zhang, Yuan Liang, Yang Zhang, Lei He, and Zhiqiang He. Tumorcp: A simple but effective object-level data augmentation for tumor segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 579–588. Springer, 2021.

[118] Shuo Yang, Lu Liu, and Min Xu. Free lunch for few-shot learning: Distribution calibration. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

[119] Jiawen Yao, Xinliang Zhu, Jitendra Jonnagaddala, Nicholas Hawkins, and Junzhou Huang. Whole slide images based cancer survival prediction using attention guided deep multiple instance learning networks. *Medical Image Analysis*, 65:101789, 2020.

[120] Xi Yin, Xiang Yu, Kihyuk Sohn, Xiaoming Liu, and Manmohan Chandraker. Feature transfer learning for face recognition with under-represented data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5704–5713, 2019.

[121] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6023–6032, 2019.

[122] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

[123] Ling Zhang, Xiaosong Wang, Dong Yang, Thomas Sanford, Stephanie Harmon, Baris Turkbey, Bradford J Wood, Holger Roth, Andriy Myronenko, Daguang Xu, et al. Generalizing deep learning for medical image segmentation to unseen domains via deep stacked transformation. *IEEE transactions on medical imaging*, 39(7):2531–2540, 2020.

[124] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *European conference on computer vision*, pages 649–666. Springer, 2016.

[125] Richard Zhang, Phillip Isola, and Alexei A Efros. Split-brain autoencoders: Unsupervised learning by cross-channel prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1058–1067, 2017.

[126] Xinyu Zhang, Qiang Wang, Jian Zhang, and Zhao Zhong. Adversarial autoaugment. *arXiv preprint arXiv:1912.11188*, 2019.

[127] Nanxuan Zhao, Zhirong Wu, Rynson WH Lau, and Stephen Lin. Distilling localization for self-supervised representation learning. *arXiv preprint arXiv:2004.06638*, 2020.

[128] Xiangyun Zhao, Raviteja Vemulapalli, Philip Andrew Mansfield, Boqing Gong, Bradley Green, Lior Shapira, and Ying Wu. Contrastive learning for label efficient semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10623–10633, 2021.