TOWARDS BETTER UNDERSTANDING AND BETTER GENERALIZATION OF LOW-SHOT CLASSIFICATION IN HISTOLOGY IMAGES WITH CONTRASTIVE LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Low-shot learning is an established topic in natural images for years, but few work is attended to histology images, which is of high clinical value since well-labeled datasets and rare abnormal samples are expensive to collect. Here, we pioneer the study of low-shot learning for histology images by setting up three cross-domain tasks that reflect real clinics problems. To enable label-efficient learning and better generalizability, we propose to incorporate *contrastive learning* (CL) with *latent augmentation* (LA) to build a few-shot system. CL learns useful representations without manual labels, while LA transfers semantic variations of base dataset in an unsupervised way. These two components fully exploit unlabeled training data and can scale gracefully to other label-hungry problems. In experiments, we find: i) models learned by CL generalize fairly better than supervised learning for histology images, and ii) LA brings consistent gains over baselines. Prior studies of self-supervised representation learning mainly focus on ImageNet-like images, which only present a dominant object in their centers. Recent attention has been paid to images with multi-objects and multi-textures (Chen & Li, 2020). Histology images are a natural choice for such study. We show the superiority of CL over supervised learning in terms of generalization for such data, and provide our empirical understanding for this observation. The findings in this work could contribute to understanding how model generalizes in the context of both self-supervised learning and histology images.

1 INTRODUCTION

Histological images provide crucial phenotypical and diagnostic information for disease assessment and prognosis (Srinidhi et al., 2020). Their examination, which nevertheless requires intensive workload and extensive expertise, is an indispensable component in practical clinics. Thus, computer-aided histological image classification systems are highly demanded but expensive to build due to the scarcity of well-annotated data. Besides, histological images diversify in many aspects, including acquisition protocols, body sites and tissue types. Such heavy domain shifts and variations pose more challenges and aggravate the data-hungry issues for histological images. How to train robust models with limited access to annotated samples becomes the key of general diagnosis systems.

The aforementioned issues have motivated us to sort for *low-shot learning*, e.g. few-shot learning (FSL) and generalized few-shot learning (GFSL), which aim to learn new concepts from only limited labeled data. Recent works have demonstrated the success of few-shot systems in natural images, but they are much unexplored in medical images, especially in histological ones. Here, we pioneer the study of GFSL and FSL for histology images by setting up 3 cross-domain tasks where there exist near-, middle- and out-domain shifts from base class to novel class. In addition, we also investigate the impact of homogeneous and heterogeneous shot selection, i.e. few-shot samples come from the same whole slide image (WSI) or different ones (more on this later), to few-shot system.

To enable label-efficient learning and better generalizability, we propose to incorporate *contrastive learning* (CL) with *latent augmentation* (LA) to build a few-shot system. Concretely, CL learns a meaningful encoder during pre-training stage without relying on label information, while LA inherits knowledge from "unlabeled" base dataset by transferring semantic variations in latent space. These two components fully exploit a base dataset by using its legacy: the learned model weights, and the captured latent variations.

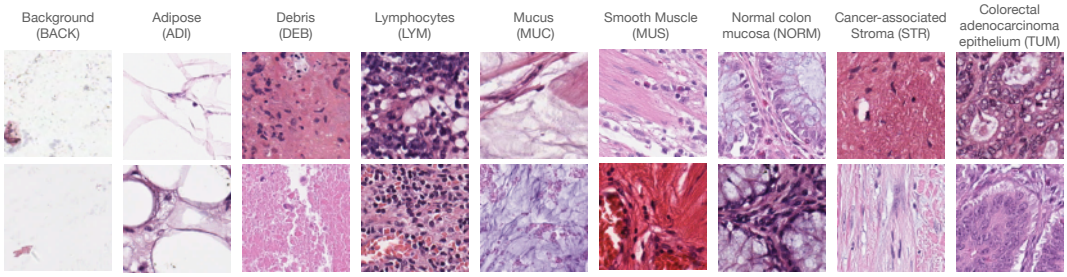


Figure 1: **Example images from NCT (Kather et al., 2018).** Each column contains two samples from the same class (column name). More example images can be found in Appendix A.1.

Also to our surprise, we show that state-of-the-art CL models generalize fairly better than the supervised counterparts for histological images. Previously studies of CL mainly focus on *iconic* natural images where only a dominant object occupies image centers, while images like histological ones, where multiple small objects (e.g. cells, nucleus) and various textures (e.g. muscle, mucus) are densely presented, are largely unexplored. We also aim to fill the gap of studying CL for *non-iconic, multi-object and multi-texture* images, and empirically explain why the large generalization gap between CL models and supervised ones exists for such images.

To summarize, our key findings and contributions are:

- We pioneer the study of low-shot classification problem for histological images. Extensive investigations w.r.t. domain-specific problems, i.e. domain shifts across tissue types and body sites, heterogeneous and homogeneous few-shot samples selection, are conducted.
- We propose a simple label-efficient method for low-shot learning, which incorporates contrastive learning and latent augmentation to fully exploit training data in an unsupervised way. Extensive experiments confirm their consistent gains and improved generalizability.
- We show that, at slight odds with findings in iconic natural images, contrastive learning learned models generalize fairly better than supervised counterparts for *non-iconic, multi-object and multi-texture* histology images. We analyze and provide our empirical explanation for this observation, which we believe could contribute to understanding how model generalizes to novel samples in the context of contrastive representation learning and histology images.

2 Preliminaries and Problem Formulation

Whole-Slide Image (WSI). Whole-slide images are digital scans of pathology tissue slides collected by biopsy or surgery. Given micron-size pixels and centimeter-size slides, a WSI is usually of Gigapixel size and thus have to be divided into hundreds or thousands of small "patches" for computational analysis. A patch reflects the basic unit in patch-level classification problem. As part of tissue processing protocols, a slide needs to be stained before imaged for diagnosis. The tissues context and staining quality could vary across WSIs, so the extracted patches' styles are confronted to their source WSIs, leading to inter-WSI domain shift. Besides, unlike iconic natural images which only present a dominant object in their centers, histological patches contain multiple small objects and multiple texture-like tissues (see Figure 1). To diagnose, a pathologist needs to inspect both the local cellular details and the overall global patterns, making the process different from the major recognition systems which only need to focus on dominant objects.

Few-shot Learning (FSL). Few-shot classification is to learn from a large "base" dataset, and afterwards generalize to unseen classes with only limited labeled data. Formally, a 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 a sample set and $\mathcal{Y}_{base}$ is their label space. 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 $\mathcal{Y}_{novel}$ is the novel label space. A few-shot learner is trained on $\mathcal{D}_{base}$ and evaluated on a series of *meta-tasks* sampled from $\mathcal{D}_{novel}$. Such 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, called *support* set, and $\mathcal{Q} = \{\mathbf{x}_i\}_{i=1}^{NQ} \sim \mathcal{X}_{novel}$ is another small test set, called *query* set, and I is the number of tasks. This formulation is termed as N-way K-shot (Q-query) task, since N classes are sampled from $\mathcal{Y}_{novel}$, each with K labeled

samples for training and Q unlabeled samples for testing. Usually, K is less than Q, e.g. $K = 1, 5$, and $Q = 15$. The evaluation stage is often referred as meta-testing stage.

Generalized few-shot learning (GFSL). Unlike in FSL, GFSL samples meta-tasks from a joint dataset $\mathcal{D}_{joint} = \mathcal{D}_{base} \cup \mathcal{D}_{novel}$, with a joint label space $\mathcal{Y}_{joint} = \mathcal{Y}_{base} \cup \mathcal{Y}_{novel}$. Now, support sets and query sets contain both seen base classes and unseen novel classes.

3 METHODS

Consider a few-shot classifier as $f = f_\theta \circ f_\phi$[1], where f_ϕ is a embedding function, i.e. a feature extractor, that maps the high-dimensional input image data $\mathbf{x}$ into a low-dimensional latent space $\mathbb{R}^d$, and f_θ is a classifier trained on the support set $\mathcal{S}$ and predicts the results for query set $\mathcal{Q}$. ϕ and θ are their corresponding parameters. 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 meta-testing stage. Figure 2 shows the overview of our methods. We elaborate them in the followings.

3.1 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 do standard fully supervised classification pre-training without splitting data. The former one emphasizes the idea of meta-learning, i.e. "learning to learn" for fast adaption and generalization (Schmidhuber, 1987; Finn et al., 2017), while the latter one attributes the success of FSL to feature reuse (Raghu et al., 2019) and good representations (Chen et al., 2019; Tian et al., 2020). We follow the latter for simplicity and we believe better learned encoders lead to stronger generalizability. Figure 2-(a) shows the pre-training step.

Fully-supervised pre-training (FSP). Given a base dataset $\mathcal{D}_{base} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N_{base}}$, we jointly train the feature extractor f_ϕ and a proxy classifier f_ψ using standard cross-entropy loss:

$$\phi^*, \psi^* = \arg\min_{\phi,\psi} \mathbb{E}_{\mathbf{x},y\sim\mathcal{D}_{base}} \left[y \log(f_\psi \circ f_\phi(\mathbf{x})|\phi, \psi) \right]. \tag{1}$$

The proxy classifier f_ψ is then deprecated for downstream evaluation, while the feature extractor f_ϕ is kept and *fixed* for all tasks[2]. We term the embedding functions learned by FSP as f_ϕ^{FSP}.

Contrastive-learning pre-training (CLP). Self-supervised learning methods alleviate the need for data annotations. Here we focus on a contrastive learning method – MoCo-v3 (Chen et al., 2021b), which currently holds state-of-the-art performance. It 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 $\mathcal{D}_{base}^u = \{\mathbf{x}_i\}_{i=1}^{N_{base}}$, the model learns to minimize the contrastive loss function w.r.t. unlabeled batch data:

$$\phi^*, g^*, q^* = \arg\min_{\phi,g,q} \mathbb{E}_{\mathbf{x},\mathbf{x}' \overset{t}{\sim} \mathcal{D}_{base}^u} \left[\mathcal{L}_{CLP} \left(f_q \circ f_g \circ f_\phi(\mathbf{x}), f_{\tilde{g}} \circ f_{\tilde{\phi}}(\mathbf{x}'); \phi, g, q \right) \right], \tag{2}$$

where $\mathcal{L}_{CLP}$ is a contrastive loss function; $\mathbf{x}, \mathbf{x}'$ are two views of the same images obtained by applying random data augmentation t; $\tilde{\phi}$ and $\tilde{g}$ denote the momentum updated copies of ϕ and g. In short, contrastive learning aims to maximize the similarity between positive pairs (two augmented views of a same image), while minimize the similarity between negative pairs (two different images). We leave more detailed description of MoCo-v3 and $\mathcal{L}_{CLP}$ to Appendix B. Once CLP is done, two auxiliary heads, f_g and f_q, are deprecated, while f_ϕ is kept and *fixed*, termed as f_ϕ^{CLP}.

3.2 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 (Wang et al., 2018) since they are all pathology-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.

[1]Operator $\circ$ denotes a composite function, e.g. $g \circ f(x) = g(f(x))$.

[2]Without loss of generality, we abbreviate all the * symbols in body text, e.g. $\phi \leftarrow \phi^*$.

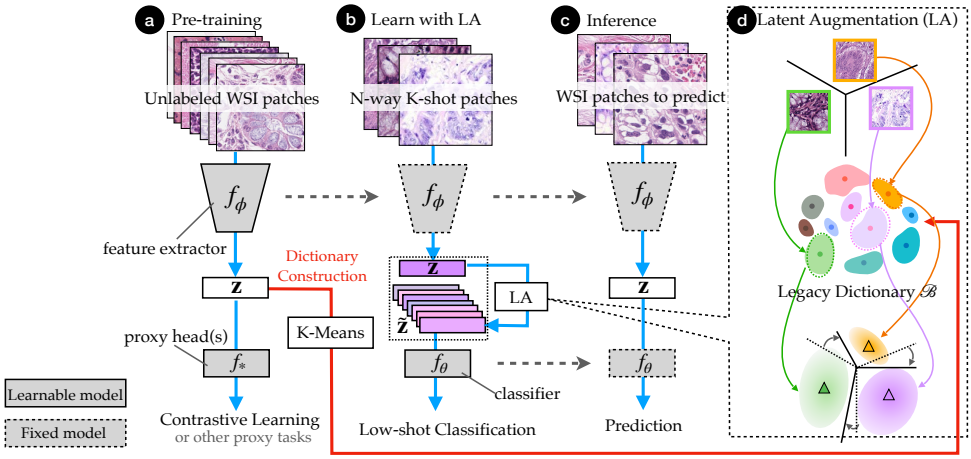


Figure 2: **Overview.** With pre-trained feature extractor (a), N-way K-shot classifiers are learnt (b) based on LA (d) to classify WSI patches (c). Given a novel representation $\mathbf{z}$, LA generates its new features from the most likely variation in the legacy dictionary, so few-shot novel samples can be proliferated in a reasonable way, and the decision boundary can therefore be improved.

Legacy dictionary and Latent augmentation (LA). Our goal is to fully exploit training data. This is done not only by reusing pre-trained model weights f_ϕ, but also by transferring possible semantic shifts of clustered representations. Given an unlabeled base dataset, we perform K-Means on the representations $\mathbf{z} = f_\phi(\mathbf{x})$ to obtain C clusters (Figure 2-(a), red arrows). The *legacy dictionary* is constructed as $\mathcal{B} = \{(\mathbf{c}_i, \mathbf{\Sigma}_i)\}_{i=1}^{C}$, where $\mathbf{c}_i$ is the i-th cluster prototype (i.e. average representation) and $\mathbf{\Sigma}_i$ denotes its intra-cluster covariance matrix. Roughly, $\mathcal{B}$ captures how the pre-trained f_ϕ thinks samples from base dataset will vary in latent space conditioned on cluster i, e.g. using a multivariate Gaussian $\mathcal{N}(\mathbf{c}_i, \mathbf{\Sigma}_i)$. Given the legacy dictionary $\mathcal{B}$, during meta-testing stage, LA uses original representations $\mathbf{z}$ to query the most likely variations from $\mathcal{B}$, followed by additive augmentation $\tilde{\mathbf{z}} = \mathbf{z} + \boldsymbol{\delta}$ (Figure 2-(c)). This is done by sampling $\boldsymbol{\delta} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_{i^*})$ where $i^* = \arg\max_i \mathbf{z} \cdot \mathbf{c}_i / \|\mathbf{z}\| \|\mathbf{c}_i\|$. The classifier f_θ is then trained on both the original representations $\mathbf{z}$ and the augmented representations $\tilde{\mathbf{z}}$ (Figure 2-(b)).

3.2.1 Intuitions and motivations on latent augmentation

Why transferring variations works. LA aims at transferring the knowledge of variations. Such knowledge brings *semantic* diversity from base class to novel class. For example, tumorous cells are mutated from normal cells; when given limited tumorous samples, LA can replicate how normal cells alter under this mutation using the captured variations in legacy dictionaries. This closely resembles how a pathologist builds his/her knowledge on unseen phenotype from seen phenotype. From the view of under-representation learning (Yin et al., 2019), replicating latent variations encourages under-represented distributions to be closer to regular ones. From the view of low-data learning, distribution of few samples is not well calibrated (Yang et al., 2021), so using established distributions in base class to calibrate untrustworthy novel class can help. Besides, LA can be seen as a consistency regularization technique. Enforcing the classifiers' predictions to be consistent across different perturbation is known to help in low-data regime (Bachman et al., 2014; Berthelot et al., 2019; Sohn et al., 2020). In fact, LA is a stronger alternative against data augmentation (DA), as we later show in §4.3 that LA outperforms DA by a large margin and can cover the role played by DA.

Why linear additive augmentation is meaningful. Deep representation space is shown to be highly linearized (Bengio et al., 2013; Upchurch et al., 2017), demonstrating the rationality behind *linearly* interpolating or extrapolating features, i.e. using "add" operation to generate new features.

Why legacy dictionary construction is warranted in both FSP and CLP. FSP uses classification task as a proxy task to learn useful encoders f_ϕ. During optimization, features are incentivized to maximize their dot-product similarity with class weights in f_ψ, thus forming meaningful metric space. In the regard of CLP, the contrastive loss, a form of metric-based loss, brings similar features closer while spreading dissimilar representations farther, which also results in an informative metric space. Thus, feature distance in the representation space of both FSP and CLP is meaningful, justifying the use of unsupervised clustering method to construct legacy dictionary.

4 EXPERIMENTS

4.1 SETUP

To better study low-shot classification problem for histological image, we use three public histology datasets to construct three tasks with different degrees of domain shift, elaborated as below.

Datasets. Since tissue type and appearance vary across body sites, we use three datasets from different body sites. They are: NCT-CRC-HE-100K dataset collected from colon site (Kather et al., 2018), LC25000 dataset collected from lung and colon sites (Borkowski et al., 2019), and PAIP19 dataset collected from liver site (Kim et al., 2021); we term them as NCT, LC-25K and PAIP respectively. NCT consists of 9 classes with 100k non-overlapping patches in total, each of size 224×224. LC-25K has 5 classes with 5,000 patches in each class; each patch is of size 768×768. PAIP is composed of 50 WSIs, each of size about 45k$\times$45k with 3 mask annotated classes. For LC-25K, all patches are resized to 224×224. For PAIP, large images are cropped into patches of size 224×224 with the same pixel resolution as NCT. Only patches containining foreground tissues are kept (75k patches) and their labels are assigned by majority voting. More details are in Appendix A.

We define three tasks to better reflect real-world scenarios:

i) **Near-domain task (to study GFSL).** NCT bears slight domain shift since samples are collected by various organizations with different staining intensity[3]. We randomly split NCT by 80%/20% to construct a training set (80k images) and a test set (20k images). Then, we do leave-one-class-out-as-novel-class to the training set to construct 9 base datasets and use the test set as $\mathcal{D}_{joint}$ for evaluation. This results in 9 sub-tasks, each of which has one class regarded as novel class and samples belong to it are excluded from pre-training datasets. See Appendix A.1 for more details.

ii) **Mixture-domain task (to study FSL).** LC-25K collects images from lung and colon sites. Two of five classes in LC-25K are colon-related, which suffers moderate domain shift due to difference in imaging protocols and pixel resolutions; we see them as middle-domain novel classes. The rest three lung-related classes are regarded as out-domain novel classes. We use the training set of NCT as $\mathcal{D}_{base}$, and LC-25K as $\mathcal{D}_{novel}$. See Appendix A.2 for more details.

iii) **Out-domain task (to study FSL).** Liver has different tissue organization and cell types compared with colon. Similar to mixture-domain task, we use the NCT training set as $\mathcal{D}_{base}$ and PAIP as target dataset $\mathcal{D}_{novel}$. To study *heterogeneous* and *homogeneous* shot selection problem, we use WSI ID information 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). During evaluation, support samples and query samples are drawn from the support WSI set and the query WSI set respectively. *Heterogeneous* strategy selects few-shot samples from different support WSIs, while *homogeneous* strategy selects them from a single randomly chosen support WSI. See Appendix A.3 for more details.

Evaluation. If not otherwise specified, for near-domain task, we evaluate methods in 1000×9 (9 sub-tasks) random meta-tasks; for mixture- and out-domain task, we evaluate methods in 1000 randomly sampled meta-tasks. All meta-tasks use 15 samples per class as the query set. We report the average F1-score and 95% confidence interval. To handle the unequal numbers of base classes and novel classes, we follow convention in GZSL (Xian et al., 2018) and GFSL (Shi et al., 2020) to report their average harmonic mean. More details about evaluation metrics are in Appendix A.

Implementations. **I. Pre-training.** We use ResNet-18 as the embedding function f_ϕ, and follow previous arts in FSL (Tian et al., 2020; Chen et al., 2019) to use l_2-normalized features for clustering and downstream meta-tasks. **II. Latent Augmentation.** We use faiss (Johnson et al., 2019) to perform K-means with a fix seed. The number of prototypes in legacy dictionary is 16 ($C = 16$, discussed in ablation §4.3). In each meta-task, each sample is augmented 100 times (including the original one) by LA. We defer more implementation details of pre-training and LA to Appendix C.

Compared methods. Recent works (Chen et al., 2019; Tian et al., 2020), including a concurrent work for histology image (Shakeri et al., 2021), show that sophisticated episodic training (meta-training) is no better than standard pre-training. Hence, we summarizes methods using standard pre-training as: 1) NearestCentroid. It computes class centroids from support sets and classifies query samples to their nearest centroids. Related works using such strategy includes Wang et al. (2019a), Snell et al. (2017), and Chen et al. (2020c), to list a few; 2) LinearClassifier. It trains a new

[3]See Figure A.1 in Appendix A.1 for a visual comparison.

Table 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. "LA" denotes latent augmentation. The bold numbers denote the best while the underscored numbers denote the second best. More 10-shot results are in Table D.1.

	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**

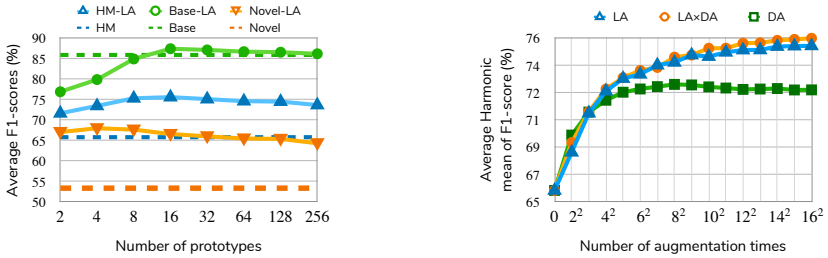
fully-connected layer with different loss functions (Chen et al., 2019; Lee et al., 2019) w.r.t. support samples, or directly uses linear models in scikit-learn (Pedregosa et al., 2011), e.g. LogisticRegression (Yang et al., 2021; Tian et al., 2020). For the ease of implementation and consistency, we use NearestCentroid, and two l_2-regularized linear classifiers — LogisticRegression and RidgeClassifier, all from scikit-learn APIs (Pedregosa et al., 2011).

4.2 Main Results

Main results in near-domain task and mixture-domain task are presented in Table 1. We discuss the overall findings here and leave the discussion of design choices to ablation study in §4.3, and the choice of linear classifiers (i.e. LogisticRegression/RidgeClassifier) to §4.4.

Fully-supervised f_ϕ^{FSP} v.s. Self-supervised f_ϕ^{CLP}. Results in Table 1 show that CLP generalizes fairly better to novel classes than FSP. Comparing the best vanilla entries (w/o. LA) using two types of pre-training methods, we observe an average advantage in HarmMean of 4%, 5% and 8% in 1-/5-/10-shot settings by CLP in near-domain task and 10%, 19%, 16% in mixture-domain task. Besides, CLP representations benefit more from the increase of number of shots than FSP's in both tasks, e.g. +17% vs. +11% and +12% vs. +10% when 1-shot $\rightarrow$ 5-shot for linear classifiers in near-domain, and mixture-domain tasks respectively. Despite the inevitable advantage of FSP in base classes under full supervision, CLP demonstrates stronger generalizability compared to FSP. Furthermore, Table 2 also confirms the superiority of CLP over FSP in out-domain task where large domain shift exists. Such generalization gap between FSP and CLP in pathology images is *at slight odds with* observations in natural images, where they show similar generalizability. We study and discuss it at §4.4. We also provide the linear evaluation results of all the 20 pre-trained models in Appendix C.3 for a reference to see how each model perform in NCT dataset.

Latent augmentation brings consistent improvement. Regardless of pre-training methods, LA brings consistent gains over baseline linear classifiers, confirming its effectiveness. With legacy dictionaries, limited few-shot samples are able to proliferate in a reasonable way by transferring latent variations. Such boost maintains its significance from near-domain task to mixture-domain task (Table 1) but becomes humble in out-domain task (Table 2). This is in our expectation since the three



(a) Effect of number of prototypes (b) Effect of number of augmentation times

Figure 3: **Ablations on latent augmentation.** We do 300 meta-tasks in 5-shot setting for this ablation. (a) The effect of different number of prototypes. Dash lines are the baselines for the solid lines with same colors. (b) The effect of number of augmentation times. The harmonic mean are plotted. "LA×DA" denotes T latent augmentations are applied after T traditional data augmentations (leads to T^2 times in total). **Best viewed by zoom-in with color.**

Table 2: **Results in out-domain tasks.** Average F1-scores from 1000 meta-tasks are reported.

Out-Domain Task						
RidgeClassifier	Homogeneous selection			Heterogeneous selection		
3-way K-shot	FSP	CLP	CLP+LA	FSP	CLP	CLP+LA
$K = 1$	36.90±1.06	42.56±1.20	**43.14±1.21**	/	/	/
$K = 5$	39.00±1.04	48.91±1.13	49.83±1.09	43.35±0.85	52.25±0.82	**53.67±0.81**
$K = 10$	40.26±1.03	50.57±1.11	51.62±1.08	45.91±0.78	55.96±0.74	**58.35±0.72**
$K = 50$	41.53±1.00	51.76±1.12	53.71±1.07	50.54±0.70	61.88±0.61	**65.38±0.61**
$K = 100$	41.23±0.98	52.74±1.07	54.25±1.03	52.45±0.66	64.03±0.60	**67.56±0.59**

classes (non-tumor, viable-tumor and other) defined in PAIP are extremely coarse-grained[4]: it may include couples of real fine-grained classes. Few samples are unable to well represent their entangled semantics. Hence, this observation does not repudiate the effectiveness of latent augmentation but re-ensures its tenability.

4.3 Ablations

To ablate design choices, we exclude two cancer-related classes, i.e. cancer-associated stroma (STR) and colorectal adenocarcinoma epithelium (TUM), from NCT to be novel classes, and use the rest as base classes. All ablations are conducted on CLP backbones with RidgeClassifier if not specified.

Number of prototypes in legacy dictionary. We do this ablation in 5-shot setting. Figure 3 (a) shows how the number of prototypes, i.e. C, affects recognition performance. We observe the similar tendency between base class and novel class, where their harmonic mean 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 exits 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 exposed in training. Nevertheless, LA demonstrates its robustness by consistent improvement over baselines (solid lines vs. dashed lines of same color in Figure 3-(a)).

DA vs. LA, and number of augmentation times. Consistency regularization helps in low-data regime as discussed in §3.2.1. Here we compare LA with data augmentation (DA), and their combination. Results are shown in Figure 3-(b). It can be seen 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 slightly polish 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 worthy to emphasize the computation budget involved in LA (addition in $\mathbb{R}^d$ space) is significantly lower than LA (image augmentation in $\mathbb{R}^{3HW}$ space and encoder forwards). Therefore, we run all experiments only with unsupervised LA, as one of our core contributions go with it, and its combination with DA is beyond the scope of this work.

Heterogeneous v.s. Homogeneous patch selection. In real clinics, WSIs' appearance varies across patients. Patches from a single WSI are confronted to its stain degree (impacts color) and

[4]Figure A.4 in Appendix shows visual examples to demonstrate the coarse granularity of classes in PAIP.

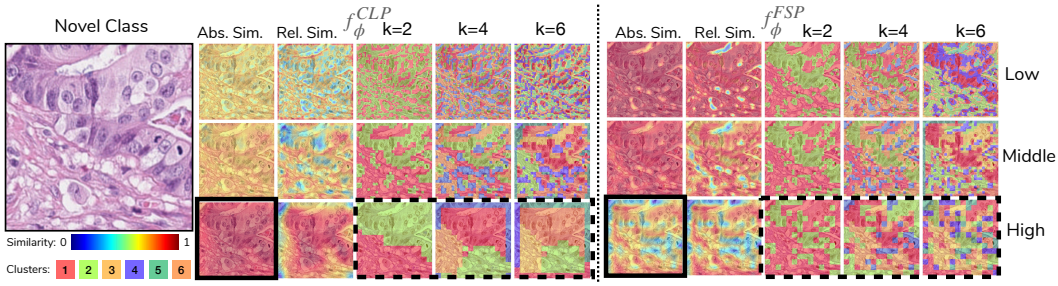


Figure 4: **Visualization of representations learned by CLP and FSP.** We visualize the absolute/relative cosine similarity between the global average pooled feature and the local features before pooling ("Abs./Rel. Sim." columns accordingly). To inspect correlation between local features, we conduct K-means clustering with different k values (columns with "k=*"). "Low", "middle" and "high" represent using features from stage-2, 3, and 4 from ResNet-18.Different colors indicate different clusters. Compared to FSP model, CLP model produces more global-local consistent high-level features (solid boxes) and can better differentiate tissue structures (dashed boxes). More examples are provided in Appendix D.3. **Best viewed by zoom-in with color.**

body site (impacts pattern and context). We therefore investigate the hetero-/homo-geneous patch selection strategies as defined in out-domain task (§4.1). Table 2 shows the results. We observe: i) heterogeneous selection is favored over homogeneous one, as it shows higher baselines; and ii) LA brings more gains for heterogeneous selection over baselines. Heterogeneous few-shot patches provide reliable and more diversified "anchor" samples than homogeneous ones, thus can benefit more from bootstrapping the legacy dictionary.

4.4 More discussion

Disparity between f_ϕ^{CLP} and f_ϕ^{FSP} influences the choice of base learner. In Table 1, we find i) stronger baselines for CLP and FSP vary, and ii) simple NearestCentroid can sometime outperform the vanilla l_2-regularized[5] linear classifiers for FSP. Here we briefly discuss our understandings. Representations produced by CLP can have different distributions compared to FSP, as also noticed by He et al. (2020). With limited training samples, different classifiers can have their own biases in building decision boundary, leading to different generalizability. Besides, in our implementation, no regularization techniques are used during FSP, e.g. weight decay (Krogh & Hertz, 1992), DropBlock (Lee et al., 2019; Tian et al., 2020) or "distill" regularization (Tian et al., 2020). The linear classifiers, though with l_2 penalty, may still be overfitted in such non-regularized representation space when only limited samples are provided. Subsequently, the simplest NearestCentroid model can yield better results than these overfitted linear models, as it has the least complexity.

Why do CLP models generalize better than FSP ones in pathology? To study why the large generalization gap exists, we follow Chen & Li (2020) to see how features are aggregated in space. Specifically, we run K-means on features from different layers (i.e. stage 2, 3, 4 of ResNet) with different cluster numbers (Figure 4). The CLP model extracts low-level features that are edge-related, and afterwards successively agglomerates adjacent similar structures (dashed boxes in Figure 4). In contrast, the FSP model can differentiate nuclei in low- and middle-levels but fails to encode tissue-structure-related features in deeper layer. These findings differ from those in ImageNet-like images (Chen & Li, 2020), where FSP and CLP show no difference across layers[6]. We also visualize the similarity between global features and local ones. The FSP model maintains global-local consistency, i.e. high similarity between global&local features, in low-/middle-levels, while the CLP model holds it in high-level (solid boxes in Figure 4). Unlike iconic natural images which only present a dominant object in centers, histology patches contain multiple small objects as well as texture-like tissues. The lack of dominant objectiveness makes FSP tend to overfit severely in high level, but CLP can capture the overall patterns from the crudely scattered tissues. We believe this is the main reason why CLP models generalize better than FSP models, *in pathology*. We provide detailed visualization procedures and more examples in Appendix D.3.

[5]RidgeClassifier and LogisticRegression are by default with an additional l_2 regularization term as loss.
[6]See Chen & Li (2020) or https://contrastive-learning.github.io/intriguing/ for a comparison.

5 RELATED WORK

Few-shot learning (FSL) and generalized few-shot learning (GFSL). **FSL** has been tackled from different perspectives, e.g. *metric-based* and *optimization-based* (Finn et al., 2017; Rusu et al., 2018; Munkhdalai & Yu, 2017). Here we focus on metric-based ones. They aim to learn task-dependent metric space by "meta-training", where base datasets are divided into many meta-tasks during training, and predictions are made by comparing distance between support samples and query samples (Gidaris & Komodakis, 2018; Snell et al., 2017; Vinyals et al., 2016). Later methods is attracted to *transfer learning*, i.e. *pre-training & fine-tuning*, which learns shared metric space by standard fully-supervised pre-training (Tian et al., 2020; Chen et al., 2019; Wang et al., 2018). **GFSL** is usually studied in the context of low-shot learning (Wang et al., 2018; Hariharan & Girshick, 2017), with adopted meta-learning few-shot methods (Gidaris & Komodakis, 2018; Shi et al., 2020). This paper follows "pre-training & fine-tuning" scheme for both FSL and GFSL, but we propose to incorporate self-supervised pre-training to enable label-efficient learning, and show that it can yield stronger generalization than supervised pre-training.

FSL in medical images. FSL in medical images is at its early stage, especially for histological images. Mahajan et al. (2020) investigate FSL methods in skin-disease classification, while Chen et al. (2021a) tackle COVID-19 CT image classification using contrastive pre-training and prototypical network fine-tuning. In the regard of histology image, Medela et al. (2019) use a triplet loss (Schroff et al., 2015) to pre-train an encoder with a followed fine-tuned SVM classifier for low-shot domain adaptation. Concurrent to our work, Shakeri et al. (2021) simultaneously propose a benchmark for few-shot classification of histological images. Our work has similar but different settings, with more investigations conducted, e.g. GFSL task, and hetero-/homo-geneous few shots selection.

Self-supervised learning. Self-supervised learning aims to learn good representations without golden labels. Recent state-of-the-art variants can be categorized as contrastive-based learning (Chen et al., 2021b; 2020a; He et al., 2020), cluster-based learning (Caron et al., 2018; 2020) and expectation prediction based learning (Grill et al., 2020; Chen & He, 2021). However, this line of works focus on pre-training on ImageNet-like images, and recent attention has been attracted to images with *multi-objects* and *multi-texture* presented (Chen & Li, 2020). We see histology image as a natural choice for such study, and show that contrastive learning can agglomerate structural "part-whole" information and maintain "global-local consistency", which make it generalize better for such data than supervised counterparts (see §4.4 and Table C.1 in Appendix C.3).

Representation variation augmentation. The idea of exploiting representation variations dates back to a decade ago (Heller et al., 2009; Salakhutdinov et al., 2012) or more earlier. Recent variants further develop this idea. For example, Hariharan & Girshick (2017) use a generator to generate "hallucinated" novel features from variation of given base samples. This method is later extended by not relying given base samples (Wang et al., 2018). Wang et al. (2019b) use class variance to perform semantic augmentation for classification and segmentation (Wang et al., 2021), while Yin et al. (2019) and Liu et al. (2020) utilize intra-class variance of head classes to augment tail classes for "long-tail" face recognition problem. Yang et al. (2021) use the distribution information, i.e. mean and variance, of base classes to calibrate novels' distribution. This work follows the line of Yang et al. (2021); Wang et al. (2021); Liu et al. (2020), but in contrast to them, we obtain and transfer variations without relying on any label information, allowing our method to scale gracefully to other label-hungry problems.

6 CONCLUSION

This work has studied, as an early attempt, the low-shot classification problem for histological images. We incorporate contrastive learning and latent augmentation to fully exploit training data in an unsupervised way, which means our method can gracefully scale to other large label-hungry problems. More importantly, we show that state-of-the-art contrastive learning pre-training method can generalize significantly better than supervised pre-training for histological images, which is in slight contrast to observations in iconic natural images. We analyze the underlying reasons and provide our empirical understandings. We hope our work could contribute to the study of representation learning and generalization for both self-supervised learning community as well as histology image community.

REFERENCES

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

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

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.

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. URL https://github.com/tampapath/lung_colon_image_set/.

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)*, pp. 132–149, 2018.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. 2020.

Ting Chen and Lala Li. Intriguing properties of contrastive losses. *arXiv preprint arXiv:2011.02803*, 2020.

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

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.

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, 2021a.

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

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

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

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

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

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

Spyros Gidaris and Nikos Komodakis. Dynamic few-shot visual learning without forgetting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4367–4375, 2018.

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.

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

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*, pp. 9729–9738, 2020.

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

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

Jakob Nikolas Kather, Niels Halama, and Alexander Marx. 100,000 histological images of human colorectal cancer and healthy tissue, April 2018. URL https://doi.org/10.5281/zenodo.1214456.

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.

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

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*, pp. 10657–10665, 2019.

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*, pp. 2970–2979, 2020.

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*, pp. 730–731, 2020.

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)*, pp. 1860–1864. IEEE, 2019.

Tsendsuren Munkhdalai and Hong Yu. Meta networks. In *International Conference on Machine Learning*, pp. 2554–2563. PMLR, 2017.

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

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.

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.

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.

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*, pp. 195–206. JMLR Workshop and Conference Proceedings, 2012.

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

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*, pp. 815–823, 2015.

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.

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.

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

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.

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

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*, pp. 266–282. Springer, 2020.

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*, pp. 7064–7073, 2017.

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. *Advances in neural information processing systems*, 29:3630–3638, 2016.

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*, 2019a.

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*, pp. 7278–7286, 2018.

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, 2019b.

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.

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*, pp. 3733–3742, 2018.

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.

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.

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*, pp. 5704–5713, 2019.

A DATASET

A.1 NCT

Data. NCT-CRC-HE-100K dataset (Kather et al., 2018) contains $100,000$ non-overlapping patches extracted from hematoxylin and eosin (H&E) stained human colorectal cancer and normal tissues. Each image is of size 224×224 at 0.5 MPP ($20\times$ magnification). Tissue classes and their label index are: 0) background (BACK), 1) adipose (ADI), 2) debris (DEB), 3) lymphocytes (LYM), 4) mucus (MUC), 5) smooth muscle (MUS), 6) normal colon mucosa (NORM), 7) cancer-associated stroma (STR), 8) colorectal adenocarcinoma epithelium (TUM). Figure A.2 shows the class distribution and Figure A.1 shows 4 examples per class from NCT. These images are from National Center for Tumor Diseases (Heidelberg, Germany) and University Medical Center Mannheim (Mannheim, Germany). Their acquisition protocols differ across organizations, which lead to inter-source domain shift.

Processing. We use the "NCT-CRC-HE-100K-NONORM" dataset, which does not apply color normalization to images. We randomly split NCT dataset (100k images) into a training set (80k images) and a test set (20k images). The class distribution is the same for the training set and test set, i.e. data are sampled w.r.t. each class. Since each image in NCT is of size 224×224, we do not further resize it.

Near-domain task. NCT dataset contains tissue patches from multiple sources, which bears slight domain shift since the staining intensity varies. To better study the generalizability in FSL problem, we do leave-one-class-out-as-novel-class to the training set of NCT. This procedure closely resembles the leave-one-out cross-validation. Since pathology datasets usually have far less number of classes compared to natural image dataset (e.g. 9 classes in NCT v.s. 100 classes in *mini*ImageNet and 608 classes in *tiered*ImageNet), we believe such leave-one-class-out-as-novel-class can better reflect the generalization error bound when number of classes is small. After the leave-one-class-out-as-novel-class split, 9 sub-base datasets are constructed, each of which has one class regarded as the novel class and samples belong to it are excluded from pre-training. For each sub-base dataset, a CLP model and a FSP model will be trained on it. Therefore, a total of 18 models will be obtained (9 CLP models and 9 FSP models). All models are evaluated on the NCT test set. It contains both seen classes and unseen novel classes, so as to simulate generalizes few-shot classification.

Evaluation metric. Typical FSL methods use accuracy as metrics. However, to study GFSL problem, the performance of novel classes and base classes should be appropriately considered. Accuracy w.r.t. specific classes cannot be computed in a joint label space. Hence, we choose F1-score as our metric. Besides, the numbers of novel classes and base classes are unequal in each sub-task, so we follow convention in GFSL to compute their harmonic mean. The final metrics are

$$M_{base}(i) = \frac{1}{8} \sum_{j \in \{0..9\} \setminus \{i\}} F_1(j), \tag{A.1}$$

$$M_{novel}(i) = F_1(i), \tag{A.2}$$

$$HMean(i) = \frac{2}{1/M_{base}(i) + 1/M_{novel}(i)}, \tag{A.3}$$

$$\text{Base} = \frac{1}{9}\sum_{i=1}^{9} M_{base}(i), \quad \text{Novel} = \frac{1}{9}\sum_{i=1}^{9} M_{novel}(i), \quad \text{HarmMean} = \frac{1}{9}\sum_{i=1}^{9} HMean(i), \tag{A.4}$$

where $F_1(j)$ denotes the F1-score of j-th class. $M_{base}(i)$ (Equation A.1) computes the average F1-score of base classes in i-th sub-task. Equation A.2 and Equation A.3 compute the F1-score of novel class and the harmonic mean of M_{base} and M_{novel} in i-th sub-task respectively. Equation A.4 shows the final metrics we report in this work.

A.2 LC25000

Data. The LC25000 (LC-25K) dataset contains 25,000 color images from 5 classes. They are: 0) colon adenocarcinoma (Colon_ACA), 1) benign colonic tissue (Colon_benign), 2) lung adenocarcinoma (Lung_ACA), 3) benign lung tissue (Lung_benign), and 4) lung squamous cell carcinoma

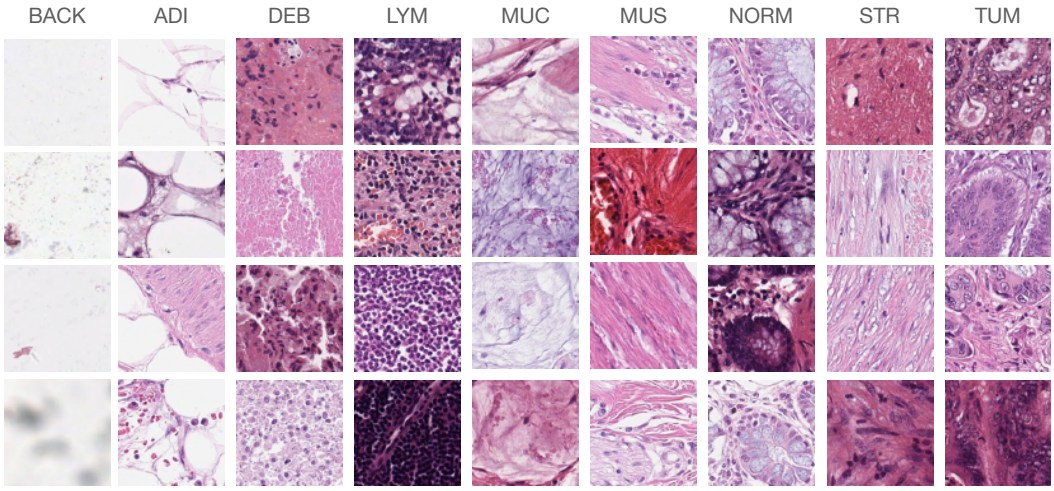


Figure A.1: Example images from NCT dataset. Each column represents a different class.

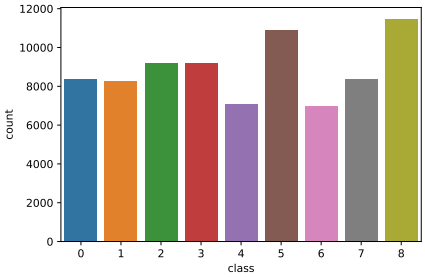


Figure A.2: Class distribution in NCT dataset. Labels denote: 0-ADI, 1-BACK, 2-DEB, 3-LYM, 4-MUC, 5-MUS, 6-NORM, 7-STR and 8-TUM. See Appendix A.1 for full names of labels.

(Lung_SCC). The class distribution is balanced, i.e. each class has 5000 images. All images are of size 768×768. LC25K is constructed by augmenting 1250 images (250 images for each class). The augmentations are: left and right rotations (within 25 degrees, p=1.0) and vertical and horizontal flips (p=0.5), where "p" represents the probability.

Processing. We resize all images from 768×768 to 224×224. No other process is taken.

Mixture-domain task. LC-25K dataset includes tissues from colon and lung sites. Two of five classes in LC-25K are colonic tissues, which suffers moderate domain shift due to difference in imaging protocol and pixel resolutions. Thus, we see them as middle-domain novel classes. The rest three lung-related classes are regarded as out-domain novel classes since they are from a different oragan. To study few-shot classification, we use the entire training set (80k images) of NCT as $\mathcal{D}_{base}$, and use LC-25K as $\mathcal{D}_{novel}$ for evaluation.

Evaluation metric. Similar to near-domain task evaluation in Appendix A.1, we compute $\text{Middle} = \frac{1}{2}(F_1(0) + F_1(1)), \text{Out} = \frac{1}{3}(F_1(2) + F_1(3)F_1(4)), \text{HarmMean} = \frac{2}{1/\text{Middle}+1/\text{Out}}$. Here $F_1(i)$ denotes the F1-score of i-th class. Class 0 and class 1 are colon-related, termed as middle-domain novel classes, and the rest three classes are lung-related, termed as out-domain novel classes.

A.3 PAIP19

Data. The PAIP 2019 (Kim et al., 2021) is composed of 50 H&E stained WSIs at $20\times$ magnification. Each WSI is approximately of size $45,000 \times 45,000$ with XML annotation. It contains three

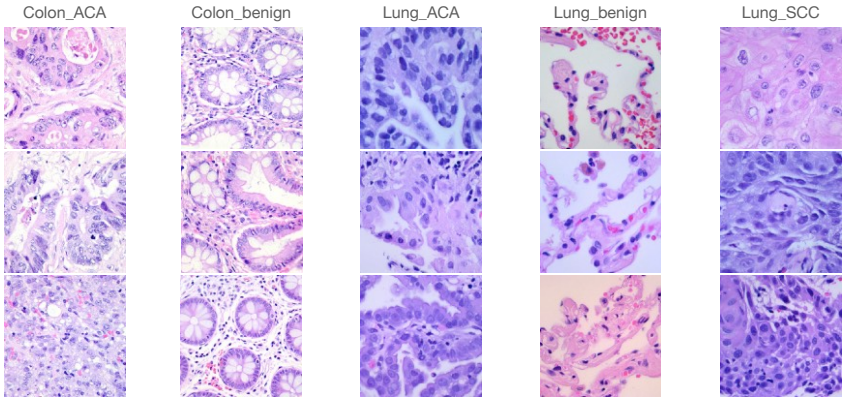


Figure A.3: Example images from LC25000 dataset. Each column represents a different class. The classes are: colon adenocarcinoma (Colon_ACA), benign colonic tissue (Colon_benign), lung adenocarcinoma (Lung_ACA), benign lung tissue (Lung_benign), and lung squamous cell carcinoma (Lung_SCC).

classes, which are: 0) non-tumor liver tissue, 1) viable tumor, and 2) other tissues in whole tumor area but not viable tumor. The class 2 may include intratumoral hemorrhage, necrosis or non-tumor tissue in whole tumor area. Figure A.4 shows an example WSI of PAIP19. This dataset provides "case number" (WSI identifier) information.

Processing. We use mask information to extract 500 randomly selected non-overlapping patches at $20\times$ magnification for each class from each WSI, leading to a dataset with $75,000$ samples. We define patches from 15 randomly selected WSIs as the support WSI set and patches from the rest 35 WSIs as query WSI set.

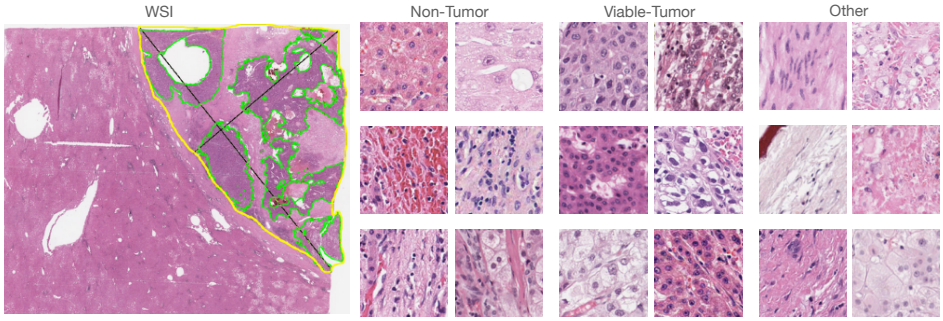


Figure A.4: Examples from PAIP 2019 dataset. The Leftmost image shows the overview of a whole-slide image, where tissues outside yellow boundary are defined as "non-tumor" class, and tissues inside green boundary are "viable-tumor" class, and tissues inside yellow but outside green are "other" class. Right columns show examples of extracted patches from PAIP19 dataset. The "non-tumor" class and "other" class are coarse-defined, since they can include many real fine-grained classes such as hepatocyte, vein, artery, stroma, fibrosis, fatty infiltration, etc.

B Contrastive Learning Pre-training: MoCo-v3

MoCo-v3 (Chen et al., 2021b) aims for an empirical study for vision transformer pre-training, but has a ResNet variant. We adopt the ResNet variant.

Overview. MoCo-v3 follows its ancestors MoCo v1/2 (He et al., 2020; Chen et al., 2020b) with straightforward modification. As shown in Figure B.1, each batch is augmented twice under stochastic augmentation t to obtain two views, denoted as $x_1, x_2 \sim t(x)$, where x represents the raw im-

ages. They are then encoded by query encoder and key encoder respectively. The query encoder is composed of three components: a backbone f_ϕ, a projector (projection head) f_g and a predictor (prediction head) f_q, formulated as $f_q \circ f_g \circ f_\phi$. The key encoder, also known as momentum encoder, consists of momentum copies of backbone and projector, formulated as $f_{\tilde{g}} \circ f_{\tilde{\phi}}$.

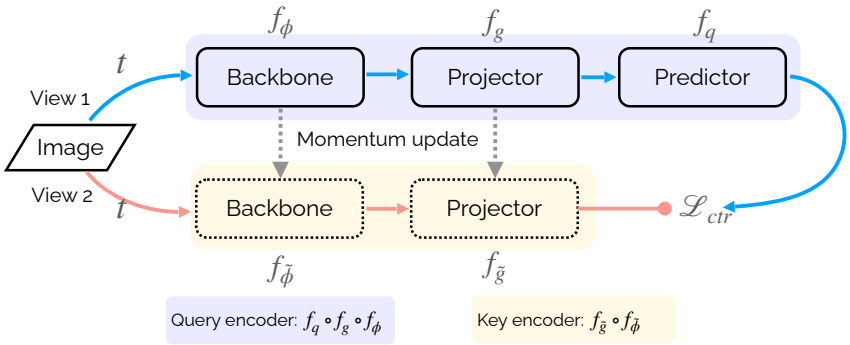


Figure B.1: Abstraction of MoCo-v3 structure. Each image is augmented twice under stochastic augmentation t to obtain two views (denoted by blue and red color). The key encoder (bottom, $f_{\tilde{g}} \circ f_{\tilde{\phi}}$ is momentum updated by parts query encoder (top, $f_q \circ f_g \circ f_\phi$). The $\tilde{}$ symbol denotes momentum updated parameters.

Momentum update. In iteration k, the momentum update rule is:

$$\tilde{\phi}_k \leftarrow m\tilde{\phi}_k + (1-m)\phi_k, \qquad \tilde{g}_k \leftarrow m\tilde{g}_k + (1-m)g_k, \tag{A.5}$$

where m is the momentum.

Loss function. The contrastive loss function is a form of InfoNCE (Oord et al., 2018):

$$\mathcal{L}_{ctr}(u, v) = -\log \frac{\exp\left(u \cdot v^+/\tau\right)}{\exp\left(u \cdot v^+/\tau\right) + \sum_{v^-} \exp\left(u \cdot v^-/\tau\right)}, \tag{A.6}$$

Here, v^+ denotes the positive sample, i.e. the other augmented view of u, and v^- denotes the negative samples, i.e. other representations in the batch. τ is a hyper-parameter, known as temperature (Wu et al., 2018). The final loss is a symmetric sum: $\mathcal{L}_{CLP} = \left[\mathcal{L}_{ctr}(z_1, \tilde{z}_2) + \mathcal{L}_{ctr}(z_2, \tilde{z}_1)\right]/2$, where the subscript indicates view source, $z = l_2\left[f_q \circ f_g \circ f_\phi(x)\right]$, $\tilde{z} = l_2\left[f_{\tilde{g}} \circ f_{\tilde{\phi}}(x)\right]$, and $l_2[\cdot]$ means l_2 normalization.

C Implementation Details

C.1 Fully-supervised Pre-training

Optimization. We use SGD optimizer with $lr = 0.5, momentum = 0.9$, and no weight decay is used, i.e. weight_decay=0. The batch size is 512. We train for 100 epochs with "step decay" learning schedule. The lr is multiplied by 0.1 at 30, 60 and 90 epochs respectively.

Data augmentation. We use `RandomResizedCrop`, `RandomHorizontalFlip`, followed by normalization (subtract `mean` and divide `std`) using ImageNet (Deng et al., 2009) statistics, i.e. $mean = (0.485, 0.456, 0.406)$ and $std = (0.229, 0.224, 0.225)$. During testing, we only resize the image to 224×224, followed by ImageNet normalization. Following previous work (Chen et al., 2019; Tian et al., 2020), we further l_2-normalize features.

C.2 Contrastive-learning Pre-training

The detailed description of MoCo-v3 is in Appendix B.

Table C.1: The linear evaluation accuracy of FSP and CLP models. FSP means fully-supervised pre-training, while CLP denotes contrastive-learning pre-training. We train LogisticRegression models on whole training features, and report the accuracy for whole test features. Base-Base means the novel class is excluded when we train and test linear classifiers. Joint-Joint means the linear classifiers are trained on all training features (80k samples) and then evaluated on all test features (20k samples). The "Average" row reports the average accuracy of 9 sub-tasks (the top 9 rows).

	Base-Base		Joint-Joint	
Novel Class	FSP	CLP	FSP	CLP
0	**98.46**	95.81	**97.73**	95.98
1	**97.80**	95.79	**97.41**	95.72
2	**98.59**	96.28	90.60	**95.10**
3	**96.67**	95.85	93.92	**95.88**
4	95.50	**96.49**	92.66	**95.67**
5	**97.60**	97.39	92.43	**94.33**
6	**98.72**	96.09	94.75	**95.22**
7	**98.49**	97.68	92.66	**95.52**
8	**97.44**	95.75	92.63	**94.80**
Average	**97.80**	96.35	93.87	**95.36**
No Novel Class	**98.43**	96.00	/	/

Architecture. The architecture of MoCo-v3 is `ResNet18-Projector-Predictor`. We use the codebase of OpenSelfSup[7]. For projector, we use 3-layer `NonLinearNeckSimCLR` (Chen et al., 2020a), with the dimension transitions of $512 \rightarrow 1024 \rightarrow 1024 \rightarrow 256$. For predictor, we use 2-layer `NonLinearNeckSimCLR`, with the dimension transitions of $256 \rightarrow 1024 \rightarrow 256$. The base momentum m to update key encoder is 0.996 and is linearly increased to 1 as training iteration goes. The temperature τ for contrastive loss is set to 1.

Optimization. We follow Chen et al. (2021b) to use `LARS` optimizer with initial learning rate of 0.3, weight decay of $1.5e-6$, $momentum = 0.9$, and use `CosineAnnearling` learning schedule. We train all models with batch size of 256 for 200 epochs.

Data augmentation. Following previous work (Grill et al., 2020; Chen et al., 2021b), we use strong data augmentation: `RandomResizedCrop`, `RandomHorizontalFlip`, and `ColorJitter` with (brightness=0.4, contrast=0.4, saturation=0.4, hue=0.4) and probability of 0.8, `RandomGrayscale` with probability of 0.2, `GaussianBlur` with probability of 0.5, and `Solarization` with probability of 0.2.

C.3 Linear evaluation for pre-training

In §4.1, we manually construct 9 sub-tasks in near-domain task, thus resulting 9 pre-trained models for each pre-training method (FSP and CLP). Besides, for mixture-domain and out-domain tasks, we use models pre-trained on the entire NCT training dataset with no class excluded.

Here we report their performance w.r.t. linear classifier. Specifically, we use the pre-trained models to extract features from NCT training set and test set. Then, we train LogisticRegression models on whole training features, and report the accuracy for whole test features. We report the results of Base-Base and Joint-Joint. Base-Base means the novel class is excluded when we train and test the linear classifiers, while Joint-Joint means the linear classifiers are trained on all training features (80k samples) and then evaluated on all test features (20k samples).

Table C.1 shows the results. With full supervision, FSP models can achieve better results in base label space (Base-Base). However, when evaluated in joint label space (Joint-Joint), they are worse than CLP models. Besides, in low-shot setting, CLP models underperform FSP models in base label space (*c.f.* the "Base" column in Table 1 near-domain task), but when the number of training samples

[7]https://github.com/open-mmlab/OpenSelfSup

Table D.1: Results of 10-shot setting in near-domain task and mixture-domain task. In near-domain task, the "Base" and "Novel" columns report the average F1-score of the base/novel class respectively; the "HarmMean" columns report their average harmonic mean. In mixture-domain task, the same metrics are reported w.r.t. middle-domain and out-domain novel classes. "LA" denotes latent augmentation. The bolded numbers denote the best while the underscored numbers denote the second best. Results in this table are in complementary to Table 1.

	10-shot Near-domain Task			10-shot Mixture-domain Task		
Methods	Base	Novel	HarmMean	Middle	Out	HarmMean
Fully-supervised pre-training (FSP)						
NearestCentroid	90.08±0.35	60.96±0.72	71.00±0.46	50.63±0.96	62.04±0.75	55.75±0.84
LogisticRegression	90.65±0.32	55.92±0.78	66.39±0.45	55.42±0.87	56.84±1.00	56.12±0.93
RidgeClassifier	90.14±0.34	49.62±0.86	60.33±0.48	62.72±0.78	58.93±1.01	60.77±0.88
LogisticRegression + LA (ours)	**92.75±0.29**	**69.10±0.63**	**78.14±0.39**	70.10±0.69	**65.26±0.78**	**67.60±0.73**
RidgeClassifier + LA (ours)	90.69±0.33	56.65±0.80	66.96±0.46	**71.09±0.66**	63.77±0.84	67.23±0.74
Contrastive-learning pre-training (CLP), ours						
NearestCentroid	85.15±0.45	68.18±0.72	75.18±0.54	85.97±0.40	68.58±0.65	76.30±0.49
LogisticRegression	87.20±0.41	69.83±0.71	76.83±0.51	85.82±0.38	66.42±0.66	74.89±0.48
RidgeClassifier	89.02±0.36	73.03±0.66	79.45±0.45	87.04±0.36	67.19±0.66	75.84±0.46
LogisticRegression + LA (ours)	**89.62±0.36**	82.40±0.48	85.48±0.40	90.44±0.34	80.08±0.51	84.95±0.41
RidgeClassifier + LA (ours)	89.45±0.36	**83.81±0.47**	**86.17±0.40**	**91.77±0.32**	**81.47±0.50**	**86.32±0.39**

Table D.2: Ablation on the effect of the number of prototypes in the legacy dictionary. This table shows the numerical results of Figure 3-(a).

	5-shot ablation task		
Number of prototypes	Base	Novel	HarmMean
No Latent Augmentation	85.85	53.27	65.74
2	76.84	66.99	71.58
4	79.83	**67.92**	73.39
8	84.87	67.59	75.25
16	**87.37**	66.51	**75.53**
32	87.08	65.94	75.05
64	86.65	65.41	74.55
128	86.52	65.32	74.44
256	86.13	64.25	73.60

increases (no longer few-shot setting), CLP models can achieve similar results as FSP models (Table C.1 "Base-Base" column).

D Detailed Results

D.1 10-shot results

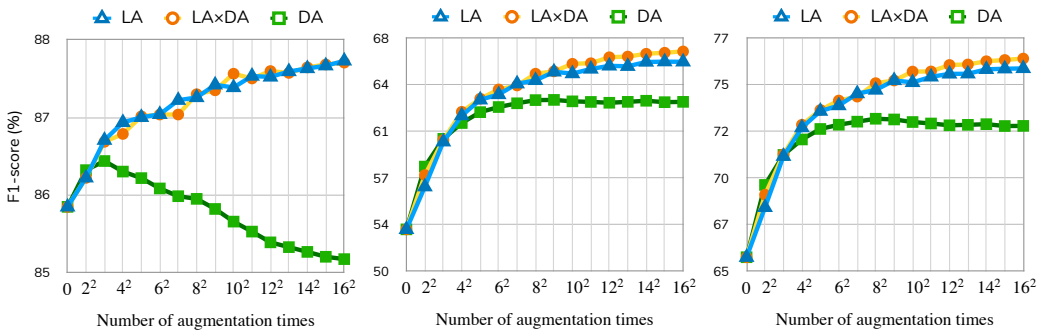
Table D.1 shows the results of 10-shot settings in near-domain and mixture-domain task, in complementary to results in Table 1.

D.2 Ablation

For ablation study, we exclude cancer-related classes, i.e. cancer-associated stroma (STR) and colorectal adenocarcinoma epithelium (TUM), from NCT training set, and use them as novel classes. An additional CLP model is trained on this base dataset for ablation. Table D.2 shows the numerical results of Figure 3-(a), which studies the effect of the number of prototypes in the legacy dictionary.

Figure D.1 shows the full results of comparison between data augmentation (DA), latent augmentation (LA), and their combination (LA$\times$ DA). It can be seen that, DA brings slight gain when the number of augmentation times is small, and brings negative gain when the number of augmentation times increases. In contrast, LA and LA$\times$DA can consistently improve the baselines for base class and novel class, resulting in best performance in harmonic mean.

The data augmentations used in this experiment are `RandomResizedCrop(scale=(0.8, 1.0))`, `RandomHorizontalFlip`, and `ColorJitter(brightness=0.4, contrast=0.4, saturation=0.4, hue=0.2)` with probability 0.8.



(a) Effect of # aug. times in Base Class (b) Effect of # aug. times in Novel Class (c) Effect of # aug. times of their HarmMean

Figure D.1: Comparison of DA, LA and LA×DA. "LA×DA" denotes T latent augmentation applied after T traditional data augmentation (leads to T^2 times in total).

D.3 Visualization

D.3.1 Visualization procedure

Models. All images are visualized with models that have never seen their classes during pre-training. For example, if the class "NORM" in NCT is regarded as novel class, samples belonging to it are excluded from this sub- pre-training dataset. Models trained on this sub-dataset are used to visualize "NORM" class.

Similarity map. To better understand why the generalization gap exists between CLP and FSP models, we visualize how CLP models and FSP models are attended to different local features for non-iconic, multi-object and multi-texture pathology images. To this end, we resize images in the NCT test set to 448×448, and use the pre-trained CLP models and FSP models to extract l_2-normalized features from stage-2, stage-3, and stage-4 of ResNet-18. To obtain similarity heat maps, we compute the cosine similarity between global representations (after pooling) and local representations (before pooling). For absolute similarity ("Abs. Sim."), we directly rescale the similarity values by multiplying 255. For normalized relative similarity ("Rel. Sim."), we first rescale the similarity values to $[0, 1]$ by subtracting the minimum and dividing the maximum, and then multiply them by 255.

Local feature agglomeration. We follow Chen & Li (2020) to see how local features are agglomerated across layers. To this end, we run K-means in scikit-learn Pedregosa et al. (2011) on the l_2 normalized local features (before pooling) from stage-2, stage-3, and stage-4 of ResNet-18 with different numbers of clusters, i.e. 2, 4, 6.

D.3.2 More examples

Figure D.2 and Figure D.3 show more example images in the NCT dataset. Figure D.4 and Figure D.5 show the results of the LC-25K dataset and the PAIP dataset respectively. We visualize the absolute and relative cosine similarity between the global average pooled feature and the local features before pooling ("Abs. Sim." and "Rel. Sim." columns accordingly). To inspect correlation between local features, we conduct k-means clustering with different k values (columns with "k=*"). "Low", "middle" and "high" represent using features from stage-2, 3, and 4 from ResNet-18.

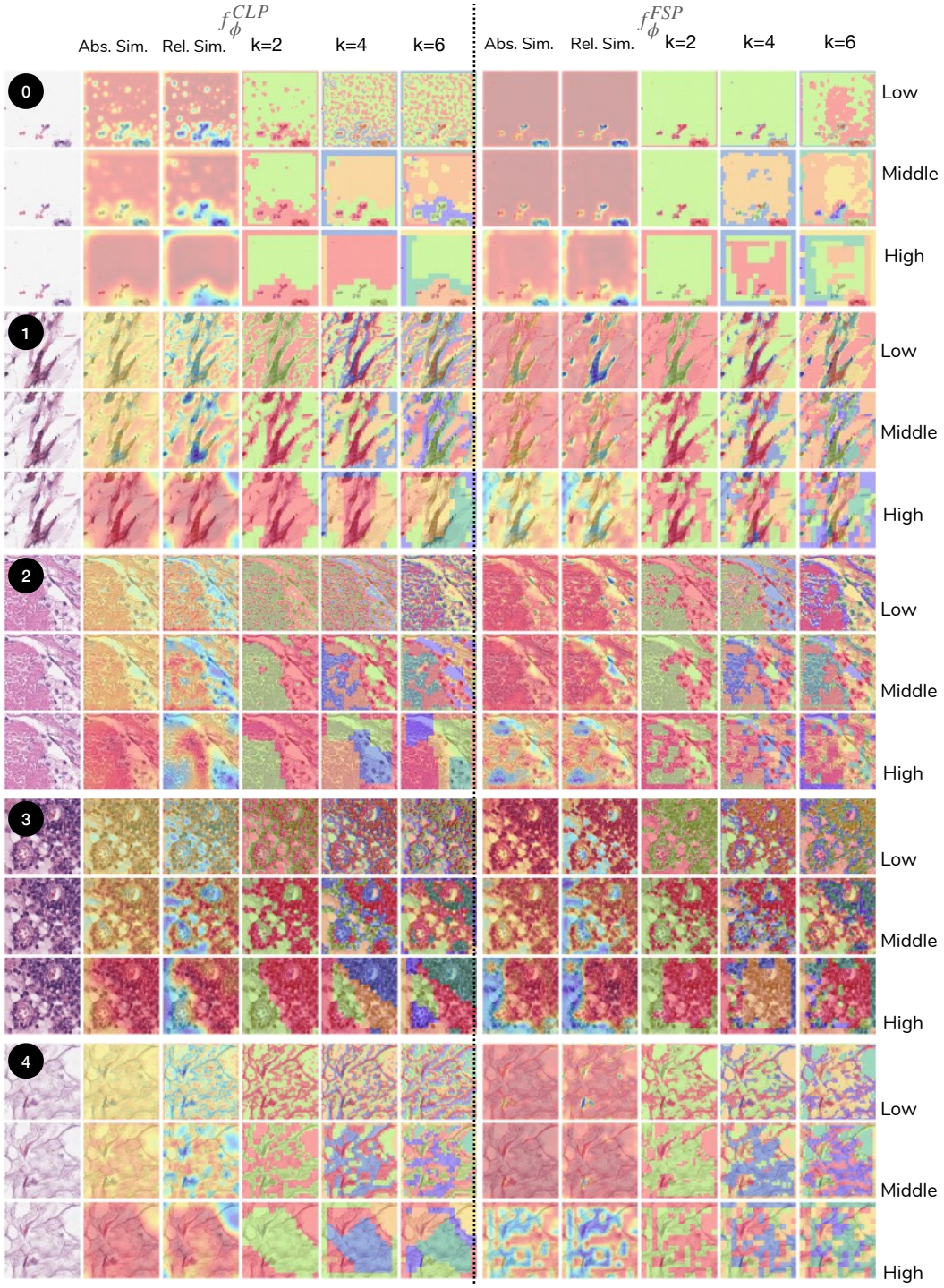


Figure D.2: **Visualization in the NCT dataset - 1.** White indexes in black circles represent class IDs. They are: 0) background (BACK), 1) adipose (ADI), 2) debris (DEB), 3) lymphocytes (LYM), and 4) mucus (MUC).

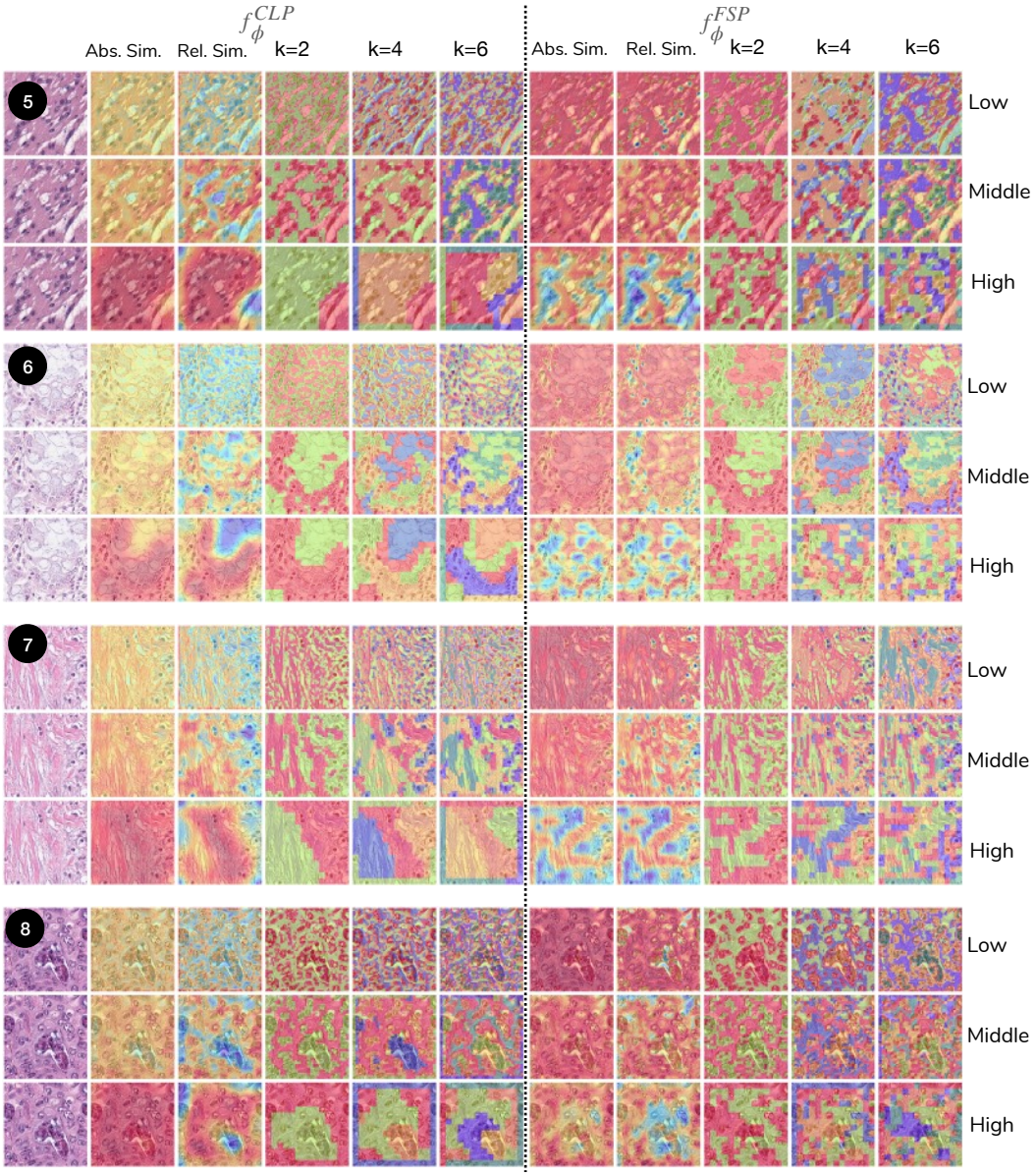


Figure D.3: **Visualization in the NCT dataset - 2.** White indexes in black circles represent class IDs. They are: 5) smooth muscle (MUS), 6) normal colon mucosa (NORM), 7) cancer-associated stroma (STR), 8) colorectal adenocarcinoma epithelium (TUM).

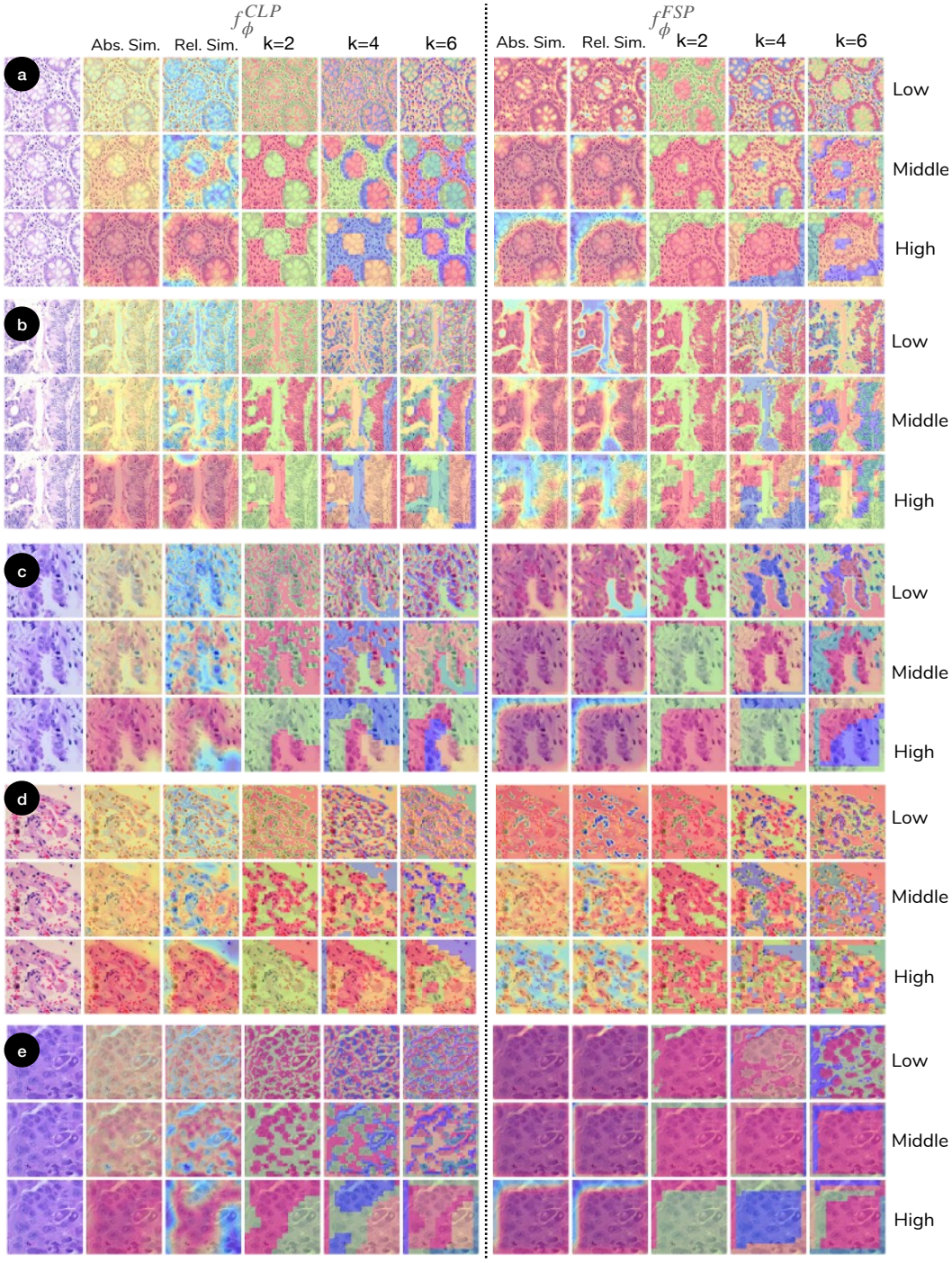


Figure D.4: **Visualization in the LC-25K dataset.** White indexes in black circles represent class IDs. They are: 0) colon adenocarcinoma, 1) benign colonic tissue, 2) lung adenocarcinoma, 3) benign lung tissue, and 4) lung squamous cell carcinoma.

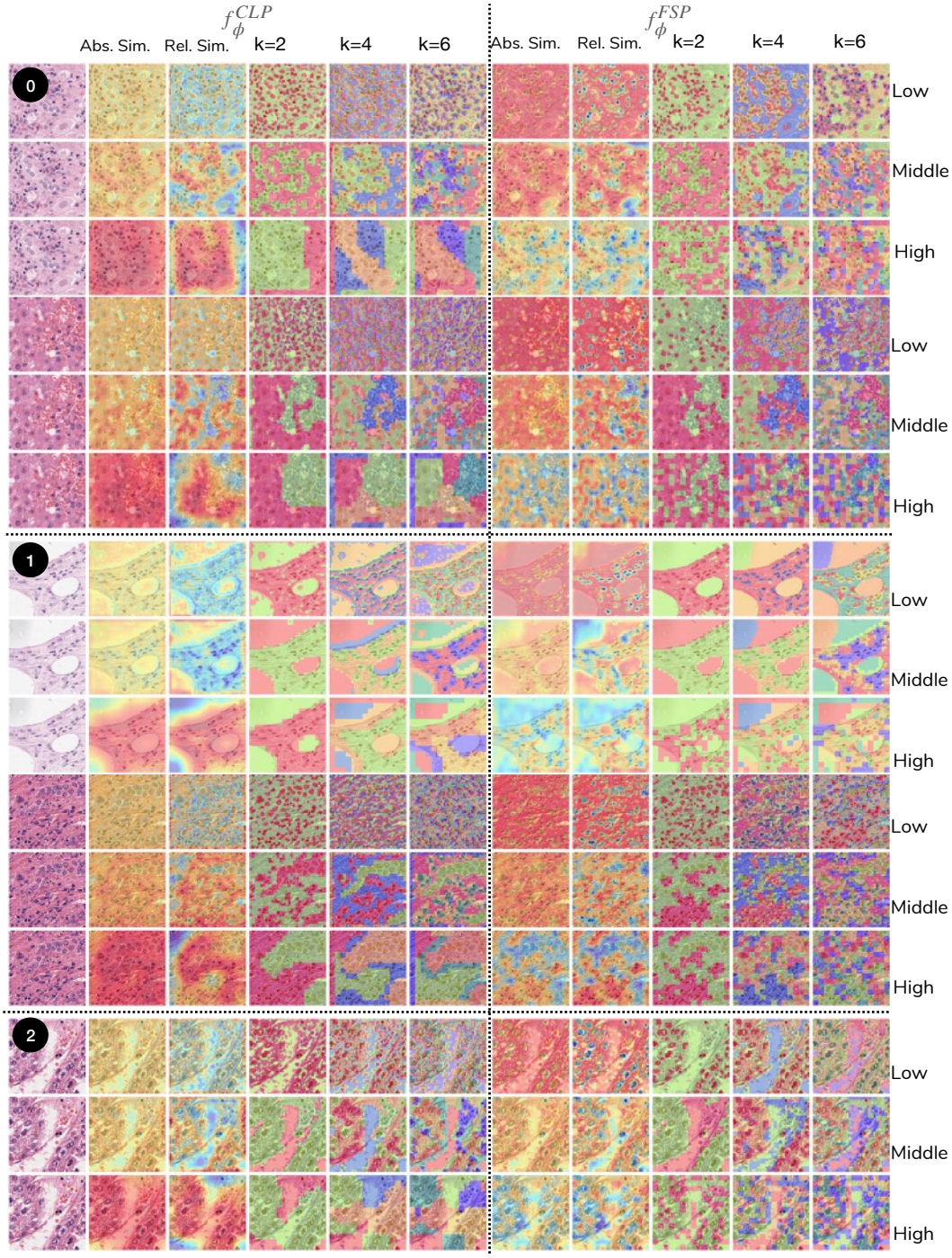


Figure D.5: **Visualization in the PAIP dataset.** White indexes in black circles represent class IDs. They are: 0) non-tumor, 1) viable-tumor and 2) other tissues.