Promptable Representation Distribution Learning and Data Augmentation for Gigapixel Histopathology WSI Analysis

Kunming Tang[1,2], Zhiguo Jiang[1,2,3], Jun Shi[4], Wei Wang[5,6], Haibo Wu[5,6], Yushan Zheng[1]*

[1]Beijing Advanced Innovation Center on Biomedical Engineering, School of Engineering Medicine, Beihang University
[2]Image Processing Center, School of Astronautics, Beihang University
[3]Tianmushan Laboratory
[4]School of Software, Hefei University of Technology
[5]Department of Pathology, the First Affiliated Hospital of USTC
[6]Intelligent Pathology Institute, Division of Life Sciences and Medicine, University of Science and Technology of China
{tkm2000,jiangzg,yszheng}@buaa.edu.cn, juns@hfut.edu.cn, weiwang@hmfl.ac.cn, wuhaibo@ustc.edu.cn

Abstract

Gigapixel image analysis, particularly for whole slide images (WSIs), often relies on multiple instance learning (MIL). Under the paradigm of MIL, patch image representations are extracted and then fixed during the training of the MIL classifiers for efficiency consideration. However, the invariance of representations makes it difficult to perform data augmentation for WSI-level model training, which significantly limits the performance of the downstream WSI analysis. The current data augmentation methods for gigapixel images either introduce additional computational costs or result in a loss of semantic information, which is hard to meet the requirements for efficiency and stability needed for WSI model training. In this paper, we propose a Promptable Representation Distribution Learning framework (PRDL) for both patch-level representation learning and WSI-level data augmentation. Meanwhile, we explore the use of prompts to guide data augmentation in feature space, which achieves promptable data augmentation for training robust WSI-level models. The experimental results have demonstrated that the proposed method stably outperforms state-of-the-art methods.

Code — https://github.com/lazytkm/PRDL

Introduction

Histopathology whole slide image (WSI) classification (Lu et al. 2021; Shao et al. 2021; Yang et al. 2022; Zheng et al. 2023) is increasingly popular in computer-aided pathological diagnosis, presenting unique challenges to the field of computer vision (Chen et al. 2022; Nakhli et al. 2023; Zhang et al. 2022b). Unlike conventional natural images, WSIs have massive image resolutions, often reaching up to billions of pixels. To tackle the gigapixel problem, a variety of methods within this domain employ multiple instance learning (MIL) frameworks (Ilse, Tomczak, and Welling 2018; Li, Li, and Eliceiri 2021; Campanella et al. 2019; Zhang et al. 2022a) to address the specific needs of WSI analysis.

Within the MIL framework, WSI analysis is usually divided into three phases: 1) Divide a WSI into patches; 2) Extract features for these patches; and 3) Aggregate these features to make a prediction for the entire WSI. Data augmentation serves as an effective expanding data strategy for training deep models (Chen et al. 2020a). It is also important in the domain of histopathology WSI analysis. As shown in Figure 1(a), if we draw a parallel to the process used for natural images, data augmentation for WSI should ideally be performed continuously after the WSI patching stage and before the patch-level feature extraction stage. However, for efficiency, the first two stages are generally conducted only once during the entire training process (Lu et al. 2021; Shao et al. 2021). This leads to the inapplicability of traditional image augmentation techniques, thereby inspiring the shift towards performing data augmentations directly in feature space.

To achieve data augmentation for WSI, several studies have proposed using generative methods (Dai et al. 2024; Zaffar et al. 2022) or Mixup techniques (Yang et al. 2022; Chen and Lu 2023; Gadermayr et al. 2023) to create data augmentations in feature space, as illustrated in Figure 1(b). However, these methods either introduce additional computational costs due to the need for training another parameterized model or cause a loss of semantic information. Additionally, the augmented results generated by these methods often lack control, in contrast to data augmentations applied directly in image space where changes are more visually intuitive and easier to manage. Currently, there lacks data augmentation methods to meet the unique demands of gigapixel image analysis that are not only computationally efficient but also preserve control.

In this paper, we introduce a novel approach named promptable representation distribution learning (PRDL) to address the challenges associated with patch-level representation learning and WSI-level data augmentation in histopathological WSI analysis. Within this framework, a representation distribution estimator is designed and trained during self-supervised representation learning. As shown in Figure 1(c), this estimator is capable of predicting a distribution of potential representation augmentations for each patch. After patch-level feature extraction, each patch within a WSI is represented as an individual distribution rather than

*Corresponding author

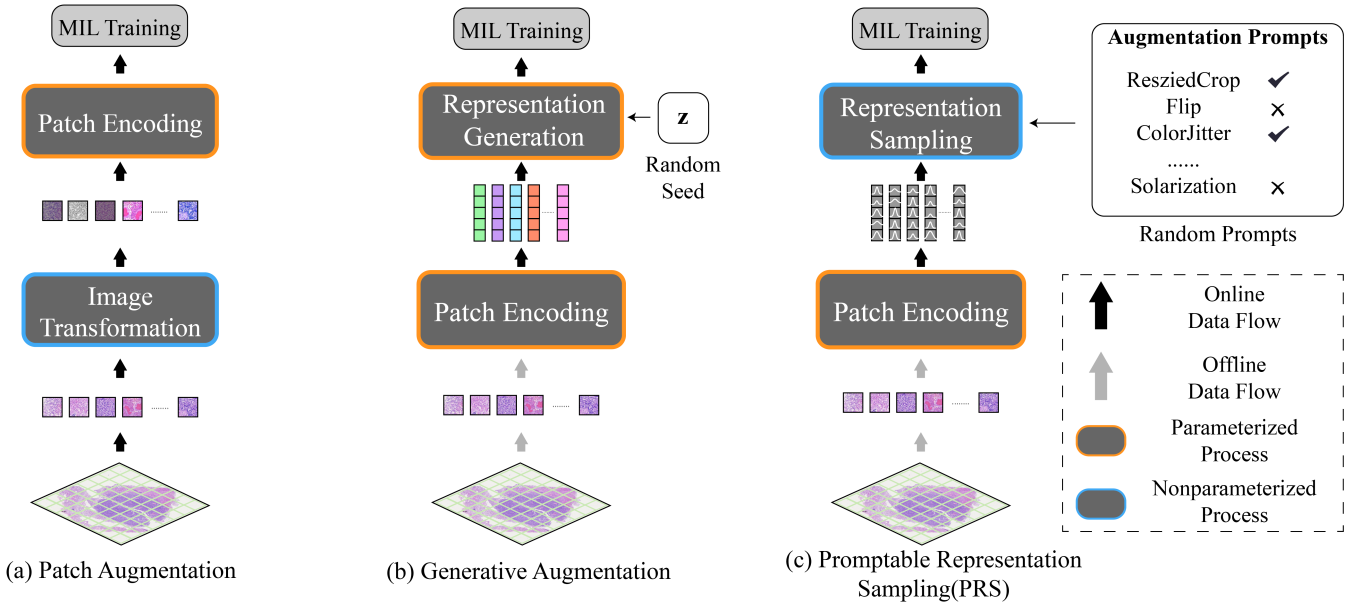


Figure 1: Comparison between existing methods for WSI data augmentation and our method. (a) represents the traditional image augmentation used in natural images, which is inefficient. (b) involves the use of generative models for data augmentation in feature space. (c) describes our promptable representation sampling strategy tailed for WSI augmentation.

a static point in feature space. Additionally, the representation distribution of each dimension is confined within a range by specific augmentation prompts, to simulate different augmentation operations that are used in image space. Finally, we implement a non-parameterized representation augmentation process through online sampling patch representations from these prompted distributions, to efficiently achieve the data augmentation for WSI-level model training. The proposed method was evaluated on a lung dataset with 754 WSIs and two public lung datasets with 696 WSIs and 3064 WSIs. The experimental results have demonstrated that the proposed method stably outperforms state-of-the-art methods. The main contributions of this paper can be summarized as follows:

(1) We proposed promptable representation distribution learning (PRDL), a novel representation learning framework with prompted representation distribution estimation for WSI classification. A promptable distribution estimator is designed to incorporate representation augmentation into representation learning. Compared with the traditional image-level data augmentation, the proposed method can provide more expansive augmentation. This significantly improves the discrimination of the patch representations and thereby enhances the performance of the subsequent WSI analysis model.

(2) We designed a promptable representation sampling (PRS) module based on PRDL. Utilizing PRS, we successfully facilitated the interchange between data augmentation and patch encoding processes, and achieved promptable and flexible data augmentation in the feature space for gigapixel histopathology image analysis. Furthermore, we leveraged the augmentation prompts in image space to guide the training of the learnable augmentation masks in feature space. This strategy enables us to conduct representation augmentation with greater control, enhancing the flexibility of the augmentation process.

Related Work

Data Augmentation for WSI analysis

The conventional method of data augmentation for WSI, similar to that used for natural images, involves continuously extracting representations of augmented patches from each "bag" (a set of patches) throughout the training process. However, this is obviously inefficient for WSI model training due to the huge amount of time required for feature extraction. Consequently, augmentation is primarily performed during WSI preprocessing, as seen in methods like AB-MIL (Ilse, Tomczak, and Welling 2018), to enhance patch diversity before training.

Data augmentation strategies developed for WSI analysis fall into two main categories. The former is realized with generative models (Zaffar et al. 2022; Dai et al. 2024), while the latter generates new subsets from bags through various mixing approaches (Yang et al. 2022; Chen and Lu 2023; Gadermayr et al. 2023). Specifically, DAGAN (Zaffar et al. 2022) involves training networks like generative adversarial networks to create synthetic data augmentations within the feature space. However, these generative models require an extra training phase separate from the representation model, as well as additional computational resources during the inference stage. ReMix (Yang et al. 2022) mixes class-specific prototypes determined by K-means clustering. Although Mixup methods can enhance data diversity, they may sometimes generate representations that deviate from the distribution of real-world data, which potentially com-

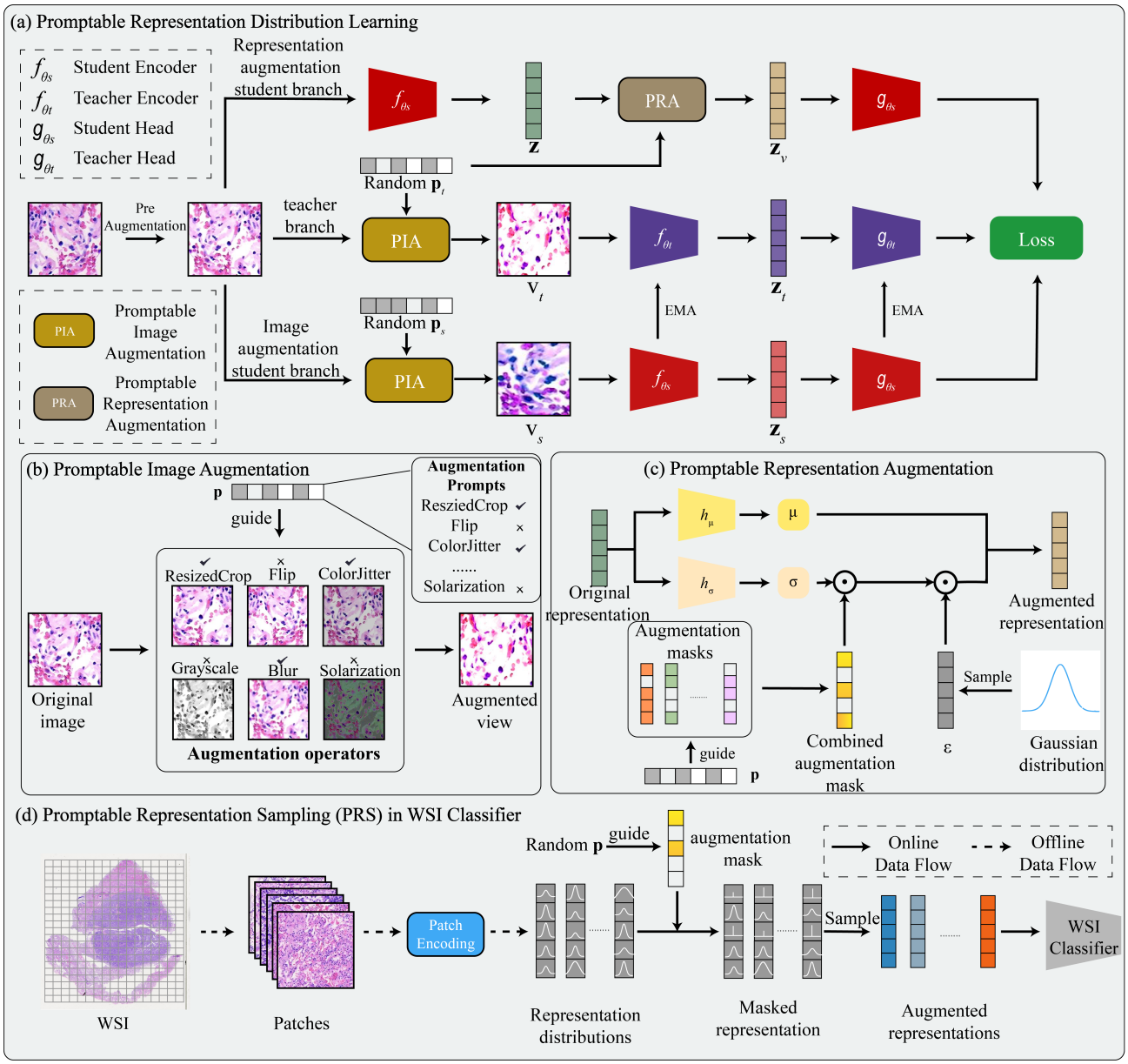


Figure 2: The proposed representation learning and WSI data augmentation framework includes (a) the process of PRDL, where the two student branches share weights in the encoder and head, (b) and (c) provide detailed descriptions of the modules in (a), and (d) shows the flowchart of WSI augmentation during training.

promises the model's performance on actual datasets. The shared shortcoming of these data augmentation methods is their inability to direct the augmentation process, leading to a lack of control.

Self-Supervised Representation Learning

In MIL methods, it is essential to select an appropriate patch encoder. CNN models pre-trained on the ImageNet (Russakovsky et al. 2015) often serve as patch encoders due to their robust feature extraction capabilities (Lu et al. 2021; Shao et al. 2021). However, there are inevitably semantic differences between pathological and natural images. Self-supervised learning (SSL) has been widely used in representation learning (Jaiswal et al. 2020; Bao et al. 2022; Huang et al. 2023; Assran et al. 2023; Zhou et al. 2022). Chen et al. (Chen et al. 2020a) propose an end-to-end model SimCLR and systematically study the impact of data augmentations. They observe that single augmentation is not sufficient to learn good representations. MoCov2 (Chen et al. 2020b) expands upon MoCo (He et al. 2020) by incorporating blur augmentation, thereby improving the baseline on ImageNet. Grill et al. (Grill et al. 2020) develop BYOL,

a unique metric-learning approach that learns representations by predicting one view from another, demonstrating the importance of color diversity in augmentations. Most recently, DINO (Caron et al. 2021) is proposed to utilize self-supervised ViT (Dosovitskiy et al. 2021) for representation learning. DINO (Caron et al. 2021) follows the data augmentations of BYOL (Grill et al. 2020) and multi-crop (Caron et al. 2020), which have also proven the advantage for patch representation pretraining in histopathology WSI analysis (Chen et al. 2022; Wu et al. 2024). Our approach leverages the DINO framework and ViT architecture, setting this combination as the baseline for our representation learning. Unlike previous methods that focus on image space, we investigate the potential of applying data augmentation in feature space to enhance representation learning.

Method

Our goal is to develop a data augmentation strategy that operates within the feature space after the encoding of image patches, rather than applying traditional image augmentation techniques before patch encoding. Additionally, it is crucial that the representation augmentation remains as promptable as traditional image augmentations, avoiding that the outcomes are unpredictable and harm the performance.

As illustrated in Figure 2a, the architecture of our proposed model is constructed on the foundation of DINO (Caron et al. 2021). We extend the model by introducing an additional student branch specifically for representation augmentation. This branch shares weights with the image augmentation student branch and incorporates a promptable representation augmentation module. Within this framework, we can employ prompts aligned with image augmentation to guide the augmentation in feature space.

Following the training of the model, as depicted in Figure 2d, we can continuously obtain diverse WSI data in feature space through online sampling representations from the distributions. These distributions can be adjusted by different combinations of the trained augmentation masks. In this way, we can perform data augmentations in feature space that are as promptable as image augmentations.

Promptable Representation Distribution Learning

In self-supervised learning, the typical approach involves generating two different sets of augmented data from the same original data to help the model learn useful features without labeled data. In this approach, we modify traditional image augmentation by integrating a set of prompts that specify the type of augmentation. Given an image, we obtain $\tilde{\mathbf{x}} \in \mathbb{R}^{w \times h \times 3}$ by pre-augmentation, including flip, color distortion and gray scaling. Assuming that we have an augmentation set $\mathcal{T}$ consisting of K augmentation operators $o_1, \ldots, o_K$, we produce two augmentation prompts $\mathbf{p}_t \in \{0,1\}^K$ and $\mathbf{p}_s \in \{0,1\}^K$ by sampling random compositions of augmentation operators, where a value of 1 appears in the i-th bin of the prompt indicates the i-th augmentation operator o_k is active.

Promptable Image Augmentation Image augmentation is the basis of traditional self-distillation. We first define the augmentations required in our image augmentation student branch and teacher branch. Guided by the augmentation prompts $\mathbf{p}_t$ and $\mathbf{p}_s$, we produce two different views $\mathbf{v}_t = t(\tilde{\mathbf{x}}|\mathbf{p}_t)$ and $\mathbf{v}_s = t(\tilde{\mathbf{x}}|\mathbf{p}_s)$. The views $\mathbf{v}_t$ and $\mathbf{v}_s$ are further encoded by f_{θ_t} and f_{θ_s} to obtain their representations $\mathbf{z}_t = f_{\theta_t}(\mathbf{v}_t)$ and $\mathbf{z}_s = f_{\theta_s}(\mathbf{v}_s)$. Subsequently, these representations are transformed into embeddings through projector heads g_{θ_t} and g_{θ_s}. Referring to the knowledge distillation paradigm, we train the student network to match the output of the given teacher network, parameterized by θ_t and θ_s, respectively.

Representation Distribution Estimation The core of the representation augmentation student branch is representation distribution estimation. In our method, we construct a neural network that can be trained to act as an estimator, using Gaussian prior to characterizing the distribution of patch representations (Zang, Huang, and Loy 2021). This architecture includes two main components: a mean head h_μ and a variance head h_σ, both composed of fully connected layers. Given the inherent non-negativity of the variance, we compute its logarithm rather than the variance itself directly. The corresponding representation mean $\boldsymbol{\mu} \in \mathbb{R}^D$ and standard deviation $\boldsymbol{\sigma} \in \mathbb{R}^D$ of each image that together define a Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma}^2)$, which can be calculated by equations:

$$\boldsymbol{\mu} = h_\mu(f_{\theta_s}(\tilde{\mathbf{x}})), \quad \boldsymbol{\sigma} = \exp\left(h_\sigma(f_{\theta_s}(\tilde{\mathbf{x}}))/2\right). \tag{1}$$

Promptable Representation Augmentation The estimated distribution $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma}^2)$ handles the potential outcomes of representation augmentations for the patch. To make the distribution reflect representations from specific augmentation operators, we introduce a set of masks K denoted by $\mathbf{M} = [\mathbf{m}_1, \mathbf{m}_2, ..., \mathbf{m}_K]^{\mathrm{T}}$, where each row $\mathbf{m}_k$ corresponds to a specific augmentation operator o_k from the set of transformations $\mathcal{T}$. To ensure $\mathbf{M} \in (0,1)^{K \times D}$, we adopt a strategy where $\mathbf{M}$ is obtained by applying the sigmoid function to another random initialized matrix $\mathbf{U} \in \mathbb{R}^{K \times D}$. By incorporating the augmentation prompt $\mathbf{p}_t$, we constrain this distribution to a more specific augmentation space. Preciously,

$$\boldsymbol{\sigma}_{p_t} = \boldsymbol{\sigma} \odot \mathbf{m}_{p_t}, \quad \mathbf{m}_{p_t} = \mathbf{p}_t\mathbf{M}/\|\mathbf{p}_t\|_1, \tag{2}$$

where $\odot$ represents Hadamard product. With the narrowed distribution $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma}_{p_t}^2)$ for a patch, we can obtain variable representations of the patch by sampling process $\mathbf{z}_v \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma}_{p_t}^2)$, which is computationally efficient. Furthermore, by configuring $\mathbf{p}_t$, we can control the specific type of augmentation to be executed. Here, we adopt the reparameterization trick (Doersch 2021) to enable backpropagation for training. Specifically, we sample representations $\mathbf{z}_v$ under the representation independence assumption

$$\mathbf{z}_v = \boldsymbol{\mu} + \boldsymbol{\sigma}_{p_t} \odot \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I}_D), \tag{3}$$

where $\mathbf{I}_D$ denotes D-dimensional identity matrix.

Objective and Optimization

Knowledge Distillation Then, we adapt the augmentation paradigm to self-supervised learning. Following the structure of DINO (Caron et al. 2021), we obtain the probability

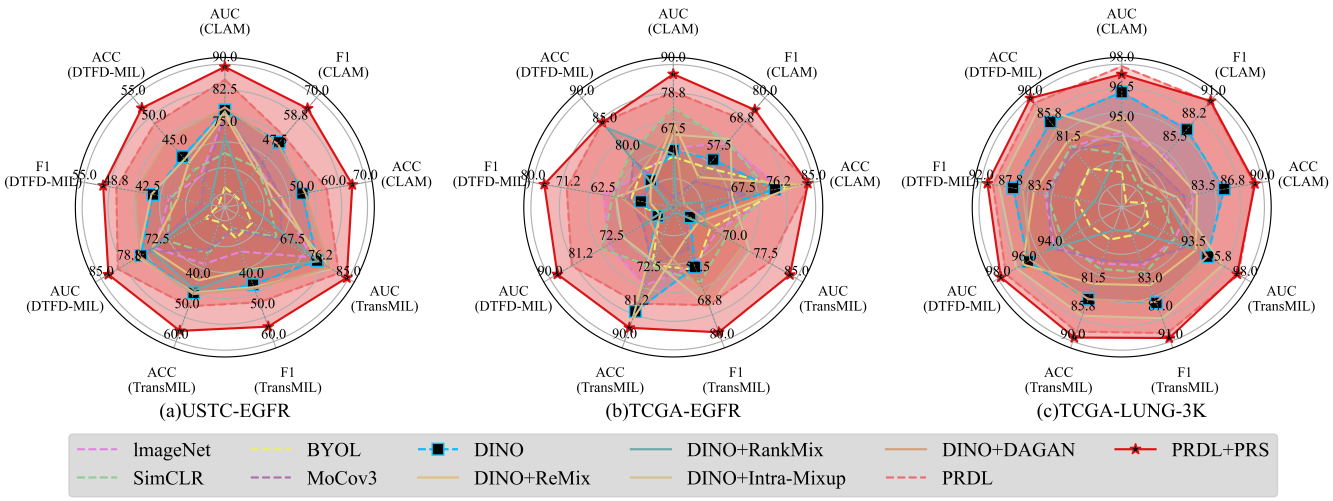


Figure 3: Comparisons with SOTA Methods. Please refer to the supplemental material in the extended version for complete numerical results.

distribution over D dimensions denoted by $\mathbf{y}_t$ by a softmax operation on the branch outputs:

$$\mathbf{y}_t = softmax(g_{\theta_t}(\mathbf{z}_t)/\tau_t), \tag{4}$$

where $\tau_t > 0$ is a temperature parameter that controls the sharpness of the probability distribution. Similar formulas hold for $\mathbf{y}_s$ and $\mathbf{y}_v$ from the two student branches with the mapping head g_{θ_s}. We minimize the basic loss function

$$\mathcal{L}_{CE} = H(\mathbf{y}_t, \mathbf{y}_s) + H(\mathbf{y}_t, \mathbf{y}_v), \tag{5}$$

where $H(\mathbf{a}, \mathbf{b}) = -\mathbf{a} \log \mathbf{b}$. It is important to note that we assign the same random prompt $\mathbf{p}_t$ for a patch to obtain $\mathbf{z}_t$ and $\mathbf{z}_v$. It guides the distillation architecture to align representations from both the image augmentation and representation augmentation for the same combination of augmentation operators. This is the basis on which we can decouple the augmentation operators and thereby control the process of the representation augmentation. Moreover, we follow the DINO (Caron et al. 2021) framework to adopt the multi-crop strategy by using 2 global views and several local views. All crops are fed into the image augmentation student branch while only the global views are fed into the teacher branch.

Representation Distribution Constraint As we model the representation distribution based on Gaussian prior, We add a Kullback-Leibler (KL) divergence constraint to the distribution estimator, which is represented as

$$\mathcal{L}_{KL} = D_{KL}(\mathcal{N}(0, \mathbf{I}_D) | \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma}^2)), \tag{6}$$

where $\mathcal{N}$ is the Gaussian distribution and D_{KL} is the K-L divergence.

Promptable Mask Constraint The prompted masks aim to identify the most specific feature dimensions associated with different augmentations. Therefore, we employ L1 normalization to induce sparsity in the augmentation masks:

$$\mathcal{L}_{sp} = \|\mathbf{m}_{p_t}\|_1, \tag{7}$$

Additionally, we introduce a variance regularization term on the standard deviation of the embeddings across the feature dimension (Bardes, Ponce, and LeCun 2022), to mitigate the issue of augmentation masks trending towards zero:

$$\mathcal{L}_{var} = max(0, 1 - \sqrt{Var(\mathbf{m}_{p_t}) + \gamma}), \tag{8}$$

where γ is a small scalar preventing numerical instabilities.

Overall Objective The final object function is composed as follows

$$\mathcal{L}_{total} = \mathcal{L}_{CE} + \beta_1 \mathcal{L}_{KL} + \beta_2 \mathcal{L}_{sp} + \beta_3 \mathcal{L}_{var}, \tag{9}$$

where β_1, β_2 and β_3 controls the weights of each term in the loss. The parameters of the student network and the distribution estimator are optimized by the gradient descent algorithm, and the teacher network is updated by the exponential moving average (EMA) mechanism

$$\theta_t \leftarrow \lambda \theta_t + (1 - \lambda)\theta_s \tag{10}$$

with the update rule of λ following a cosine schedule during training.

WSI Analysis With Representation Augmentation

Given a WSI $\mathbf{X} \in \mathbb{R}^{W \times H \times 3}$ sized by $W \times H$, consisting of patches $(\mathbf{x}_1, \mathbf{x}_2, \ldots \mathbf{x}_n)$, a common MIL model for WSI classification can be formulated as:

$$\hat{Y} = \psi_{\theta_2}(f_{\theta_1}(\mathbf{x}_1), f_{\theta_1}(\mathbf{x}_2), \ldots, f_{\theta_1}(\mathbf{x}_n)) \tag{11}$$

where $\hat{Y}$ is the WSI-level prediction, f_{θ_1} is the patch encoder and ψ_{θ_2} is the WSI-level representation aggregator.

Promptable Representation Sampling Through knowledge distillation in the promptable representation distribution learning, we achieved the goal of procedure exchange of data augmentation and patch encoding. Then we proposed a representation augmentation strategy named promptable

Methods	CLAM (Lu et al. 2021)			TransMIL (Shao et al. 2021)			DTFD-MIL (Zhang et al. 2022a)		
	AUC	F1-Score	ACC	AUC	F1-Score	ACC	AUC	F1-Score	ACC
DINO+Random Perturbation	74.9	35.0	40.8	74.1	39.2	39.9	75.6	42.8	43.9
DINO+MC Sampling	75.9	37.2	39.5	75.2	36.7	38.6	81.3	45.0	47.1
PRDL w/o PRA (DINO)	79.4	49.8	50.2	75.2	41.3	44.0	76.1	41.5	43.1
PRDL w/o $\mathcal{L}_{KL}$	83.8	51.2	51.6	81.0	43.1	46.6	80.0	39.3	41.7
PRDL w/o $\mathcal{L}_{sp}$	83.8	52.7	54.3	78.4	43.5	45.3	79.6	41.7	44.0
PRDL w/o $\mathcal{L}_{var}$	84.6	49.9	51.6	80.8	43.1	49.8	80.9	39.6	47.3
PRDL	86.4	52.5	57.9	82.0	47.1	48.4	81.3	48.7	49.8
PRDL+PRS	**89.4**	**65.1**	**65.9**	**84.5**	**55.1**	**56.5**	**83.4**	**51.3**	**52.9**

Table 1: Ablation study of the proposed framework on the USTC-EGFR Dataset.

representation sampling (PRS) for WSI augmentation. As illustrated in Figure 2d, patch-level representation distributions are obtained after patch encoding. During WSI classifier training, patch representations $\mathbf{Z} = (\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_n)$ within $\mathbf{X}$ can be replaced by representations online sampled from the corresponding distributions with random combined augmentation prompts, which can be formulated as

$$\mathbf{Z}_v = (\mathbf{z}_{v_1}, \mathbf{z}_{v_2}, \ldots, \mathbf{z}_{v_n}), \quad \mathbf{z}_{v_i} \sim \mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\sigma}_i^2). \tag{12}$$

Then, we feed $\mathbf{Z}_v$ instead of $\mathbf{Z}$ into the WSI model to achieve data augmentation for training, which is represented as

$$\hat{Y} = \psi_{\theta_2}(\mathbf{z}_{v_1}, \mathbf{z}_{v_2}, \ldots, \mathbf{z}_{v_n}), \tag{13}$$

As in natural images, WSI augmentation is only used in the training phase to improve the generalization of the model. The image augmentation student branch and the representation augmentation student branch share weights in patch-level representation learning, determining that the WSI augmentation module can be skipped in the inference phase. Therefore, we can directly feed representations extracted from the original patches into the WSI classifier for prediction:

$$\hat{Y} = \psi_{\theta_2}(\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_n). \tag{14}$$

Experiments

The proposed method was evaluated based on WSI classification tasks in three datasets: 1) **USTC-EGFR**: a private dataset with 754 WSIs categorized into 5 types. 2) **TCGA-EGFR**: a public dataset with 696 WSIs categorized into 2 types. 3) **TCGA-LUNG-3K**: a public dataset with 3064 WSIs categorized into 3 types. Please refer to the supplemental material for detailed information about the datasets.

The WSIs were segmented into non-overlapping patches in size of 224 × 224 under 20× lenses for representation learning and extraction. Each dataset was split into training, validation, and testing subsets with the ratio of 6:1:3 at the patient-level.

We followed the implementation in DINO (Caron et al. 2021) to use ViT-S/16 (Dosovitskiy et al. 2021) as the backbone and the class token of ViT as the final feature embedding. Our evaluation metrics include average accuracy, micro-average area under the curve (AUC), and macro-average F1 score. All the methods were implemented in Python 3.8 with torch 1.8.1 and run on a computer cluster with 10 Xeon 2.66GHz CPUs and 10 GPUs of Nvidia Geforce 2080Ti.

Comparison With SOTA Methods

We compared our method with 5 different WSI representing strategies, including ImageNet (Russakovsky et al. 2015), SimCLR (Chen et al. 2020a), BYOL (Grill et al. 2020), MoCov3 (Chen, Xie, and He 2021) and DINO (Caron et al. 2021). The first three take ResNet50 (He et al. 2016) as the backbone, while the others utilize ViT-S (Dosovitskiy et al. 2021). Moreover, we compared our method with 4 WSI data augmentation frameworks, including DAGAN (Zaffar et al. 2022), ReMix (Yang et al. 2022), RankMix (Chen and Lu 2023), Intra-Mixup (Gadermayr et al. 2023). Given DINO is the basis of our representation learning framework, the comparative experiments with WSI augmentations are performed on the representations extracted by DINO.

As depicted in Figure 3, our method achieves the best performance on the USTC-EGFR dataset, TCGA-EGFR dataset and TCGA-LUNG-3K dataset under CLAM (Lu et al. 2021), TransMIL (Shao et al. 2021) and DTFD-MIL (Zhang et al. 2022a) benchmarks.

Comparison With Representation Learning Methods SimCLR (Chen et al. 2020a) is the self-supervised representation learning framework widely applied in feature extraction. The results on the TCGA-EGFR dataset show that it achieves 12.5%, 2.3%, 1.9% AUC better than ImageNet, which showcases the importance of addressing the semantic gap between natural and pathological images. MoCov3 (Chen, Xie, and He 2021) and DINO (Caron et al. 2021) utilize the ViT architecture as the backbone, demonstrating superior performance on the USTC-EGFR and TCGA-LUNG-3K datasets compared to SimCLR (Chen et al. 2020a) and BYOL (Grill et al. 2020). However, the observed performance gaps of MoCov3 and DINO on the TCGA-EGFR dataset compared to top-tier results indicate that current image augmentation strategies may not be sufficiently robust for generating discriminative features needed for effective WSI classification.

Our proposed PRDL additionally introduces representation augmentation into the process of self-supervised learning and employs augmentation prompts to control the representation augmentations, which improves the performance

by effectively increasing the diversity of representations. Compared with our baseline model DINO, PRDL achieves increase in AUC of 7.0%, 18.4%, 1.3% under the CLAM (Lu et al. 2021) benchmark, 6.8%, 19.2%, 2.7% under the TransMIL (Shao et al. 2021) benchmark and 5.2%, 24.4%, 1.7% under the DTFD-MIL (Zhang et al. 2022a) benchmark on the three datasets, respectively. Moreover, the promptable distribution estimator and the augmentation masks trained in representation learning can be utilized in the downstream task.

Comparison With WSI Augmentation Methods Mixup-based methods generate varied representations by mixing features at different levels. ReMix (Yang et al. 2022) proposes to mix instance prototypes formed by clustering. The performance of ReMix in AUC on the TCGA-EGFR dataset has improved by 8.8% under the CLAM benchmark when compared to DINO. Nevertheless, the performance degradation observed across all Mixup-based methods on the TCGA-LUNG dataset under the CLAM benchmark suggests that Mixup-based strategies may compromise semantic integrity, leading to noisy training samples and a subsequent decline in model performance. DAGAN (Zaffar et al. 2022) incorporates a Generative Adversarial Network (GAN) (Goodfellow et al. 2020) model to create new representations, adding another training process. It improves performance in AUC by 0.6%, 3.8%, 0.3% on the USTC-EGFR dataset. However, the generative model is trained after representation learning so that the performance is limited by the quality of the original representations. Moreover, DAGAN introduces extra computation costs in both training and inference phases.

Our WSI augmentation strategy PRS, which samples from prompted distributions, offers flexible augmentation, enhancing WSI classification without additional parameters. This can also be viewed as a reasonable perturbation to representations, which breaks the invariance of representations during training. With the guide of the augmentation prompts, the noise generated during sampling is much less than Mixup. Moreover, the distribution estimator is concurrently trained with representation learning, avoiding additional training overhead and continuously adapting to the representation change over the training procedure. With PRS, there are observed improvements in performance across three datasets, with increases in AUC of 9.4%, 15.2%, 2.1% under the CLAM (Lu et al. 2021) benchmark, 4.8%, 9.8%, 1.3% under the TransMIL (Shao et al. 2021) benchmark, 6.6%, 18.4%, 1.2% under the DTFD-MIL (Zhang et al. 2022a) benchmark compared to the second-best data augmentation methods.

Ablation Study

We first conduct ablation study to verify the necessity of model components. The results are detailed in Table 1. Here, we first investigate WSI augmentation within the feature space through random perturbation using a standard Gaussian distribution. It shows that the performance on the USTC-EGFR dataset is decreased in F1-Score by 14.8% compared to DINO under the CLAM benchmark, while it is increased by 1.3% under the DTFD-MIL benchmark. This suggests that while perturbations affect the model's performance, they do not completely ruin the representations. Meanwhile, we utilize Monte Carlo Sampling (Zheng et al. 2023) where patch representations are randomly discarded during the training of the WSI models. As a result, the performance under the DTFD-MIL benchmark is increased in AUC by 5.2% compared to DINO, which indicates that it works in some cases. However, this strategy has a high risk of losing important information for classification, thus leading to significant degradation of performance under the CLAM benchmark. The results of these two strategies show that the impact of purposeless perturbation on the patch representations is unstable, but opens the possibility of guiding perturbations to enhance the performance of WSI classification.

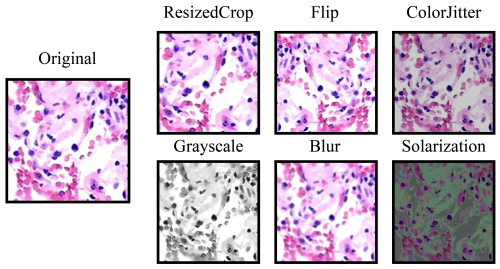
Methods	AUC	F1-Score	ACC
DINO	79.4	49.8	50.2
PRDL	86.4	52.5	57.9
PRDL+RS (w/o **M**)	86.0	59.6	59.6
PRDL+$\mathbf{m}_{ResizedCrop}$	89.3	61.2	61.9
PRDL+$\mathbf{m}_{HorizontalFlip}$	88.8	61.0	61.0
PRDL+$\mathbf{m}_{ColorJitter}$	88.9	61.9	62.3
PRDL+$\mathbf{m}_{Grayscale}$	89.3	62.1	63.2
PRDL+$\mathbf{m}_{GaussianBlur}$	89.0	62.7	62.8
PRDL+$\mathbf{m}_{Solarization}$	89.1	63.0	63.2
PRDL+PRS	**89.4**	**65.1**	**65.9**

Table 2: Impact of the different augmentation masks on the USTC-EGFR Dataset under the CLAM benchmark.

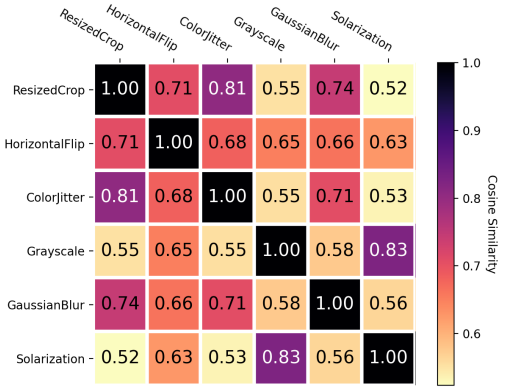
PRDL w/o PRA denotes PRDL without promptable representation augmentation related operations, where data augmentations are merely carried out in image space. The decrease in the evaluation metrics demonstrates that the PRA module plays a crucial role in representation learning. Considering that the representation distribution is estimated based on a Gaussian prior, we constrain it with KL divergence. The results of PRDL w/o $\mathcal{L}_{KL}$ show that KL divergence is essential for the training of the distribution estimator. Additionally, components like $\mathcal{L}_{sp}$ and $\mathcal{L}_{var}$ within the PRDL framework are crucial for effective learning of augmentation masks. Finally, we verify the impact of the PRS strategy on WSI classification. We observe that PRS can improve the performance of the WSI classifier even further building on PRDL, where the Micro-AUC under the CLAM benchmark increases from 0.864 to 0.894. This suggests that the representations generated by PRS are effective to the model and contain useful semantic information.

The Impact of Augmentation Prompts

As shown in Table 2, we study the impact of different augmentation prompts. The results of PRDL+RS(w/o **M**) indicate that an overly broad range of representation augmentations can result in limited improvement. Individual augmentation prompts produce better but varied results, yet they still fall short compared to PRDL+PRS that employs a random combination of the augmentation prompts. This demon-



(a) Image Augmentations



(b) Similarity Matrix

Figure 4: Dimensional impact between different augmentation prompts on the USTC-EGFR dataset, where (a) is the image augmentations corresponding to the prompts. (b) is the cosine similarities of the augmentation masks $\mathbf{M}$.

strates the importance of a guided and variable representation augmentation for improving model performance. Figure 4 shows that different augmentation prompts have different effects on the representation dimensions, and some augmentations cause similar impacts in feature space. *ReisizedCrop* and *HorizontalFlip* are both operated in spatial space and thus their corresponding prompts are similar. However, *HorizontalFlip* has less improvement in performance compared with other prompts, which indicates that the prompt obtained from this simple image transformation also has relatively limited effects on the augmented representations. *ColorJitter* and *GaussianBlur* have minimal effects on structural attributes of images, leading to similar impacts in feature space. Conversely, *Grayscale* and *Solarization* significantly alter the image, which directs focus towards key dimensions in representation, hence their prompts perform better.

Conclusion

We proposed a novel promptable representation distribution learning (PRDL) framework with a promptable representation sampling (PRS) strategy for promptable and efficient data augmentation in feature space for histopathology WSI classification. The approach involves a promptable distribution estimator and augmentation prompts for generating diverse representation augmentations, thereby improving the representation quality. This is complemented by a PRS strategy, tailored to leverage the trained estimator and prompts for effective WSI augmentation. The prompted representation augmentations significantly enhance image representations and preserve control, thus avoiding the loss of important semantic information, all of which are beneficial for WSI analysis.

Acknowledgments

This work was partly supported by Beijing Natural Science Foundation (Grant No. 7242270), partly supported by the National Natural Science Foundation of China (Grant No. 62171007, 61901018, and 61906058), partly supported by the Fundamental Research Fund for the Central Universities of China (grant No. YWF-23-Q-1075), partly supported by the Anhui Provincial Natural Science Foundation (Grant No. 2408085MF162), partly supported by Emergency Key Program of Guangzhou Laboratory (Grant No. EKPG21-32), partly supported by Joint Fund for Medical Artificial Intelligence (Grant No. MAI2023C014), and partly supported by National Key Research and Development Program of China (Grant No. 2021YFF1201004).

References

Assran, M.; Duval, Q.; Misra, I.; Bojanowski, P.; Vincent, P.; Rabbat, M.; LeCun, Y.; and Ballas, N. 2023. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15619–15629.

Bao, H.; Dong, L.; Piao, S.; and Wei, F. 2022. BEiT: BERT Pre-Training of Image Transformers. arXiv:2106.08254.

Bardes, A.; Ponce, J.; and LeCun, Y. 2022. VICReg: Variance-Invariance-Covariance Regularization for Self-Supervised Learning. arXiv:2105.04906.

Campanella, G.; Hanna, M. G.; Geneslaw, L.; Miraflor, A.; Werneck Krauss Silva, V.; Busam, K. J.; Brogi, E.; Reuter, V. E.; Klimstra, D. S.; and Fuchs, T. J. 2019. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature medicine*, 25(8): 1301–1309.

Caron, M.; Misra, I.; Mairal, J.; Goyal, P.; Bojanowski, P.; and Joulin, A. 2020. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in neural information processing systems*, 33: 9912–9924.

Caron, M.; Touvron, H.; Misra, I.; Jégou, H.; Mairal, J.; Bojanowski, P.; and Joulin, A. 2021. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, 9650–9660.

Chen, R. J.; Chen, C.; Li, Y.; Chen, T. Y.; Trister, A. D.; Krishnan, R. G.; and Mahmood, F. 2022. Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16144–16155.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020a. A simple framework for contrastive learning of visual repre-

sentations. In *International conference on machine learning*, 1597–1607. PMLR.

Chen, X.; Fan, H.; Girshick, R.; and He, K. 2020b. Improved Baselines with Momentum Contrastive Learning. arXiv:2003.04297.

Chen, X.; Xie, S.; and He, K. 2021. An Empirical Study of Training Self-Supervised Vision Transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 9640–9649.

Chen, Y.-C.; and Lu, C.-S. 2023. RankMix: Data Augmentation for Weakly Supervised Learning of Classifying Whole Slide Images With Diverse Sizes and Imbalanced Categories. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 23936–23945.

Dai, L.; Wang, Y.; Wang, H.; Zhang, Y.; et al. 2024. AugDiff: Diffusion based feature augmentation for multiple instance learning in whole slide image. *IEEE Transactions on Artificial Intelligence*.

Doersch, C. 2021. Tutorial on Variational Autoencoders. arXiv:1606.05908.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv:2010.11929.

Gadermayr, M.; Koller, L.; Tschuchnig, M.; Stangassinger, L. M.; Kreutzer, C.; Couillard-Despres, S.; Oostingh, G. J.; and Hittmair, A. 2023. Mixup-mil: Novel data augmentation for multiple instance learning and a study on thyroid cancer diagnosis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 477–486. Springer.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2020. Generative adversarial networks. *Communications of the ACM*, 63(11): 139–144.

Grill, J.-B.; Strub, F.; Altché, F.; Tallec, C.; Richemond, P.; Buchatskaya, E.; Doersch, C.; Avila Pires, B.; Guo, Z.; Gheshlaghi Azar, M.; et al. 2020. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33: 21271–21284.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9729–9738.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Huang, C.; Goh, H.; Gu, J.; and Susskind, J. 2023. MAST: Masked Augmentation Subspace Training for Generalizable Self-Supervised Priors. arXiv:2303.03679.

Ilse, M.; Tomczak, J.; and Welling, M. 2018. Attention-based deep multiple instance learning. In *International conference on machine learning*, 2127–2136. PMLR.

Jaiswal, A.; Babu, A. R.; Zadeh, M. Z.; Banerjee, D.; and Makedon, F. 2020. A survey on contrastive self-supervised learning. *Technologies*, 9(1): 2.

Li, B.; Li, Y.; and Eliceiri, K. W. 2021. 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*, 14318–14328.

Lu, M. Y.; Williamson, D. F.; Chen, T. Y.; Chen, R. J.; Barbieri, M.; and Mahmood, F. 2021. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature Biomedical Engineering*, 5(6): 555–570.

Nakhli, R.; Moghadam, P. A.; Mi, H.; Farahani, H.; Baras, A.; Gilks, B.; and Bashashati, A. 2023. Sparse Multi-Modal Graph Transformer With Shared-Context Processing for Representation Learning of Giga-Pixel Images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11547–11557.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. 2015. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115: 211–252.

Shao, Z.; Bian, H.; Chen, Y.; Wang, Y.; Zhang, J.; Ji, X.; et al. 2021. Transmil: Transformer based correlated multiple instance learning for whole slide image classification. *Advances in neural information processing systems*, 34: 2136–2147.

Wu, K.; Jiang, Z.; Tang, K.; Shi, J.; Xie, F.; Wang, W.; Wu, H.; and Zheng, Y. 2024. Pan-cancer Histopathology WSI Pre-training with Position-aware Masked Autoencoder. *IEEE Transactions on Medical Imaging*.

Yang, J.; Chen, H.; Zhao, Y.; Yang, F.; Zhang, Y.; He, L.; and Yao, J. 2022. 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*, 35–45. Springer.

Zaffar, I.; Jaume, G.; Rajpoot, N.; and Mahmood, F. 2022. Embedding Space Augmentation for Weakly Supervised Learning in Whole-Slide Images. arXiv:2210.17013.

Zang, Y.; Huang, C.; and Loy, C. C. 2021. Fasa: Feature augmentation and sampling adaptation for long-tailed instance segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3457–3466.

Zhang, H.; Meng, Y.; Zhao, Y.; Qiao, Y.; Yang, X.; Coupland, S. E.; and Zheng, Y. 2022a. Dtfd-mil: Double-tier feature distillation multiple instance learning for histopathology whole slide image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18802–18812.

Zhang, J.; Zhang, X.; Ma, K.; Gupta, R.; Saltz, J.; Vakalopoulou, M.; and Samaras, D. 2022b. Gigapixel whole-slide images classification using locally supervised

learning. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part II*, 192–201. Springer.

Zheng, Y.; Li, J.; Shi, J.; Xie, F.; Huai, J.; Cao, M.; and Jiang, Z. 2023. Kernel Attention Transformer for Histopathology Whole Slide Image Analysis and Assistant Cancer Diagnosis. *IEEE Transactions on Medical Imaging*.

Zhou, J.; Wei, C.; Wang, H.; Shen, W.; Xie, C.; Yuille, A.; and Kong, T. 2022. iBOT: Image BERT Pre-Training with Online Tokenizer. arXiv:2111.07832.