RankMix: Data Augmentation for Weakly Supervised Learning of Classifying Whole Slide Images with Diverse Sizes and Imbalanced Categories

Yuan-Chih Chen
IIS, Academia Sinica, Taiwan, ROC
willpower057@gmail.com

Chun-Shien Lu
IIS, Academia Sinica, Taiwan, ROC
lcs@iis.sinica.edu.tw

Abstract

Whole Slide Images (WSIs) are usually gigapixel in size and lack pixel-level annotations. The WSI datasets are also imbalanced in categories. These unique characteristics, significantly different from the ones in natural images, pose the challenge of classifying WSI images as a kind of weakly supervise learning problems. In this study, we propose, RankMix, a data augmentation method of mixing ranked features in a pair of WSIs. RankMix introduces the concepts of pseudo labeling and ranking in order to extract key WSI regions in contributing to the WSI classification task. A two-stage training is further proposed to boost stable training and model performance.

To our knowledge, the study of weakly supervised learning from the perspective of data augmentation to deal with the WSI classification problem that suffers from lack of training data and imbalance of categories is relatively unexplored.

1. Introduction

1.1. Background

Natural image processing tasks, including image classification and object detection, have been widely solved using deep learning models and obtain astounding results. In this study, we investigate how medical imaging can also benefit from deep learning with focus on whole slide images (WSIs). WSI scanning is commonly used in disease diagnosis [12, 34]. The demand of computer aided assessment makes deep learning widely adopted in this field [1]. Because WSI is a gigapixel image and lacks pixel-level annotations, multiple instance learning (MIL) [31, 32] is an exact solution to this weakly supervised learning problem [2]. In MIL, a WSI is often cropped into tens of thousands of patches and then an aggregator will make a prediction based on integrating these patches. Most recent works [5, 9, 10, 25, 28, 30, 37, 38, 44] focus on aggregator architecture design and improving feature extraction of the patches. However, because WSI is difficult to collect and share, we explore the possibility of data augmentation in WSI classification to increase training samples and mitigate the problem of class imbalance [19, 26, 33, 48] that WSI may have due to rare diseases (versus common diseases). In addition, patch feature extractor is often trained by self-supervised learning [24, 28] or comes from pre-trained models (such as pre-trained in ImageNet [30, 37] or WSI datasets [5, 11]). Therefore, for universality and portability, our work will focus on studying the feature domain instead of pixel domain of patches.

Traditionally, mixup methods [18, 48] are employed to mix photos of the same aspect ratio or vectors of the same dimension. Nevertheless, this is not the case for WSI as WSI intrinsically has a different number of patches, ranging from hundreds to hundred of thousands. This is because the generation of a WSI, caused by the tissue placement and the tissue size, will make WSIs of varying aspect ratios and sizes.

In addition, because a WSI tends to be a very large size (equivalent to tens of thousands of 224×224 patches or even larger) and the background often occupies a large part, it is better to use data pre-processing to remove unimportant background parts in order to save computation time and avoid possible unnecessary information [21, 39] (such as noise and artifacts). That is, the pre-processing step of cropping a WSI into patches, as shown in Fig. 1, will remove most of the background patches. The resultant WSI patches, however, will lose their absolute positions.

1.2. Challenges

The above characteristics of WSI lead to the difficulty of directly employing the traditional mixup methods. We cannot simply resize two WSIs to have the same size for the sake of performing mixup. This is because all WSIs are scanned at the same magnification (*e.g.*, 20x), the physical meanings will be lost if they are rescaled casually. More importantly, due to loss of absolute positions among the patches after removing the WSI background, resizing patches actually do not solve this problem.

Another commonly used techniques for data augmentation are based on cutting, including Cutout [14] and Cutmix [47]. The core of cutting aims at obtaining or removing parts of an images. However, the key difference between WSI images and natural images is that the main objects can occupy a major part of a natural image, but it is not the case in WSIs. For example, the tumor slide of Camelyon16 dataset [16] only has a small area of tumor (approximately $< 10\%$ of tissue area). Therefore, if a random cut is made, there is a non-negligible probability that the tumor slide will not contain any tumor patches.

To address these challenges, we propose a novel mixup method, called RankMix, for augmentation of whole slide images with diverse sizes and imbalanced categories. RankMix introduces the concepts of instance-level pseudo labeling and ranking in order to obtain meaningful WSI regions that can contribute to the WSI classification task. In order to further enhance model performance, two-stage training is proposed in that the first step is to train a stable score function by general MIL, and then the score function and mixup technique are jointly used in the second stage of training.

1.3. Our Contributions

Our contributions are summarized as follows:

- To our knowledge, MIL currently focuses on improving feature extraction and aggregator-based classification. It is relatively ignored in investigating weakly supervised learning from the perspective of data augmentation. Our proposed method can be applied to WSI classification problems and can be easily incorporated to existing MIL methods.
- In contrast to the existing mixup methods that aim at mixing natural images of the same size, our method can mix images (*e.g.*, WSIs) of different sizes.
- Because of rare diseases and the difficulty of medical image collection, the WSI classification problem is apt to suffer from lack of training data and imbalance of categories. Our proposed method is demonstrated to be feasible in addressing these challenges.

2. Related Work

In this section, we briefly introduce the techniques that are relevant to our work.

2.1. Data Augmentation

Data augmentation can improve the generalization and has been widely used in training neural networks. Mixup [48] fuses information from two images by convex combination. CutMix [47] inpaints the masking area (produced by random occlusion) with another image content at the same location. Recently, Part *et al*. [35] proposed how the majority can help the minority with re-balancing distribution of each sub-class. In addition to mixing in the pixel domain, Manifold Mixup [42] and PatchUp [18] proposed mixing the features of two images. Despite promising, the aforementioned methods are not presented for medical images.

For medical imaging, Galdran *et al*. [19] mixed two medical images sampled from different distributions, including instance-based sampling and class-based sampling. By increasing the sampling probability of the minority class, the class imbalance problem that usually occurs in medical imaging can be properly alleviated. In histopathology studies, the datasets often suffer from stain color variation because of different stain approaches, procedures, and slide scanner. To address this problem, Chang *et al*. [4] mixed two stain color matrices from stained images in order to increase the generalization of unseen colors. Chen *et al*. proposed Flow-Mixup [6] to regularize medical images with corrupted multi-labels because the annotations of medical images are costing and automatic annotation often provides corrupted labels. Gazda *et al*. [20] used the mixup technique to improve model performance for medical image segmentation. In ReMix [46], Yang *et al*. proposed using latent-space data augmentation to deal with WSI classification. However, ReMix only mixed instance prototypes of slides from the same class by K-Means while maintaining the original labels. From this perspective, ReMix acts like a kind of feature augmentation instead of general mixup.

As shown in Tab. 1, although the previous researches on medical images are diverse, their data augmentations are not feasible for WSIs with large gaps in size.

2.2. Multiple Instance Learning (MIL)

In recent MIL studies, the patches of a WSI are transformed into the features of fixed size by a feature extractor and then the patch features will be aggregated to get a final slide-level prediction. To leverage patch features, MIL-RNN [2] uses the recurrent neural network (RNN) [36] to encode position information. Because RNNs are apt to suffer from information loss over long distances, the follow-up works proposed attention-based MIL [25, 30] to calculate the contributions of instance-level features by learnable neural networks. Thanks to the success of Transformer [15, 41], self-attention based MIL methods [10, 28, 37] receive considerable attention in WSI classifications. Li *et al*. [28] proposed the non-local attention mechanism to calculate the relation between the critical feature and remaining features, but ignored the position relationship between patches. To address this problem, Shao *et al*. proposed TransMIL [37], which emphasizes the benefits of Pyramid Position Encoding Generator (PPEG), to encode spatial information by group convolution. Recently, Chikontwe *et*

Setting	Image type	Mixup input	Mixup output	Large size gap between 2 inputs
Mixup [48]	Natural image	Image	v	x
Manifold Mixup [42]	Natural image	Image, Hidden state	v	x
Balanced-MixUp [19]	Retinal image, Gastro-intestinal image	Image	v	x
Stain Mix-up [4]	Patches of WSI	Stain matrix	v	x
Flow-Mixup [6]	Chest X-ray, ECG	Hidden state	v	x
Mixup for KiTS [20]	CT image	Image	v	x
ReMix [46]	WSI	Instance prototype	x	v
RankMix (Ours)	WSI	Image feature	v	v

Table 1. Comparisons among mixup methods. Our method, RankMix, deals with mixing of not only two images with large gap in size in the feature domain but also their labels.

al. [10] proposed combining the critical feature [28] with PPEG [37], treating the tumor slides as a kind of out-of-distribution compared to normal slides, and re-calibrating the patch features of a slide according to the magnitudes of critical feature and slide label.

2.3. Self-Training and Knowledge Distillation

Knowledge distillation [23,40] uses soft labels instead of hard labels (which are the maximum of class probabilities) from the teacher model to train the student model. The student model, aiming for model compression, is a smaller one to mimic the output of its teacher model. Instead of compressing models, Xie *et al*. [45] integrated self-training with knowledge distillation to improve the image classification task with unlabeled data. The authors propagated pseudo soft labels of unlabeled data from the teacher model to the student model. The researches [3, 8, 17] further used self-training and knowledge distillation to deal with the self-supervised learning problem.

To sum up, our work, RankMix, has similar concepts with previous works, but we study how the general MIL model can be adopted as the teacher model. This makes RankMix act like a post-processing of traditional MIL, as illustrated in Fig. 2, and can be simply plugged into the exist approaches.

3. Preliminary

In this section, we introduce the classification problem of WSIs, as well as some baseline techniques and notations, to make this paper self-contained.

3.1. Problem Formulation

In the WSI classification problem, we have a series of WSIs $X = \{X_1, X_2, \ldots, X_n\}$ and corresponding slide labels $Y = \{Y_1, Y_2, \ldots, Y_n\}$ as dataset $\mathcal{D} = \{(X_i, Y_i), \quad i = 1, \ldots, n\}$. Our goal is to train an NN model based on dataset $\mathcal{D}$ for binary slide label $Y_i \in \{0, 1\}$ prediction of an incoming WSI image X_i. In MIL, a WSI can be treated as a bag (slide) containing the instances (patches) as $X_i = \{x_{i,1}, x_{i,2}, \ldots, x_{i,m(i)}\}$. Note that the WSI X_i's have different numbers $m(i)$'s of instances and the instance-level labels $\{y_{i,1}, y_{i,2}, \ldots, y_{i,m(i)}\}$ are unknown. Without loss of generality, we will mostly use m in place of $m(i)$ for notation simplicity. The slide label Y_i of a WSI X_i is defined to be negative ($Y_i = 0$) when all instances in a bag are negative (without tumors), *i.e*., $y_{i,j} = 0$ for all j. If at least one instance in X_i is positive (with tumor), its slide label will be positive ($Y_i = 1$). The slide label of WSI X_i is defined as:

$$Y_i = \begin{cases} 0, \text{iff } \sum y_{i,j} = 0 \\ 1, \text{otherwise} \end{cases}. \tag{1}$$

3.2. MIL as Baseline Model

In MIL, as shown in Fig. 1, the slide X_i will first be cut into many instances (patches) as $X_i = \{x_{i,1}, x_{i,2}, \ldots, x_{i,m}\}$ and then passed through an instance feature extractor G_θ to get the so-called features (embeddings) $H_i = \{h_{i,1}, h_{i,2}, \ldots, h_{i,m}\} \in \mathbb{R}^{m \times d}$ via Eq. (2) as:

$$h_{i,j} = G_\theta(x_{i,j}), \tag{2}$$

where d denotes the length of a feature vector and m denotes the number of features. As mentioned before, different WSIs will have different numbers of features but feature length d is the same. Finally, the aggregator will output a slide label prediction $\hat{Y}_i$ by Eq. (3) as:

$$\hat{Y}_i = aggregator(h_{i,1}, h_{i,2}, \ldots, h_{i,m}). \tag{3}$$

In this work, we use DSMIL [28] and FRMIL [10] as the backbone models to represent SOTA permutation-variant and permutation-invariant MIL models, respectively, where the former only considers the relationship between each embedding and does not consider relative positions, and the latter uses the pooling multi-head self-attention (PMSA) module and positional encoding module (PEM) to fuse relative

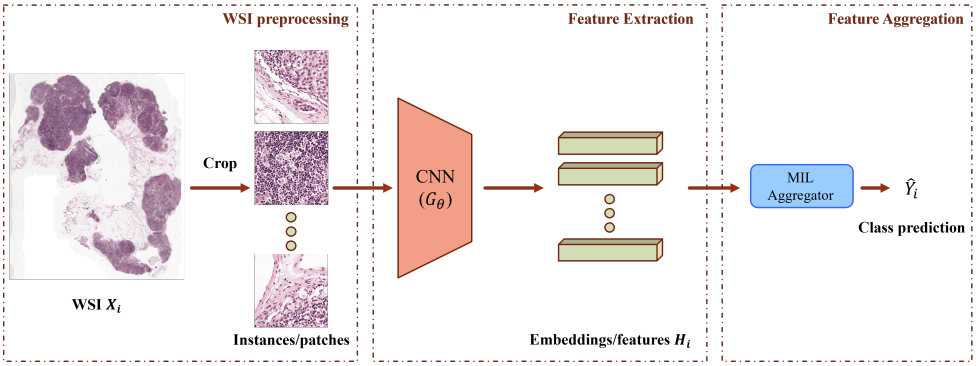


Figure 1. The workflow of general MIL.

spatial information. We study the impact of RankMix on these two models and demonstrate that our method can be plugged into the current MIL methods. The general loss function of an MIL model can be formulated as:

$$\mathcal{L}_{\mathcal{MIL}} = w_1 * \mathcal{L}_{bag}(\hat{Y}_i, Y_i) + \sum_{\ell} w_\ell * \mathcal{L}_\ell, \tag{4}$$

where $\mathcal{L}_{bag}$ is a binary cross-entropy (BCE) loss, $\mathcal{L}_\ell$ is other loss term that depends on different MIL models, and w_ℓ's are the corresponding balance weights.

3.3. Mixup

The mixup technique was first proposed in [48] as:

$$x_{mxp} = \lambda x_1 + (1-\lambda)x_2 \tag{5}$$

$$y_{mxp} = \lambda y_1 + (1-\lambda)y_2, \tag{6}$$

where the input sample x_i $(i = 1, 2)$ is drawn from the training dataset, the label corresponding to the input sample x_i is y_i, and $\lambda \in [0, 1]$ is sampled from $\sim$ Beta(α, α).

The main goal of mixup is to make linear combinations of inputs and outputs, respectively, and to ensure that the mapping of the mixed input and mixed output can maintain linear constraints. Therefore, the mixup technique is able to boost the generalization and robustness of an NN model. In addition, mixup can improve the performance of a model encountering the class imbalance problem [19].

4. Proposed Method: RankMix

RankMix is mainly composed of pseudo labeling, ranking, mixing ranked features, and self-training, as illustrated in Fig. 2. We will describe the motivation of our method and the role of each component in RankMix.

4.1. Motivation

A WSI is usually composed of normal tissue and special areas, as shown in Fig. 3a. Special areas can be tumors or defects, etc. If a pathologist obtains a WSI that contains tumors, his/her expertise is sufficient to make a correct decision of determining whether there is a tumor inside it. Based on this premise, even a pathologist obtains a partial image of the same WSI that contains most of the tumor areas, as shown in Fig. 3b, it suffices to determine that the image contains tumors. By contrast, if an entire WSI or a partial WSI contains only normal tissues, it can be easily determined that the WSI is normal, as shown in Figs. 3c and 3d, respectively.

Therefore, the hypothesis here is that cropping a WSI into patches will not affect tumor detection, not to mention the fact that dealing with the entire WSI as a whole is indeed impractical when computing power and memory consumption are taken into consideration.

Moreover, based on the problem introduced in Sec. 1.2, it is easy to lose physical meaning from picking patches and resizing them to have the same size. If we can select the areas of arbitrary sizes to represent the original slide, we can prepare two partial but representative slides of the same size for use in the mixup technologies. Nevertheless, the prerequisite is based on the fact that we can know which parts of the slide are sufficient to represent the original slide. This conflicts with the scenario that we only have the label of entire slide and do not have the instance-level labels corresponding to patches in the slide, as described in Sec. 3.1.

In view of the above challenges, we will explain in following subsections how to pick out fragments that are sufficient to represent a slide without instance-level labels. We will demonstrate how to get the pseudo instance-level labels in Sec. 4.2, how to select representative features in Sec. 4.3, and how to combine representative features with a mixup mechanism in Sec. 4.4. The module of pseudo labeling and ranking in RankMix (Fig. 2) is further detailed in Fig. 4.

4.2. Pseudo Labeling

Given $H_i = \{h_{i,1}, h_{i,2}, \ldots, h_{i,m}\} \in \mathbb{R}^{m \times d}$ from a WSI X_i by Eq. (2), the pseudo patch-level labels will be predicted to determine which patches are useful for mixup later. Specifically, if we have a score function $f \in \mathbb{R}^{d \times 1}$

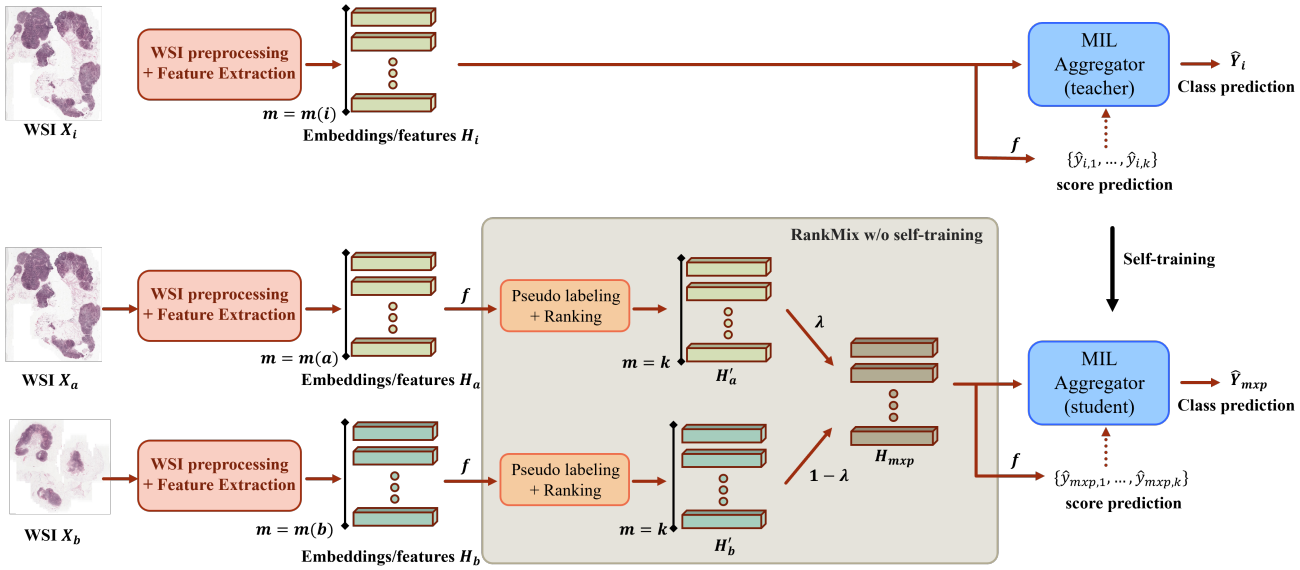


Figure 2. The flowchart of MIL with RankMix. We train the general MIL model with the training set X in the first stage (Top), and then the score function f obtained in the first stage is used to make pseudo labels for $\{X_a, X_b\}$ and train the MIL model with mixup mechanism in the second stage (Bottom). The dashed arrows indicate the patch scores that may be used by the aggregator depending on the MIL method.

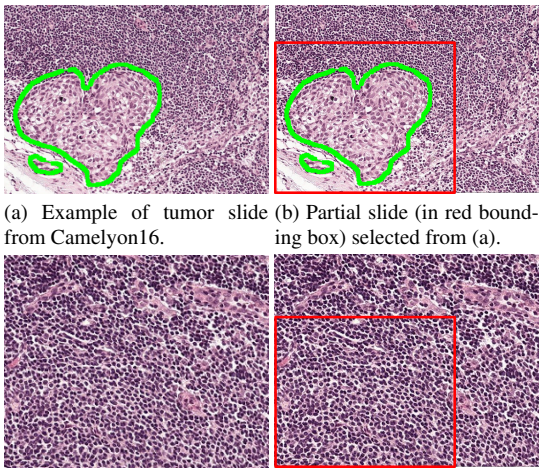
implemented by a multilayer perceptron (MLP) as:

$$\hat{y}_{i,j} = score_{i,j} = f(h_{i,j}), \quad \forall i = 1, \ldots, n \tag{7}$$

we can define a loss function $\mathcal{L}_{max}$ as:

$$\mathcal{L}_{max} = \text{BCE}(\hat{y}_{i,j^*}, Y_i), \tag{8}$$

where $j^* = \arg\max_j(\hat{y}_{i,j})$ and Y_i is obtained from Eq. (1). Note that the score function here can be calculated individually to deal with the multi-class classification problems (*e.g.*, TCGA-Lung dataset in Sec. 5.1) and can be replaced by any similar mechanisms of MIL, which calculates the importance of each patch to yield a final slide-level prediction (*e.g.*, attention mechanism [25, 28, 30] or the distance between clusters [9, 38, 44]). So far, we can get the class probability of every patch in a WSI by the score function f.



(a) Example of tumor slide from Camelyon16. (b) Partial slide (in red bounding box) selected from (a).

(c) Example of normal slide from Camelyon16. (d) Partial slide (in red bounding box) selected from (c).

Figure 3. Illustration of juding if a whole slide or partial slide contains tumors or not. The green annotation indicates the tumor regions and red bounding box indicates the partial region of WSI that is considered to have the same label as the original WSI.

4.3. Ranking

After obtaining the score function f, we already have the basic ability to achieve the motivation mentioned in Sec. 4.1. If we want to obtain a representative portion that can represent the original WSI, we need to score each feature $\{\hat{y}_{i,1}, \ldots, \hat{y}_{i,m}\} \in \mathbb{R}^{m \times 1}$ in the WSI and then sort the patch scores as:

$$\begin{aligned} &Z_i = \{z_{i,1}, z_{i,2}, \ldots, z_{i,m}\} \\ &\text{s.t. } \hat{y}_{i,z_{i,1}} > \hat{y}_{i,z_{i,2}} > \cdots > \hat{y}_{i,z_{i,m}} \quad \forall i = 1, \ldots, n. \end{aligned} \tag{9}$$

According to the number k (k is an arbitrary positive integer) of features we need, we can obtain the desired features $\bar{H}_i \in \mathbb{R}^{k \times d}$ from the original features $H_i \in \mathbb{R}^{m \times d}$ as:

$$\begin{aligned} &\bar{H}_i = \{h_{i,z_{i,1}}, h_{i,z_{i,2}}, \ldots, h_{i,z_{i,k}}\}. \\ &\text{for } k \leq m, \forall i = 1, \ldots, n \end{aligned} \tag{10}$$

Finally, in order to maintain the relative position information among patches, the features in $\bar{H}_i$ are rearranged in the original order to get the final representative features

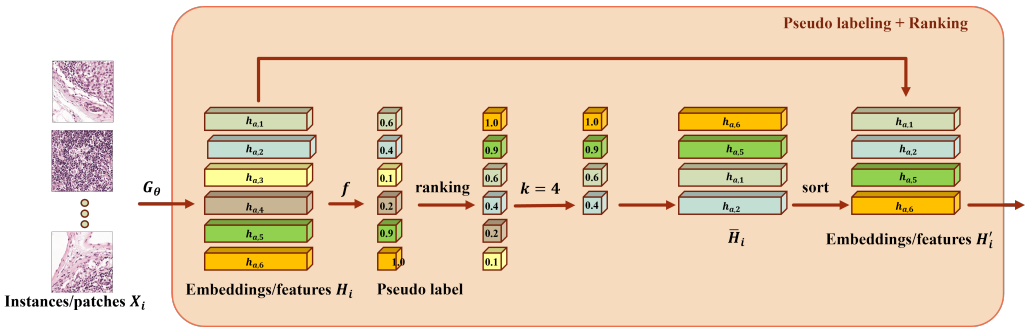


Figure 4. Illustration of the pseudo labeling and ranking mechanism.

$H'_i \in \mathbb{R}^{k\times d}$ as:

$$H'_i = \{h_{i,z'_{i,1}}, h_{i,z'_{i,2}}, \ldots, h_{i,z'_{i,k}}\}, \quad \text{s.t. } z'_{i,1} < z'_{i,2} < \cdots < z'_{i,k} \quad \forall i = 1, \ldots, n \tag{11}$$

where H'_i indicates the k features selected from the original i^{th} slide with the order preserved.

4.4. Mixup in Ranked Features

According to Eq. (5) and Eq. (6), we can formulate the mixup of features from two examples, (X_a, Y_a) and (X_b, Y_b), sampled from the dataset $\mathcal{D}$, as follows:

$$H_{mxp} = \lambda H'_a + (1-\lambda) H'_b \tag{12}$$

$$Y_{mxp} = \lambda Y_a + (1-\lambda) Y_b, \tag{13}$$

where $H'_a \in \mathbb{R}^{k\times d}$ and $H'_b \in \mathbb{R}^{k\times d}$ are obtained from $H_a \in \mathbb{R}^{m(a)\times d}$ and $H_b \in \mathbb{R}^{m(b)\times d}$, respectively, by Eq. (9)$\sim$ Eq. (11), and $m(a)$ and $m(b)$ denote the numbers of features obtained from X_a and X_b, respectively. Different from the existing mixup techniques, we conduct mixup in ranked features.

4.5. MIL Model with RankMix

As described in Sec. 3.2 and Sec. 4.2, RankMix can be applied to MIL models with the total loss being defined as:

$$\mathcal{L} = w_1 \mathcal{L}_{max}(\hat{y}_{mxp,j^*}, Y_{mxp}) + w_2 \mathcal{L}_{bag}(\hat{Y}_{mxp}, Y_{mxp}) + \sum_{\ell} w_\ell \mathcal{L}_\ell, \tag{14}$$

where the first term comes from Eq. (8) and the remaining terms come from Eq. (4). So far, we have explained how to accomplish the middle component of RankMix in MIL in Fig. 2, *i.e.*, the first stage of training.

4.6. Self-Training: Second Stage of Training

WSI classification faces the tough problem that each WSI is only given a slide-level label; that is, tens of thousands of patches correspond to one label. This kind of weakly supervised learning is more difficult to train than supervised learning. If the RankMix proposed so far is used, unstable training will be encountered. However, if the score function is more reliable (*i.e.*, can distinguish patch labels) at the beginning, then RankMix is found to achieve better results. We can imagine that the score function acts like expert annotations, and higher quality annotations will naturally result in better performance.

Inspired by self-training [40, 45] and BERT-based model [13, 27, 29], we design a score function by learning from a pre-trained task in advance (*i.e.*, can output the class probability distribution of a patch) as:

$$p(\hat{Y}_i|H_i) = aggregator(H_i) = p(\hat{Y}_i|h_{i,1}, \ldots, h_{i,m}). \tag{15}$$

Then, by using the concept of self-training, the stronger model is used as a teacher to extract the area that can represent the original WSI via score function for the student model to mixup ranked features as:

$$\begin{aligned}\hat{Y}_{mxp} &= p(Y_{mxp}|H_{mxp}) = p(Y_{mxp}|h_{mxp,1}, \ldots, h_{mxp,k}) \\ &= p(Y_{mxp}|\lambda H'_a + (1-\lambda) H'_b) \\ &= aggregator(\lambda H'_a + (1-\lambda) H'_b).\end{aligned} \tag{16}$$

From the perspective of BERT-based models, we use the teacher model to learn the WSI classification problem (also known as the pretext task) as pretraining, and treat the mixup of ranked features as the downstream task to further improve the performance. Please see Fig. 2 for the illustration of two-stage training procedure.

5. Experiments and Results

In this section, we will describe the datasets used for experiments in Sec. 5.1, experimental setting and performance metrics in Sec. 5.2, main results in Sec. 5.3, and ablation studies in Sec. 8.1 (Appendix).

Method/Dataset	Camelyon16			WSI-usability			TCGA-Lung		
	ACC	AUC	AUPRC	ACC	AUC	AUPRC	ACC	AUC	AUPRC
DSMIL [28]	86.82%	93.32%	92.68%	76.11%	86.60%	24.51%	93.81%	97.89%	97.75%
+ ReMix [46]	82.17%	86.89%	83.86%	83.19%	85.83%	25.59%	94.29%	97.62%	97.29%
+ RankMix w/o self-training	87.60%	92.07%	92.43%	90.27%	87.07%	25.66%	94.29%	98.00%	97.76%
+ RankMix	**89.92%**	**93.47%**	**92.74%**	**90.27%**	**88.16%**	**28.41%**	**94.29%**	**98.04%**	**97.79%**
FRMIL [10]	89.15%	94.57%	93.66%	83.19%	87.69%	45.99%	90.95%	95.38%	94.96%
+ ReMix [46]	82.59%	87.29%	87.35%	89.25%	80.63%	33.09%	92.22%	96.99%	97.04%
+ RankMix w/o self-training	90.70%	94.11%	93.68%	80.53%	84.27%	38.55%	93.33%	95.84%	97.01%
+ RankMix	**91.47%**	**94.59%**	**93.99%**	**93.81%**	**93.61%**	**47.65%**	**93.33%**	**97.00%**	**97.04%**

Table 2. Comparison of WSI classification between RankMix and vanilla models under three datasets.

5.1. Datasets

We introduce the WSI datasets used for experiments here, and a pre-processing setting was adopted to extract patches from each WSI. Each WSI was cropped into different numbers of 224×224 patches at 20x magnification without overlapping. We follow [10,28] to discard the background patches with tissue entropy less than 15% of a WSI.

Camelyon16 [16] is a public and well-known dataset proposed for metastasis detection in breast cancer. The dataset contains 270 training slides and 130 testing slides, and roughly has 4.1 million patches at 20x magnification with the maximum of 44000 patches per WSI and a minimum of 1200 patches per WSI in our pre-processing setting. If a WSI contains at least one tumor region, it is regarded as a positive slide. On the contrary, if the entire area of a WSI is normal, it is negative. In Camelyon16, the tumor area only accounts for approximately less than 10% of the tissue area in the positive slide.

TCGA-Lung is a public dataset that has two types of lung cancer, Lung Adenocarcinoma (TCGA-LUAD) and Lung Squamous Cell Carcinoma (TCGA-LUSC), from the Caner Genome Atlas (TCGA). Please refer to https://pubmed.ncbi.nlm.nih.gov/25691825/ for details. There are in total 1046 diagnostic WSIs, including 534 TCGA-LUAD and 512 TCGA-LUSC, were split into 836 training slides and 210 testing slides. After pre-processing, TCGA-Lung roughly has a minimum of 50 patches per slide and a maximum of 12700 patches per slide, with a total of 3.2 million patches.

WSI-usability is a private dataset used for distinguishing whether a WSI is usable or not. The purpose of this dataset is used for evaluating automatic quality control (QC). In order to have sufficient practicality, two pathologist experts each annotated 250 WSIs from TCGA, of which 50 WSIs were redundant, resulting in 450 WSIs in total. If it is a bad (labeled as positive) slide, it means that the WSI is not usable; otherwise, it is a good WSI and annotated as a negative slide. WSI-usability has only 23 positive slides, so it possesses a severe class imbalance problem. After pre-processing, WSI-usability roughly has a minimum of 700 patches per slide and a maximum of 120000 patches per slide, with a total of 5.7 million patches.

Summing up the above, the adopted datasets cover the imbalance/balanced classes, unbalanced/balanced bags, and single/multiple-class problem. Among them, WSI-usability has majority and minority categories with a huge gap, which indicates the rare diseases and other costly data. For the positive slides in Camelyon16, tumors occupy quite smaller areas than normal tissues, while the negative slide is all filled with normal tissues. Different from the other two datasets, TCGA-Lung has two types of positive slides, namely TCGA-LUSD and TCGA-LUSC, leading to a multi-class classification problem.

5.2. Experimental Setup and Evaluation Metrics

ResNet18 [22] was adopted as our backbone model that was trained by SimCLR [7] and Adam optimizer to obtain feature extractor G_θ with a mini-batch size of 512, learning rate of $1e-4$, and weight decay of $1e-5$. For Camelyon16, we used the weights of ResNet18 trained by Lee *et al*. [28], as described in [10]. Then, the embeddings $H_i \in \mathbb{R}^{m(i)\times 512}$ of a WSI X_i were obtained from the global average pooling layer of feature extractor G_θ.

For training aggregator, we used the Adam optimizer for 200 epochs with the mini-batch size of 1 (bag), learning rate of $2e-4$, and weight decay of $5e-3$. Note that the mini-batch for performing mixup of ranked features was 2, where the sampling methods of DSMIL and FRMIL followed [19] and [10], respectively. The number k of features used in RankMix was set to $min(m(a), m(b))$ for two WSIs X_a and X_b.

For two-stage training, we trained teacher models by the general MIL (*i.e.*, DSMIL and FR-MIL). Then, the teacher models were used to guide the training of student models by RankMix. The structures of both the teacher model and student model are the same, which can be considered as a kind of self-distillation mechanism. The temperature parameter τ commonly used in knowledge distillation was set to 1.0 (see Eq. (20) later).

The classification performance was evaluated in terms of

Method/Dataset	Camelyon16		WSI-Usability		TCGA-Lung	
	Accuracy	AUC	Accuracy	AUC	Accuracy	AUC
DSMIL [28]	86.82%	93.32%	76.11%	86.60%	93.81%	97.89%
DSMIL + Direct Mixup	84.50%	90.51%	89.38%	71.18%	94.29%	98.03%
DSMIL + Shrink Mixup	86.05%	92.42%	88.50%	69.94%	93.81%	97.95%
DSMIL + Duplicate Mixup	84.50%	91.96%	89.38%	72.43%	93.81%	97.99%
DSMIL + Random Mixup	87.60%	91.89%	88.50%	83.33%	94.29%	97.99%
DSMIL + RankMix	**89.92%**	**93.47%**	**90.27%**	**88.16%**	**94.29%**	**98.04%**
FRMIL [10]	89.15%	94.57%	83.19%	87.69%	90.95%	95.38%
FRMIL + Direct Mixup	88.37%	92.50%	59.29%	80.37%	92.38%	95.96%
FRMIL + Shrink Mixup	89.92%	93.16%	68.14%	57.09%	92.86%	96.88%
FRMIL + Duplicate Mixup	87.60%	92.78%	86.73%	74.77%	92.38%	96.73%
FRMIL + Random Mixup	90.70%	94.23%	77.88%	80.69%	93.33%	96.98%
FRMIL + RankMix	**91.47%**	**94.59%**	**93.81%**	**93.61%**	**93.33%**	**97.00%**

Table 3. Comparisons between RankMix and four custom mixup techniques.

the accuracy and area under the curve (AUC).

5.3. Main Results

First, we applied the proposed method, RankMix, to two baseline models, DSMIL [28] and FRMIL [10], and compared the performance with the vanilla models in Tab. 2. We can observe that self-training indeed boosts the overall model performance, as described in Sec. 4.6. Without self-training (*i.e.*, mixup in the model without being derived from a teacher model), RankMix can sometimes achieve slightly better results, but as the model and dataset vary, it can be seen that the performance is slightly degraded in few cases. On the contrary, when self-training is introduced, the performance can be improved further. A recent work, ReMix [46], which is a SOTA augmentation method in WSI, was included for comparison. It can be found that RankMix outperforms ReMix remarkably.

Second, we compared our method with the general mixup techniques under different custom settings in Tab. 3. This is because the existing mixup researches, to our knowledge, were conducted on the same feature number, they cannot be readily employed to be compared with our method under the scenario that two feature of different numbers were mixed. Here, we discuss four situations for mixing up two WSIs (H_a and H_b) with different numbers ($m(a)$ and $m(b)$) of features in the following.

1) Direct Mixup: Perform mixup directly regardless of feature numbers ($m(a)$ and $m(b)$), where the feature with a smaller number is padded with zeros. So, the feature number after mixing is defined as:

$$k_d = max(m(a), m(b)). \tag{17}$$

2) Shrink Mixup: Extract the same number k_s of features for mixing up by

$$k_s = \begin{cases} m(b), \text{iff } m(a) > m(b) \\ m(a), \text{otherwise} \end{cases}. \tag{18}$$

3) Duplicate Mixup: Duplicate the feature with a smaller number to have the same size k_{du} as the larger feature as:

$$k_{du} = \begin{cases} m(a), \text{iff } m(a) > m(b) \\ m(b), \text{otherwise} \end{cases}. \tag{19}$$

4) Random Mixup: By randomly performing both the duplicate and shrink mixups.

Compared with the above four straightforward cases, it can be seen from Tab. 3 that RankMix is a sophisticated design according to the characteristics of WSI, described in Sec. 1.2, and generally performs better than others.

Finally, we conduct ablation studies to further evaluate RankMix in two aspects. On the one hand, we examine how the score functions learned under different pre-trained models, as described in Sec. 4.6, affect RankMix. On the other hand, we examine how different knowledge transfer techniques affect the student model performance in WSI classification. Please refer to Sec. 8.1 in Appendix for details.

6. Conclusion

In this work, we investigate weakly supervised learning from the perspective of data augmentation to deal with the WSI classification problem that suffers from lack of training data and imbalance of categories. A new data augmentation method, RankMix, of is proposed to mix ranked features in a pair of WSIs with different sizes. RankMix is composed of pseudo labeling and ranking for extracting key WSI regions, and two-stage training for boosting stable training and model performance.

7. Acknowledgement

This work was supported in part by the Ministry of Science and Technology of Taiwan, ROC, under Grant MOST 110-2634-F-006-022.

References

[1] Sugata Banerji and Sushmita Mitra. Deep learning in histopathology: A review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 12(1), 2022. 1

[2] Gabriele Campanella, Matthew G Hanna, Luke Geneslaw, Allen Miraflor, Vitor Werneck Krauss Silva, Klaus J Busam, Edi Brogi, Victor E Reuter, David S Klimstra, and Thomas J Fuchs. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature medicine*, 25(8):1301–1309, 2019. 1, 2

[3] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9650–9660, 2021. 3, 11

[4] Jia-Ren Chang, Min-Sheng Wu, Wei-Hsiang Yu, Chi-Chung Chen, Cheng-Kung Yang, Yen-Yu Lin, and Chao-Yuan Yeh. Stain mix-up: Unsupervised domain generalization for histopathology images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 117–126. Springer, 2021. 2, 3

[5] Chengkuan Chen, Ming Y Lu, Drew FK Williamson, Tiffany Y Chen, Andrew J Schaumberg, and Faisal Mahmood. Fast and scalable search of whole-slide images via self-supervised deep learning. *Nature Biomedical Engineering*, pages 1–15, 2022. 1

[6] Jintai Chen, Hongyun Yu, Ruiwei Feng, Danny Z Chen, et al. Flow-mixup: Classifying multi-labeled medical images with corrupted labels. In *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 534–541. IEEE, 2020. 2, 3

[7] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 7

[8] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E Hinton. Big self-supervised models are strong semi-supervised learners. *Advances in neural information processing systems*, 33:22243–22255, 2020. 3, 11

[9] Philip Chikontwe, Meejeong Kim, Soo Jeong Nam, Heounjeong Go, and Sang Hyun Park. Multiple instance learning with center embeddings for histopathology classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 519–528. Springer, 2020. 1, 5

[10] Philip Chikontwe, Soo Jeong Nam, Heounjeong Go, Meejeong Kim, Hyun Jung Sung, and Sang Hyun Park. Feature re-calibration based multiple instance learning for whole slide image classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 420–430. Springer, 2022. 1, 2, 3, 7, 8, 11, 12

[11] Ozan Ciga, Tony Xu, and Anne Louise Martel. Self supervised contrastive learning for digital histopathology. *Machine Learning with Applications*, 7:100198, 2022. 1

[12] Toby C Cornish, Ryan E Swapp, and Keith J Kaplan. Whole-slide imaging: routine pathologic diagnosis. *Advances in anatomic pathology*, 19(3):152–159, 2012. 1

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186, 2019. 6, 11

[14] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017. 2

[15] Alexey Dosovitskiy et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 2

[16] Babak Ehteshami Bejnordi, Mitko Veta, Paul Johannes van Diest, Bram van Ginneken, Nico Karssemeijer, Geert Litjens, Jeroen A. W. M. van der Laak, , and the CAMELYON16 Consortium. Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer. *JAMA*, 318(22):2199–2210, 2017. 2, 7

[17] Zhiyuan Fang, Jianfeng Wang, Lijuan Wang, Lei Zhang, Yezhou Yang, and Zicheng Liu. Seed: Self-supervised distillation for visual representation. In *International Conference on Learning Representations*, 2020. 3

[18] Mojtaba Faramarzi, Mohammad Amini, Akilesh Badrinaaraayanan, Vikas Verma, and Sarath Chandar. Patchup: A feature-space block-level regularization technique for convolutional neural networks. In *AAAI*, volume 36, pages 589–597, 2022. 1, 2

[19] Adrian Galdran, Gustavo Carneiro, and Miguel A González Ballester. Balanced-mixup for highly imbalanced medical image classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 323–333. Springer, 2021. 1, 2, 3, 4, 7

[20] Matej Gazda, Peter Bugata, Jakub Gazda, David Hubacek, David Jozef Hresko, and Peter Drotar. Mixup augmentation for kidney and kidney tumor segmentation. In *International Challenge on Kidney and Kidney Tumor Segmentation*, pages 90–97. Springer, 2022. 2, 3

[21] Maryam Haghighat et al. Automated quality assessment of large digitised histology cohorts by artificial intelligence. *Scientific Reports*, 12(1):1–16, 2022. 1

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 7

[23] Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2(7), 2015. 3, 11

[24] Ziwang Huang, Hua Chai, Ruoqi Wang, Haitao Wang, Yuedong Yang, and Hejun Wu. Integration of patch features through self-supervised learning and transformer for survival analysis on whole slide images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 561–570. Springer, 2021. 1

[25] Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *International conference on machine learning*, pages 2127–2136. PMLR, 2018. 1, 2, 5

[26] Lie Ju, Xin Wang, Lin Wang, Tongliang Liu, Xin Zhao, Tom Drummond, Dwarikanath Mahapatra, and Zongyuan Ge. Relational subsets knowledge distillation for long-tailed retinal diseases recognition. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 3–12. Springer, 2021. 1

[27] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations. In *International Conference on Learning Representations*, 2019. 6, 11

[28] Bin Li, Yin Li, and Kevin W Eliceiri. Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14318–14328, 2021. 1, 2, 3, 5, 7, 8, 11, 12

[29] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. 6, 11

[30] Ming Y Lu, Drew FK Williamson, Tiffany Y Chen, Richard J Chen, Matteo Barbieri, and Faisal Mahmood. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature biomedical engineering*, 5(6):555–570, 2021. 1, 2, 5

[31] Ming Y Lu, Melissa Zhao, Maha Shady, Jana Lipkova, Tiffany Y Chen, Drew FK Williamson, and Faisal Mahmood. Deep learning-based computational pathology predicts origins for cancers of unknown primary. *arXiv preprint arXiv:2006.13932*, 2020. 1

[32] Oded Maron and Tomás Lozano-Pérez. A framework for multiple-instance learning. *Advances in neural information processing systems*, 10, 1997. 1

[33] Yassine Marrakchi, Osama Makansi, and Thomas Brox. Fighting class imbalance with contrastive learning. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 466–476. Springer, 2021. 1

[34] Liron Pantanowitz, Paul N Valenstein, Andrew J Evans, Keith J Kaplan, John D Pfeifer, David C Wilbur, Laura C Collins, and Terence J Colgan. Review of the current state of whole slide imaging in pathology. *Journal of pathology informatics*, 2(1):36, 2011. 1

[35] Seulki Park, Youngkyu Hong, Byeongho Heo, Sangdoo Yun, and Jin Choi. The majority can help the minority: Context-rich minority oversampling for long-tailed classification. In *CVPR*, pages 6887–6896. IEEE/CVF, 2022. 2

[36] Mike Schuster and Kuldip K Paliwal. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11):2673–2681, 1997. 2

[37] Zhuchen Shao, Hao Bian, Yang Chen, Yifeng Wang, Jian Zhang, Xiangyang Ji, et al. Transmil: Transformer based correlated multiple instance learning for whole slide image classification. *Advances in Neural Information Processing Systems*, 34:2136–2147, 2021. 1, 2, 3

[38] Yash Sharma, Aman Shrivastava, Lubaina Ehsan, Christopher A Moskaluk, Sana Syed, and Donald Brown. Cluster-to-conquer: A framework for end-to-end multi-instance learning for whole slide image classification. In *Medical Imaging with Deep Learning*, pages 682–698. PMLR, 2021. 1, 5

[39] Gijs Smit, Francesco Ciompi, Maria Cigéhn, Anna Bodén, Jeroen van der Laak, and Caner Mercan. Quality control of whole-slide images through multi-class semantic segmentation of artifacts. *Medical Imaging with Deep Learning*, 2021. 1

[40] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pages 10347–10357. PMLR, 2021. 3, 6, 11

[41] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2

[42] Vikas Verma, Alex Lamb, Christopher Beckham, Amir Najafi, Ioannis Mitliagkas, David Lopez-Paz, and Yoshua Bengio. Manifold mixup: Better representations by interpolating hidden states. In *International Conference on Machine Learning*, pages 6438–6447. PMLR, 2019. 2, 3

[43] Longhui Wei, An Xiao, Lingxi Xie, Xiaopeng Zhang, Xin Chen, and Qi Tian. Circumventing outliers of autoaugment with knowledge distillation. In *European Conference on Computer Vision*, pages 608–625. Springer, 2020. 11

[44] Chensu Xie, Hassan Muhammad, Chad M Vanderbilt, Raul Caso, Dig Vijay Kumar Yarlagadda, Gabriele Campanella, and Thomas J Fuchs. Beyond classification: Whole slide tissue histopathology analysis by end-to-end part learning. In *Medical Imaging with Deep Learning*, pages 843–856. PMLR, 2020. 1, 5

[45] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10687–10698, 2020. 3, 6, 11

[46] Jiawei Yang, Hanbo Chen, Yu Zhao, Fan Yang, Yao Zhang, Lei He, and Jianhua Yao. Remix: A general and efficient framework for multiple instance learning based whole slide image classification. In *Medical Image Computing and Computer Assisted Intervention–MICCAI*, pages 35–45, 2022. 2, 3, 7, 8

[47] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6023–6032, 2019. 2

[48] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018. 1, 2, 3, 4