Attention-Guided Prototype Mixing: Diversifying Minority Context on Imbalanced Whole Slide Images Classification Learning

Farchan Hakim Raswa[1] Chun-Shien Lu[2] Jia-Ching Wang[1]
[1] National Central University, Taiwan, ROC [2] IIS, Academia Sinica, Taiwan, ROC
farchan.hakim.r@g.ncu.edu.tw, lcs@iis.sinica.edu, jcw@csie.ncu.edu.tw

Abstract

Real-world medical datasets often suffer from class imbalance, which can lead to degraded performance due to limited samples of the minority class. In another line of research, Transformer-based multiple instance learning (Transformer-MIL) has shown promise in addressing the pairwise correlation between instances in medical whole slide images (WSIs) with gigapixel resolution and non-uniform sizes. However, these characteristics pose challenges for state-of-the-art (SOTA) oversampling methods aiming at diversifying the minority context in imbalanced WSIs.

In this paper, we propose an Attention-Guided Prototype Mixing scheme at the WSI level. We leverage Transformer-MIL training to determine the distribution of semantic instances and identify relevant instances for cutting and pasting across different WSI (bag of instances). To our knowledge, applying Transformer is often limited by memory requirements and time complexity, particularly when dealing with gigabyte-sized WSIs. We introduce the concept of prototype instances that have smaller representations while preserving the uniform size and intrinsic features of the WSI.

We demonstrate that our proposed method can boost performance compared to competitive SOTA oversampling and augmentation methods at an imbalanced WSI level.

1. Introduction

The demand for computer-assisted diagnosis is increasing steadily. Whole-slide scanning, a widely used tool in disease diagnosis, facilitates the visualization of tissue sections. This scanning process involves transforming tissues on glass into digital whole slide images (WSIs) [20, 44]. However, assisting in the diagnosis of WSIs still poses challenges: 1) the gigapixel resolution and 2) the lack of pixel-level annotations. Weakly supervised multiple instance learning (MIL) offers an effective solution for handling WSIs by dividing them into small instances and then constructing an aggregator classifier to make disease predictions.

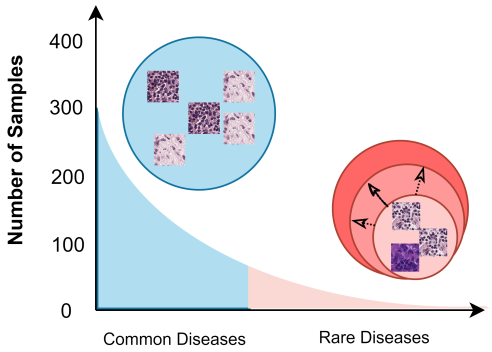

Figure 1. The imbalanced distribution in real medical settings.

Current research efforts heavily focus on the design of aggregators and the enhancement of feature extraction for WSI instances [21, 28, 40]. In this approach, each WSI is treated as a bag containing multiple instances. A WSI (bag) is labeled as disease-positive if any of its instances are disease-positive. The aggregator classifier examines the instance-level predictions and predicts the slide-level labels. To enhance WSI classification, a Transformer module has been integrated with the aggregator [28, 30, 41, 42]. This module incorporates a self-attention mechanism, allowing it to attend to multiple instances within a bag. This provides an advantage to the aggregator, as it can consider the correlation information between instances when making disease predictions.

In another line of research, real-world datasets often exhibit imbalanced or long-tailed class distributions [10, 12, 14, 32]. Some classes have a large number of samples (majority), while others have a scarcity of samples (minority), as illustrated in Figure 1. When aggregator classifiers are trained on imbalanced classes, they can become biased towards the majority class and tend to have poor abilities in recognizing the minority class.

To address this imbalance issue, oversampling augmentation has been proposed to enhance the classifier's performance. Balanced-MixUp [15] and CMO [27] are employed to mix images with uniform sizes from the majority and minority classes. Given an original image from a minority class, they randomly select regions of various sizes and paste them onto different images from the majority class. The label pairs are mixed according to a specific combination ratio. This approach addresses the problem of naive oversampling [29], which can intensify overfitting by duplicating samples without introducing sufficient diversity. While such methods have shown substantial performance gains on imbalanced natural image datasets, we argue that they may not be as helpful for medical images, especially for whole slide images (WSIs). This is because the representation of WSI in the MIL classifier involves a bag of instances that exhibit non-uniform sizes [9], and the WSI that contain tumor instances are limited ($< 20\%$ of the bag) and randomly distributed [2].

To address the above challenges, this work presents a self-attention mechanism to identify the most relevant instances with its corresponding slide label based on the Transformer-MIL classifier. Specifically, our key contributions are summarized as follows:

- We introduce a novel attention-guided approach to diversify the minority context in imbalanced whole slide image (WSI) classification learning. Unlike existing methods such as CMO and Balanced-MixUp, which randomly mix the minority context and maintain uniform size, our proposed approach achieves diversity by selectively mixing the context and allowing for non-uniform sizes (*e.g.*, WSI).

- To our knowledge, the Transformer module is known to have limitations due to its complexity and is only suitable for handling shorter sequences (*e.g.*, less than 100) [28, 33, 41]. We introduce the concept of prototype instances derived from parameterized instances with neural network on each individual bag. These prototype instances have smaller representations while preserving the uniform aspect size and intrinsic features of a WSI.

- We empirically demonstrate the effectiveness of our proposed method through experiments and ablation studies in imbalanced WSI classification.

2. Related works

We introduce the prior studies relevant to our work in this section, and remark the main difference between our method and previous works.

2.1. MIL and Transformer for WSI classification

In the early stages, MIL aggregators were designed using handcrafted approaches such as mean-pooling and max-pooling [24]. With the advent of deep neural network, assigning aggregators with neural network weights has proven to be more advantageous. An attention-based aggregator model associated with a neural network was proposed in ABMIL [18]. CLAM [23] introduced an attention mechanism in front of the slide-level classifier to enable the aggregator model to identify relevant regions within the slide image. DSMIL [21] proposed a non-local attention mechanism to calculate the relationships between critical instances and the remaining instances. DTFD-MIL [40] presented pseudo instance labels and a double-tier MIL approach. In case of imbalanced bags, MuRCL [43] investigated the latent relationships among instances through reinforcement learning to discriminate negative and positive instances. Liu *et al.* [22] proposed using pseudo instance labels obtained from the aggregator to fine-tune a feature extractor, enabling it to distinguish between negative and positive instances. These methods assume independence among instances within each bag. However, in real-world clinical settings, pathologists consider both the surrounding areas around a single instance and the correlations among multiple instances when making a diagnosis [1, 19].

Recently, the Transformer module is designed to measure multidimensional relationship between pairs of sequence instances [6, 13], as shown in Figure 2. TransMIL [28] introduced the integration of Transformers with the aggregator MIL classifier. Kernel attention Transformer (KAT) [42] presented a cross-attention paradigm to maintain near-linear computational efficiency when processing giga-sized WSIs. Zheng *et al.* [41] proposed a graph Transformer (GTP) that combines a graph-based representation of a WSI with a Transformer for disease prediction. H2T [31] proposed prototypical patterns for constructing holistic WSI-level representations. The Transformer module is utilized to retrieve information and dependencies of each prototypical pattern. However, applying the Transformer encounters the issues on memory requirements and time complexity, especially when dealing with gigabyte-sized WSIs [34, 35].

2.2. Oversampling-Augmentation Methods

The issue of class imbalance in computer vision tasks has been addressed by employing oversampling to modify the training distributions. The simplest form of oversampling is random oversampling (ROS) [29], which involves replicating the minority samples until its number of samples is equal to that in the majority class. While this method is straightforward and applicable to various applications, it could lead to overfitting due to the duplications of identical samples. An advanced oversampling method,

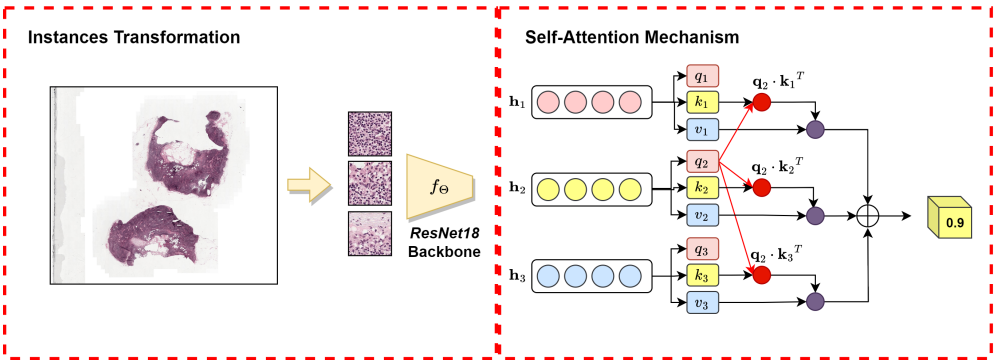

Figure 2. Self-attention mechanism via Transformer module of multiple instance learning [28].

known as the synthetic minority oversampling technique (SMOTE) [5], was proposed to generate the new minority samples by mixing the existing minority samples with their nearest minority neighbors. DeepSMOTE [11] introduced an oversampling technique at the instance-level. This approach involves encoding the data into instances using a generator module and then utilizing SMOTE to generate new synthetic instances. Generative adversarial minority sampling (GAMO) [26] utilized a generator module to generate new samples from the minority classes. The generator-generated samples are a convex combination of existing samples with the aim of misleading both the discriminator and classifier into misclassifying the generated samples. Balanced-MixUp [15] combined imbalanced (instance-based) and balanced (class-based) sampling. It incorporates MixUp [39] as a regularization technique and employs a modified oversampling strategy for imbalanced datasets. Context-Rich minority oversampling (CMO) [27] offered a simple approach that not only duplicated minority samples but also diversified their contexts. This goal can be achieved by selecting random regions from the minority images and pasting them onto the majority images, thereby creating new minority samples with varied contexts.

2.3. Augmention and MixUp Methods

Spatial augmentation methods have significantly enhanced computer vision performance. MixUp [39] introduced a weighted combination of random samples from the dataset. CutMix [38] filled a random region with a region from another sample. TransMix [7] mixed the labels based on the attention score of the Transformer module.

For WSI-level augmention, ReMix [37] introduced latent augmentation for WSIs. This method involved mixing the instance prototypes of WSIs (bags) from the same class using K-Means clustering while preserving the original labels. RankMix [9] proposed an augmentation technique that incorporates mixing ranked features within a bag. RankMix utilizes the concepts of pseudo-labeling and ranking to extract key regions from WSIs. Both ReMix and RankMix utilized interpolation for bag mixing, but this kind of approach possibly leads to unnatural combinations [38].

2.4. Remarks

Our proposed method shares a similar conceptual framework with the Balanced-MixUp [15] and CMO [27]; however, it exhibits two fundamental differences to be suitable for imbalanced WSI classification: 1) Instead of randomly cutting instances from the minority sample, our method focuses on identifying which relevant instances can be cut and pasted with other WSI (bag). It tackles the issue by limiting the tumor instance to limited ($< 20\%$ of the bag) and randomly distributed. 2) For those methods, [15, 27] that can be only applicable to mixed imbalanced samples with uniform size, our method can deal with mixed imbalanced samples with non-uniform size, such as WSIs.

On the other hand, in contrast to TransMix [7], our method diversifies the sample contexts by mixing based on the attention score. Moreover, our approach differs from TransMIL [28] that focuses on the Pyramid Position Encoding Generator (PPEG), KAT with a cross-attention paradigm [42], GTP that incorporates a minicut pooling layer [41], and H2T that utilizes clustering to select centroid instances as prototypical patterns [31]. On the contrary, we draw inspiration from the successful capture of intrinsic features achieved by parameterized instances with neural network in DSMIL [21] and DTFD-MIL [40]. Our approach introduces the concept of prototype instances derived from parameterized instances with neural network on each individual bag. These prototype instances have smaller representations while preserving the uniform size and intrinsic features of a WSI.

3. Preliminary

We briefly introduce Transformer-based MIL and CMO to make this paper self-contained.

3.1. Transformer-based MIL

In MIL, the slide image W is divided into small patches $X = \{x_1, x_2, x_3, \ldots, x_n\}$. These patches are then transformed into instance embeddings $H = \{h_1, h_2, h_3, \ldots, h_n\} \in \mathbb{R}^{n \times d}$, called as bag of instances, using the following process:

$$h_n = f_\theta(x_n) \tag{1}$$

where n denotes the number of instance embeddings within a single bag and d represents the length of the transformed instance. To extract features, we can utilize ResNet [17] as a feature extractor f_θ, and employ contrastive learning as a loss function to transform X into H [8].

The instance embeddings are used as input tokens and then concatenated together with a learnable classification token x_{class}. This x_{class} will be served as the output prediction of Transformer module thereafter [13]. The embedded sequence $\mathbf{x}$ can be expressed as:

$$\mathbf{x} = [x_{\text{class}}; H] \tag{2}$$

The embedded sequence $\mathbf{x}$ is then passed through the Transformer-based MIL. In our study, we utilize TransMIL [28] and GTP [41] as the aggregator of Transformer-based MIL models.

TransMIL consists of two Transformer modules and a position encoding layer, specifically designed for aggregating morphological information and encoding spatial information, respectively. On the other hand, GTP represents an input as a hierarchical representation. The WSI (bag) is constructed as the graph, followed by the Transformer as the aggregator classifier.

We investigate the impact of our proposed method on these two aggregator and demonstrate its effectiveness in addressing imbalanced class distribution within the current Transformer-based MIL framework.

3.2. Context-rich Minority Oversampling (CMO)

CMO [27] utilizes the information from majority samples to enhance the limited context of minority samples. Consider a training sample and its label, represented by x and y, respectively. The CMO creates a novel sample $(\tilde{x}, \tilde{y})$ by mixing two training samples (x^M, y^M) and (x^m, y^m), where x^M is randomly selected from the majority class and x^m is randomly selected from the minority class. CMO adopts the CutMix approach [38] to augment these pairs as follows:

$$\tilde{x} = \mathbf{M} \odot x^M + (\mathbf{1} - \mathbf{M}) \odot x^m \tag{3}$$

$$\tilde{y} = \lambda y^M + (1 - \lambda) y^m \tag{4}$$

where $(\mathbf{1} - \mathbf{M})$ is a binary mask that identifies the region to be selected from the minority sample and then merged with the majority sample, $\mathbf{1}$ represents a binary mask filled entirely with ones, and $\odot$ denotes element-wise multiplication. The mixing ratio λ between two samples is randomly sampled from a beta distribution $Beta(\alpha, \alpha)$.

4. Proposed Method: Attention-Guided Prototype Mixing

We propose a new oversampling augmentation for imbalanced whole slide images (WSIs). Our method, dubbed Attention-Guided Prototype Mixing, learns the distribution of the semantic WSI (prototype instances) using the learned attention score $\mathbf{A}$ of the Transformer module without modifying the aggregator design.

As shown in Figure 3, the prototype instances belonging to the tumor class (minority) are mixed with prototype instances from the normal class (majority). In our Transformer-based MIL module, we project the prototype instances P of the tumor class to generate the queries $\mathbf{q}$, keys $\mathbf{k}$, and values $\mathbf{v}$. The self-attention mechanism (Figure 2) of the module attends to the correlation between instances by computing the relevance between $\mathbf{q}$ and $\mathbf{k}$, summarizing which instances are most attentive on $\mathbf{v}$ for the final aggregator classifier. These most attentive instances can be selected and pasted with prototype instances from the normal class to diverse minority contexts.

Our proposed method is primarily composed of three components, including the re-weighting class distribution (**Q**) in Section 4.1, prototype instances building (**P**) in Section 4.2, and the most attentive instances mixing (**M**) in Section 4.3.

We provide more details regarding the definition of mathematical symbols, total loss function, and interpretation of proposed method in the supplement.

4.1. Re-weighting class distribution (Q)

Before we select the prototype instances from tumor slide image (bag), we first re-weight the class distribution Q of imbalanced WSI classes. We are inspired by the CMO [27] to assign the weight inversely proportional to class frequencies.

Specifically, given W_k as the number of slide image in the k-th class, the sampling weight $Q(W, k)$ is defined as:

$$Q(W, k) = \frac{\frac{1}{W_k}}{\sum_{k'=1}^{C} \frac{1}{W_{k'}}}, \tag{5}$$

where C denote as number of classes and the k-th class has a sampling weight inversely proportional to W_k. With this, the numbers of slide images in both the tumor and normal classes will be the same.

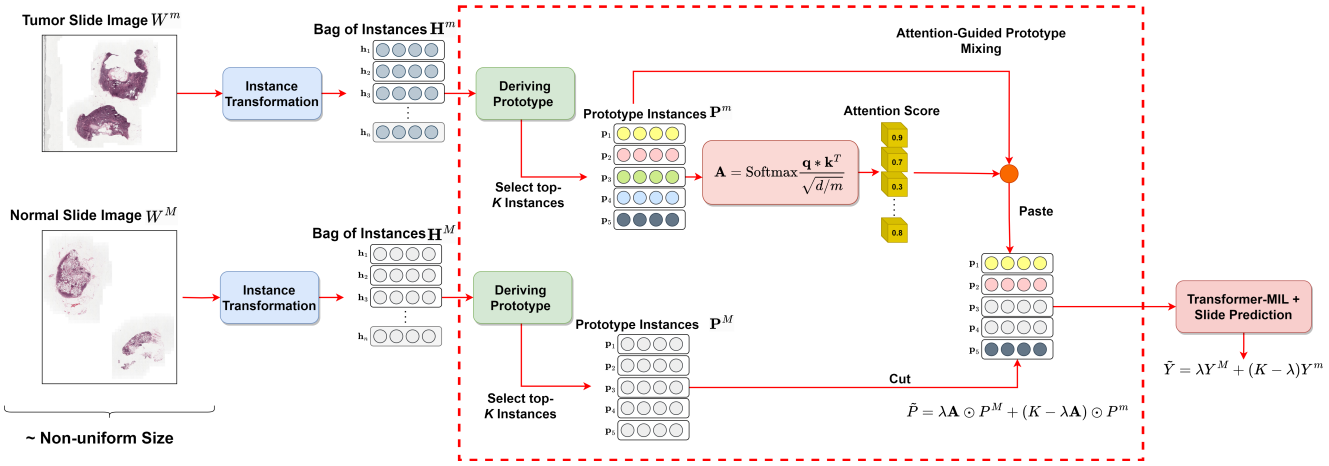

Figure 3. Overview of the proposed method-Attention-Guided Prototype Mixing. A bag of instances from both the tumor and normal classes is fist processed in the prototype instance building stage to obtain prototype instances (P^m and P^M). The P^m are then fed to the Transformer module for getting $\mathbf{A}$, where $\mathbf{A}$ produces scores for prototype instances. Finally, P^M and the most attentive instances of P^m are mixed.

4.2. Deriving prototype instances (P)

After obtaining the equal distribution of imbalanced WSI classes, we divide each slide image W into small patches X and then transform using f_θ (Eq. (1)) into the instance embeddings H, called a bag of instances. To address the challenge of Transformer module in computing the long-sequence instances, we introduce the concept of prototype instances.

Specifically, for each bag, we parameterize each instance $h_n \in H$ with the weight vector $\mathbf{w}_n \in \mathbb{R}^{d\times 1}$, followed by sorting, expressed as:

$$P = \text{sort}\left[\mathbf{w}_1 \cdot h_1, \ldots, \mathbf{w}_n \cdot h_n\right][1:K]. \tag{6}$$

Eq. (6) determines the instances with K-highest scores (*i.e.*, prototype instances), denoted as $P = \{p_1, p_2, p_3, \ldots, p_K\} \in \mathbb{R}^{K\times d}$, where K represents the number of prototype instances. Actually, P acts like a representative feature of WSI (W) but with a smaller and uniform size. In our empirical study, it is found that the optimal value of K is 64.

4.3. Mixing the most attentive instances (M)

Multi-Head Attention (MHA). According to Eq. (2) that represents H as input tokens, which is then aggregated with the class token x_{class}, in our proposed approach, we treat the prototype instances P as input tokens. Thus, the Transformer operates on the embedded sequences as follows:

$$\mathbf{x} = \left[x_{\text{class}}; P\right] \in \mathbb{R}^{(K+1)\times d} \tag{7}$$

where $K+1$ is the length of P plus x_{class} and d is the dimension of each instance.

Given a Transformer-MIL (*e.g.*, TransMIL [28] or GTP [41]) with m heads and input embedded sequences $\mathbf{x}$, we first multiply $\mathbf{x}$ with $\mathbf{w}_q$, $\mathbf{w}_k$, and $\mathbf{w}_v$ for getting queries $\mathbf{q}$, keys $\mathbf{k}$, and values $\mathbf{v}$, respectively. Then, an attention score $\mathbf{A}$ can be obtained as:

$$\begin{aligned} \mathbf{q} &= \mathbf{x} * \mathbf{w}_q \\ \mathbf{k} &= \mathbf{x} * \mathbf{w}_k \\ \mathbf{v} &= \mathbf{x} * \mathbf{w}_v \end{aligned} \tag{8}$$
$$\mathbf{A}(\mathbf{q}, \mathbf{k}) = \text{Softmax}\left(\frac{\mathbf{q}\mathbf{k}^T}{\sqrt{d/m}}\right).$$

The attention score $\mathbf{A} \in [0,1]^K$ is obtained by establishing a mapping between $\mathbf{q}$ and $\mathbf{k}$, and representing the output of the multi-head attention by multiplication with $\mathbf{v}$, *i.e.*, $\mathbf{MHA}(\mathbf{q}, \mathbf{k}, \mathbf{v}) = \mathbf{A}(\mathbf{q}, \mathbf{k}) * \mathbf{v}$. Since we have m heads, we compute the average across all attention heads to derive $\mathbf{A} \in [0,1]^K$. In our empirical study, we obtain $\mathbf{A}$ in Eq. (8) from the last layer of Transformer-MIL module [28] [41].

Due to our method utilizes both TransMIL [28] and GTP [41] as the aggregator models, we modify their architectures for our use, including:

- **TransMIL [28].** TransMIL designed Pyramid Position Encoding Generator (PPEG) to acquire spatial information. Therefore, we project our prototype instances P onto PPEG and subsequently feed them into the Transformer module.
- **GTP [41].** GTP designed Graph Convolution Network (GCN) to learn inter-relationship between instances. Therefore, we project our prototype instances P onto GCN and subsequently feed them into the Transformer module.

Mixing P based on the attention score. We expand upon the approach of mixture sampling from the minority and majority classes in CMO [27]. Once we obtain $\mathbf{A}$, we can address the challenge posed by WSI, which involves limited and randomly distributed tumor instances. To begin with, we sort the prototype instances P of tumor samples based on their corresponding scores in $\mathbf{A} = \{a_1, a_2, a_3, \ldots, a_K\}$, expressed as:

$$P = \begin{bmatrix} p_{\sigma(1)} \\ p_{\sigma(2)} \\ p_{\sigma(3)} \\ \vdots \\ p_{\sigma(K)} \end{bmatrix} \in \mathbb{R}^{K \times d} \tag{9}$$

where σ represents a permutation that rearranges the indices of the elements in P such that the corresponding scores in $\mathbf{A}$ are in a non-decreasing order and $p_{\sigma(i)}$ denotes the i-th element of P after sorting.

According to Eq. (3) and (4), we formulate the mixing operation of minority and majority samples at WSI-level. Let P and Y represent prototype instances and its slide-label, respectively. We propose generating a new prototype $(\tilde{P}, \tilde{Y})$ by combining two prototypes, (P^M, Y^M) and (P^m, Y^m), where P^M is sampled from the normal class and P^m comes from the tumor class. Unlike the previous method in Eq. (3) and (4), $\mathbf{M}$ is replaced with $\mathbf{A}$ as the guidance for mixing, expressed as:

$$\tilde{P} = \lambda \mathbf{A} \odot P^M + (K - \lambda \mathbf{A}) \odot P^m \tag{10}$$

$$\tilde{Y} = \lambda Y^M + (K - \lambda) Y^m \tag{11}$$

where $(K - \lambda \mathbf{A})$ represents a mask that indicates which instances can be selected from P^m and pasted onto P^M. The combination ratio λ between these two prototypes is sampled from $0 \leq \lambda \leq K$.

Mixing Consideration. Due to the dependency of our mixing scenario on classifier construction, we need to carefully determine the appropriate starting point during training. Directly applying our Attention-Guided Mixing at the beginning may result in unstable training. This instability arises from the fact that both the TransMIL [28] and GTP [41] models have not yet attained stable performance. However, after several epoches, we have observed that the models achieve stable performance. In our empirical study of employing TransMIL [28] and GTP [41], we have observed the stable performance after around ten epoches. Therefore, we then initiate our approach after ten epochs of the training process.

5. Experimental Settings

5.1. Datasets

We introduce the WSI datasets for our experiments. Each WSI was divided into 224×224 patches at $20x$ magnification.

Camelyon16 [2] is a public dataset proposed for metastasis detection in breast cancer. The dataset contains 270 training slides (111 tumor and 159 normal images) and 129 testing slides (49 tumor and 80 normal images). Tumor slides in this dataset contain small portions of tumor regions ($< 20\%$ per slide). The original Camelyon16 is not so imbalanced, so we modify it to create two imbalanced datasets, *i.e.*, Medium-Imbalanced and High-Imbalanced datasets, for our study. Specifically, the Medium-imbalanced dataset contains 187 training slides (28 tumor and 159 normal images) while the High-Imbalanced dataset contains 172 training slides (13 tumor and 159 normal images).

HistoQC is an in-house dataset designed for the computer-aided histopathology research. The pathologist experts annotated 448 WSIs (335 in the training set and 113 in the testing set) from The Cancer Genome Atlas Program (TCGA) lung cancer dataset. However, HistoQC has only 23 tumor slides (17 for the training set and 6 for the testing set), posing an imbalanced dataset problem. Due to the limited number of tumor slides compared to normal slides, it is categorized as High-Imbalanced.

5.2. Settings

We utilized the ResNet18 configuration as a feature extractor, denoted as f_θ (Eq. 1), which was obtained through training SimCLR [8, 21]. This configuration allows us to obtain a bag of instances for each WSI. We defined the number of epoches as 200, the learning rate as $2e-4$, and the weight decay as $1e-5$. We selected $K = 64$ instances as the prototype P emprically. In the Transformer module, the number m of attention heads was set to 8, the dimension d of each instance was 512, and dropout rate was 0.1.

During the inference step, we employed the sigmoid function to normalize the predicted diagnosis scores. Our experiments were conducted using an NVIDIA 3060Ti GPU with 12GiB RAM.

Evaluation Metrics. The performances are mainly reported as balanced accuracy (Acc) [4], Area Under Curve (AUC) score [16], and Precision-Recall Area Under Curve (PR AUC) score [3].

Comparisons. We conducted a comprehensive comparison between our approach and state-of-the-art (SOTA) methods: 1) No Oversampling, 2) Random Oversampling (ROS) [29], 3) Balanced-MixUp [15], 4) CMO [27], and 5) ReMix [37]. However, the SOTA methods were designed to mix the samples of uniform sizes. Thus, we adopted zero padding on the smaller-sized image to ensure that all WSIs have the same size.

In addition, we investigated the loss function strategies, including 1) Binary Cross Entropy (BCE) and 2) Balanced-Binary Cross Entropy (Bal-BCE) [36], to address the imbalanced class problem.

Table 1. **Comparisons with State-of-the-art Methods on Imbalanced Camelyon16** in (%).

	Vanilla-Balanced			Medium-Imbalanced			High-Imbalanced		
	Acc	AUC	PR AUC	Acc	AUC	PR AUC	Acc	AUC	PR AUC
TransMIL [28]	84.49	84.80	84.30	77.51	80.21	79.00	73.64	74.50	73.00
+ Bal-BCE [36]	84.49	84.80	84.30	82.17	84.00	83.70	79.40	80.01	79.50
+ ROS [29]	84.49	85.00	84.52	82.17	84.00	83.70	81.39	83.50	82.16
+ Balanced-MixUp [15]	87.59	89.35	88.21	86.04	87.80	87.00	84.96	86.26	85.24
+ CMO [27]	88.37	90.00	89.35	86.81	88.21	87.20	85.27	86.90	85.52
+ **Ours+BCE**	**89.92**	**93.00**	**92.50**	**89.14**	**92.00**	**92.60**	**89.14**	**92.00**	**92.60**
+ **Ours+Bal-BCE**	**90.69**	**92.75**	**93.00**	**90.69**	**92.75**	**93.00**	**89.92**	**93.00**	**92.50**
GTP [41]	84.49	85.00	84.40	75.96	76.50	76.00	72.09	76.00	73.43
+ Bal-BCE [36]	85.27	86.80	85.00	81.39	84.35	83.25	78.29	79.90	79.35
+ ROS [29]	85.27	86.90	85.52	83.72	84.50	83.70	81.39	83.50	82.16
+ Balanced-MixUp [15]	88.37	90.00	89.35	87.59	89.35	88.21	85.27	86.90	85.52
+ CMO [27]	88.37	90.00	89.35	88.37	90.00	89.35	86.81	88.21	87.20
+ **Ours+BCE**	**90.69**	**92.75**	**93.00**	**89.92**	**93.00**	**92.50**	**89.92**	**93.00**	**92.50**
+ **Ours+Bal-BCE**	**91.47**	**93.77**	**92.95**	**90.69**	**92.75**	**93.00**	**90.69**	**92.75**	**93.00**

Table 2. **Comparison with State-of-the-art WSI-Augmentation** on Vanilla Balanced Dataset.

	Acc (%)		
	Vanilla	ReMix [37]	Ours
TransMIL [28]			
+BCE	89.92	90.69$_{\uparrow 0.77}$	**91.47**$_{\uparrow 1.55}$
+Bal-BCE [36]	90.69	90.69$_{\uparrow 0.00}$	**91.47**$_{\uparrow 0.78}$
GTP [41]			
+BCE	90.69	90.69$_{\uparrow 0.00}$	**91.47**$_{\uparrow 0.78}$
+Bal-BCE [36]	91.47	91.47$_{\uparrow 0.00}$	**92.24**$_{\uparrow 0.77}$

6. Experimental Results

We provide experimental results and comparisons with SOTA merhods based on the public Camelyon16 dataset in Sec. 6.1 and in-House HistoQC dataset in Sec. 6.2.

We integrated the proposed Attention-Guided Prototype Mixing with two Transformer-MIL models, including TransMIL [28] and GTP [41], and compared the performance with the state-of-the-art (SOTA) methods.

6.1. Camelyon16

Comparison with SOTA Oversampling.

We trained our proposed model on Camelyon16 using different imbalance ratios: Vanilla Balanced (111 tumors and 159 normal slide images), Medium-Imbalanced (28 tumors and 159 normal slide images), and High-Imbalanced (13 tumors and 159 normal slide images). Then, we evaluated on testing slides (49 tumor and 80 normal images), as shown in Table 1.

We can observe from Table 1 that both our proposed method and the SOTA methods achieve significant performance on the datasets with a Vanila-balanced ratio. However, under Camelyon16 with a higher imbalance ratio, the SOTA methods fail to maintain the performance. For instance, Balanced-MixUp [15] and CMO [27] experience reductions in accuracy of up to 4% and 3% at High-Imbalanced, respectively. By contrast, our method demonstrate the ability to maintain performance in the presence of high imbalances with only a 1% reduction in accuracy. Overall, our method outperforms SOTA methods with a significant performance gap.

The main reason is that ROS [29] solely duplicates tumor bags while CMO [27] and Balanced-MixUp [15] indiscriminately mix bags without considering the specific instances involved. By contrast, our method introduces a novel strategy that significantly enhances the mixing performance by intelligently determining which instances can be cut and pasted within other bags. This strategy overcomes the limitations of previous methods and enables more accurate discrimination between normal and tumor slides.

Comparison with SOTA WSI-Augmentation. We conducted a comparative analysis between our proposed method and ReMix [37] for augmenting Vanilla Camelyon16 (159 normal and 111 tumor slide images), as shown in Table 2. The results demonstrate that our method improves ReMix with an enhanced accuracy of up to $0.7\% \sim 1.55\%$.

We hypothesize that in contrast to ReMix that employs simple interpolation-based mixing without considering specific instances, our method incorporates mixing while retaining the original features. This observation highlights the importance of carefully selecting specific instances to preserve and improve performance.

6.2. HistoQC

Comparison with SOTA Oversampling. We conducted experiments on the in-house HistoQC dataset. Unlike Camelyon16, the HistoQC dataset exhibits highly imbalance not only in the training slides but also in the testing slides, which consists of 6 tumor images and 107 normal images.

As shown in Table 3, our method demonstrates a significant performance boost in this High-Imbalanced testing, achieving a stable average evaluation accuracy of up to 93%. By contrast, SOTA methods such as Balanced-MixUp [15] and CMO [27] achieve accuracy lower than 86%.

In the subjective evaluation, it is found that our method can effectively handle the challenge of diverse minority contexts even within the highly imbalanced categories, thanks to attention-guided prototype mixing. Furthermore, the advantage of the Transformer-MIL approach is its ability in capturing the pairwise correlation between prototype instances within each WSI (bag).

Table 3. **Comparisons with State-of-the-art Methods on Imbalanced HistoQC** in (%).

	High Imbalanced		
	Acc	AUC	PR AUC
TransMIL [28]	76.10	77.05	76.80
+ Bal-BCE [36]	78.76	79.90	79.65
+ ROS [29]	79.64	80.05	79.80
+ Balanced-MixUp [15]	83.18	83.90	83.65
+ CMO [27]	84.95	85.75	85.20
+ **Ours+BCE**	**91.15**	**93.00**	**92.30**
+ **Ours+Bal-BCE**	**92.03**	**93.20**	**92.85**
GTP [41]	78.76	79.90	79.65
+ Bal-BCE [36]	79.64	80.05	80.00
+ ROS [29]	83.18	84.00	83.50
+ Balanced-MixUp [15]	84.07	85.00	84.50
+ CMO [27]	85.84	86.20	86.00
+ **Ours+BCE**	**92.03**	**93.20**	**92.85**
+ **Ours+Bal-BCE**	**92.92**	**94.00**	**93.20**

7. Ablation Study

We evaluated different numbers of prototype instances P as input sequences. In this experiment, we used Vanilla Camelyon16, as shown in Table 4. These results indicate the significant impact of deriving prototype instances to preserve the intrinsic features of long-sequence whole-slide images (WSIs) while being applicable for the Transformer-MIL framework. For instance, TransMIL [28] and GTP [41] achieve accuracies of only 87.59% and 89.14%, respectively, with high computational complexity measured in FLOPs. In our study, we observed similar performance between $P = 64$ and $P = 256$, but with different FLOPs. Consequently, we selected $P = 64$ with lower FLOPs while still maintaining excellent performance.

Table 4. Comparison with different P on Vanilla Camelyon16. The $*$ denote the results are reproduced by our experiment. M means megaFLOPS.

Method	Acc ↑	AUC ↑	PR AUC ↑	FLOPs ↓
TransMIL [28] $*$	87.59	89.00	91.20	120.050M
$+_{P=1024}$	89.14	92.00	92.60	99.031M
$+_{P=256}$	89.92	93.00	92.50	32.577M
$+_{P=64}$	**89.92**	**93.00**	**92.50**	14.391M
$+_{P=16}$	73.64	74.50	73.00	**9.204M**
GTP [41] $*$	89.14	92.00	92.60	70.248M
$+_{P=1024}$	89.92	93.00	92.50	50.610M
$+_{P=256}$	90.69	92.75	93.00	22.129M
$+_{P=64}$	**90.69**	**92.75**	**93.00**	10.091M
$+_{P=16}$	77.51	80.21	79.00	**7.034M**

8. Limitation

In our study, the process for selecting instances is based on attention guidance, which means that this mixing framework is dependent on classifier construction. If the classifier model fails to achieve convergence in classifying a specific medical dataset, the mixing framework may not perform well. Future work should focus on designing improvements to the mixing framework in an independent way, without depending on classifier construction.

9. Conclusion

This paper introduces Attention-Guided Prototype Mixing, a novel approach for learning imbalanced whole slide images (WSIs) classification. Our approach focuses on diversifying minority context within WSIs, even in the presence of an imbalanced class distribution. Through extensive experiments, we demonstrate the effectiveness of our proposed method compared to state-of-the-art (SOTA) oversampling methods. We hope that our study can serve as a strong baseline for designing oversampling and augmentation techniques to address imbalanced medical datasets.

10. Acknowledgement

This work was supported by the National Science and Technology Council (NSTC), Taiwan, ROC, under Grants NSTC 111-2634-F-006-022 and MOST 110-2221-E-001–020-MY2. We also thank Taiwan Cloud Computing (TWCC) for providing computational and storage resources.

References

[1] V. Baxi, R. Edwards, M. Montalto, and S. Saha. Digital pathology and artificial intelligence in translational medicine and clinical practice. *Modern Pathology*, 35(1):23–32, 2022.

[2] B. E. Bejnordi, M. Veta, P. J. Van Diest, B. Van Ginneken, N. Karssemeijer, G. Litjens, J. A. Van Der Laak, M. Hermsen, Q. F. Manson, M. Balkenhol, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *Jama*, 318(22):2199–2210, 2017.

[3] K. Boyd, K. H. Eng, and C. D. Page. Area under the precision-recall curve: point estimates and confidence intervals. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2013, Prague, Czech Republic, September 23-27, 2013, Proceedings, Part III 13*, pages 451–466. Springer, 2013.

[4] K. H. Brodersen, C. S. Ong, K. E. Stephan, and J. M. Buhmann. The balanced accuracy and its posterior distribution. In *2010 20th international conference on pattern recognition*, pages 3121–3124. IEEE, 2010.

[5] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[6] C.-F. R. Chen, Q. Fan, and R. Panda. Crossvit: Cross-attention multi-scale vision transformer for image classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 357–366, 2021.

[7] J.-N. Chen, S. Sun, J. He, P. H. Torr, A. Yuille, and S. Bai. Transmix: Attend to mix for vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12135–12144, 2022.

[8] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, pages 1597–1607. PMLR, 2020.

[9] Y.-C. Chen and C.-S. Lu. 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*, pages 23936–23945, 2023.

[10] C. Cong, Y. Yang, S. Liu, M. Pagnucco, and Y. Song. Imbalanced histopathology image classification using deep feature graph attention network. In *IEEE 19th International Symposium on Biomedical Imaging (ISBI)*, pages 1–4. IEEE, 2022.

[11] D. Dablain, B. Krawczyk, and N. V. Chawla. Deepsmote: Fusing deep learning and smote for imbalanced data. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[12] S. Deepak and P. Ameer. Brain tumor categorization from imbalanced mri dataset using weighted loss and deep feature fusion. *Neurocomputing*, 520:94–102, 2023.

[13] A. Dosovitskiy et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.

[14] S. Fotouhi, S. Asadi, and M. W. Kattan. A comprehensive data level analysis for cancer diagnosis on imbalanced data. *Journal of biomedical informatics*, 90, 2019.

[15] A. Galdran, G. Carneiro, and M. A. González Ballester. Balanced-mixup for highly imbalanced medical image classification. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pages 323–333. Springer, 2021.

[16] J. A. Hanley and B. J. McNeil. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143(1):29–36, 1982.

[17] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[18] M. Ilse, J. Tomczak, and M. Welling. Attention-based deep multiple instance learning. In *ICML*, pages 2127–2136. PMLR, 2018.

[19] J. Jessup, R. Krueger, S. Warchol, J. Hoffer, J. Muhlich, C. C. Ritch, G. Gaglia, S. Coy, Y.-A. Chen, J.-R. Lin, et al. Scope2screen: Focus+ context techniques for pathology tumor assessment in multivariate image data. *IEEE transactions on visualization and computer graphics*, 28(1):259–269, 2021.

[20] N. Kumar, R. Gupta, and S. Gupta. Whole slide imaging (wsi) in pathology: current perspectives and future directions. *Journal of Digital Imaging*, 33(4):1034–1040, 2020.

[21] B. Li, Y. Li, and K. W. Eliceiri. Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In *CVPR*, pages 14318–14328, 2021.

[22] K. Liu, W. Zhu, Y. Shen, S. Liu, N. Razavian, K. J. Geras, and C. Fernandez-Granda. Multiple instance learning via iterative self-paced supervised contrastive learning. In *Proceedings of the IEEE/CVF Conference*

on Computer Vision and Pattern Recognition, pages 3355–3365, 2023.

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

[24] O. Maron and T. Lozano-Pérez. A framework for multiple-instance learning. *Advances in neural information processing systems*, 10, 1997.

[25] L. McInnes, J. Healy, and J. Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

[26] S. S. Mullick, S. Datta, and S. Das. Generative adversarial minority oversampling. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1695–1704, 2019.

[27] S. Park, Y. Hong, B. Heo, S. Yun, and J. Y. Choi. The majority can help the minority: Context-rich minority oversampling for long-tailed classification. In *CVPR*, pages 6887–6896, 2022.

[28] Z. Shao, H. Bian, Y. Chen, Y. Wang, J. Zhang, X. 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.

[29] J. Van Hulse, T. M. Khoshgoftaar, and A. Napolitano. Experimental perspectives on learning from imbalanced data. In *ICML*, pages 935–942, 2007.

[30] Q. D. Vu, K. Rajpoot, S. E. A. Raza, and N. Rajpoot. Handcrafted histological transformer (h2t): Unsupervised representation of whole slide images. *Medical Image Analysis*, page 102743, 2023.

[31] Q. D. Vu, K. Rajpoot, S. E. A. Raza, and N. Rajpoot. Handcrafted histological transformer (h2t): Unsupervised representation of whole slide images. *Medical Image Analysis*, page 102743, 2023.

[32] Q. Wang, X. Zhou, C. Wang, Z. Liu, J. Huang, Y. Zhou, C. Li, H. Zhuang, and J.-Z. Cheng. Wgan-based synthetic minority over-sampling technique: improving semantic fine-grained classification for lung nodules in ct images. *IEEE Access*, 7:18450–18463, 2019.

[33] X. Wang, S. Yang, J. Zhang, M. Wang, J. Zhang, W. Yang, J. Huang, and X. Han. Transformer-based unsupervised contrastive learning for histopathological image classification. *Medical Image Analysis*, 81:102559, 2022.

[34] Z. Xia, X. Pan, S. Song, L. E. Li, and G. Huang. Vision transformer with deformable attention. In *CVPR*, pages 4794–4803, 2022.

[35] Y. Xiong, Z. Zeng, R. Chakraborty, M. Tan, G. Fung, Y. Li, and V. Singh. Nyströmformer: A nyström-based algorithm for approximating self-attention. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14138–14148, 2021.

[36] Z. Xu, R. Liu, S. Yang, Z. Chai, and C. Yuan. Learning imbalanced data with vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15793–15803, 2023.

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

[38] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, and Y. Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *CVPR*, pages 6023–6032, 2019.

[39] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

[40] H. Zhang, Y. Meng, Y. Zhao, Y. Qiao, X. Yang, S. E. Coupland, and Y. Zheng. Dtfd-mil: Double-tier feature distillation multiple instance learning for histopathology whole slide image classification. In *CVPR*, pages 18802–18812, 2022.

[41] Y. Zheng, R. H. Gindra, E. J. Green, E. J. Burks, M. Betke, J. E. Beane, and V. B. Kolachalama. A graph-transformer for whole slide image classification. *IEEE transactions on medical imaging*, 41(11):3003–3015, 2022.

[42] Y. Zheng, J. Li, J. Shi, F. Xie, J. Huai, M. Cao, and Z. Jiang. Kernel attention transformer for histopathology whole slide image analysis and assistant cancer diagnosis. *IEEE Transactions on Medical Imaging*, 2023.

[43] Z. Zhu, L. Yu, W. Wu, R. Yu, D. Zhang, and L. Wang. Murcl: Multi-instance reinforcement contrastive learning for whole slide image classification. *IEEE Transactions on Medical Imaging*, 2022.

[44] A. Zuraw and F. Aeffner. Whole-slide imaging, tissue image analysis, and artificial intelligence in veterinary pathology: An updated introduction and review. *Veterinary Pathology*, 59(1):6–25, 2022.