Boosting Whole Slide Image Classification from the Perspectives of Distribution, Correlation and Magnification

Linhao Qu[1,2*] Zhiwei Yang[1,2*] Minghong Duan[1,2]
Yingfan Ma[1,2] Shuo Wang[1,2] Manning Wang[1,2†] Zhijian Song[1,2†]
[1]Digital Medical Research Center, School of Basic Medical Science, Fudan University
[2]Shanghai Key Lab of Medical Image Computing and Computer Assisted Intervention

Abstract

Bag-based multiple instance learning (MIL) methods have become the mainstream for Whole Slide Image (WSI) classification. However, there are still three important issues that have not been fully addressed: (1) positive bags with a low positive instance ratio are prone to the influence of a large number of negative instances; (2) the correlation between local and global features of pathology images has not been fully modeled; and (3) there is a lack of effective information interaction between different magnifications. In this paper, we propose MILBooster, a powerful dual-scale multi-stage MIL framework to address these issues from the perspectives of distribution, correlation, and magnification. Specifically, to address issue (1), we propose a plug-and-play bag filter that effectively increases the positive instance ratio of positive bags. For issue (2), we propose a novel window-based Transformer architecture called PiceBlock to model the correlation between local and global features of pathology images. For issue (3), we propose a dual-branch architecture to process different magnifications and design an information interaction module called Scale Mixer for efficient information interaction between them. We conducted extensive experiments on four clinical WSI classification tasks using three datasets. MILBooster achieved new state-of-the-art performance on all these tasks. Codes will be available at https://github.com/miccaiif/MILBooster.

1. Introduction

Computer-aided diagnosis based on pathology Whole Slide Images (WSIs) has important clinical significance, but using deep-learning techniques in WSI classification faces great challenges [29, 38, 23, 26, 21]. On one hand, WSIs are very large, typically being muti-gigapixel images, so they must be divided into many small patches to be processed by deep neural networks. On the other hand, although the label of the whole slide can often be obtained, obtaining fine-grained annotations for the huge number of small patches is extremely expensive, which makes supervised learning at the patch level infeasible. Therefore, WSI classification is usually formulated as a Multiple Instance Learning (MIL) problem [34, 9, 6, 22, 24], in which each slide is treated as a bag and the patches cut from the slide are treated as instances of the bag. In the MIL formulation, all instances in a negative bag are negative, while at least one instance in a positive bag is positive. The goal of WSI classification is to accurately classify each bag, while the label of each instance is unavailable.

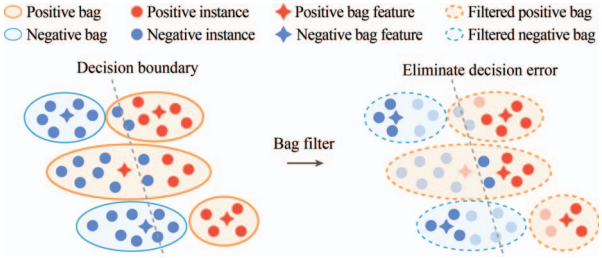

Figure 1. Motivation of our proposed bag filter. Left figure: the positive instance ratio of the middle positive bag is low so its bag feature is biased towards negative instances and it is mis-classified. Right figure: after filtering a large number of negative instances (and possibly some positive instances), the bag feature of the middle positive bag is rectified, and it is correctly classified.

Deep learning-based MIL methods can generally be divided into instance-based methods [3, 10, 30, 15, 41] and bag-based methods [14, 13, 45, 43, 36, 17, 44, 35, 18, 7, 24, 19]. Instance-based methods typically assign a pseudo-label to each instance, and then train an instance classifier to predict the positive score of instances. Finally, bag classification is achieved by aggregating the positive scores of all instances in a bag. Since the true label of each instance is unknown, the pseudo-labels used in this method often contain a large amount of noise, which limits their performance

*Co-first author. †Co-corresponding author.

[30, 17, 32, 25, 31]. Bag-based methods first extract features for each instance, and then use aggregation functions to aggregate the instance features to obtain the bag feature, which is used to train a bag classifier based on the true label of the bag. By aggregating features at the bag level, bag-based methods avoid the problem of noisy instance labels, and can often achieve higher classification performance. However, there are still three important problems that have not been fully addressed. *(1) The aggregated bag feature tends to be biased towards the negative side when the positive instance ratio is low in a positive bag, which increases the risk of decision error.* As shown in Figure 1, intuitively, when the positive instance ratio in a positive bag is low, the aggregated bag feature is easily affected by the massive negative instances, which increases the difficulty of training the bag classifier and inferencing on test bags. If a large number of negative instances can be removed from positive bags, both training and inferencing will become easier. *(2) The correlation between local and global features has not been fully modeled.* As shown in Figure 2 (A), classical attention-based methods [14, 36] assume that each instance in a bag is independent, hence lacking correlation modeling among instances [35]. Recent studies [35, 18, 7] have used Vision Transformer (ViT) frameworks [11] to model the relationship between instances, but they neglect the modeling of local context. In pathology images, tumors are often distributed in a continuous region with an area much larger than the patch size, so proper modeling of regional context of patches is essential. *(3) There is a lack of effective information interaction between different magnification levels.* In clinical practice, pathologists often zoom in and out on WSIs to make a diagnosis of tumors [1, 12, 39], which indicates the need of examining WSIs at different magnification levels. However, most existing studies only work on a single level. As shown in Figure 2 (B), several recent studies have proposed methods such as feature concatenation [17] and weighted fusion [13] for combining multi-magnification information, but they are too simple to effectively fuse information of different magnifications.

To tackle the aforementioned problems, we propose MILBooster, a powerful dual-scale, multi-stage MIL framework that significantly enhances the performance of WSI classification from three aspects: distribution, correlation, and magnification. *In MILBooster, we propose bag filter to effectively address problem (1) from the perspective of feature distribution modeling.* Specifically, the bag filter is a plug-and-play pre-processing module that is applied to both training and test sets to effectively increase the positive instance ratio in positive bags and eliminate the negative impact of a large number of negative instances on feature aggregation. *For problem (2), we propose a window-based PiceBlock (Part Interact with Entirety Block) to model the correlation between local and global features in pathology images.* In PiceBlock, we arrange the features according to the position of their corresponding patches and divide the feature map into several windows. We design a Self-window Transformer Block (SWTB) to model the correlation of instance features within each window, a Self-Window Merging Block (SWMB) to merge the instance features within each window into a region feature, and a Cross-window Transformer Block (CWTB) to model the correlation between each instance and each region feature. In this way, PiceBlock achieves efficient modeling of the correlation between local features and global features. *For problem (3), we propose a dual-scale interaction module called Scale Mixer to achieve efficient information interaction between different magnifications.* For the two scales (high scale and low scale) being used, the Scale Mixer takes the feature map of either the high or the low scale as the main information and partitions the feature map of the other scale into windows for interaction with the main information using SWMB and CWTB. The output feature maps of SWMB and CWTB are then fused with the main information flow for enhancement.

We conducted extensive single-scale and dual-scale experiments on four clinical tasks across three datasets containing different types of cancer. The results showed that MILBooster can significantly improve the performance of WSI classification and achieve new SOTA.

2. Related Work

2.1. Instance-based MIL Methods

Instance-based methods typically assign a pseudo-label to each instance, and then train an instance classifier to predict a positive score for each instance. Finally, the positive scores of all instances in a bag are aggregated to predict the bag label. Early methods [42, 40, 16, 28] directly assign bag labels to each instance of it as their pseudo instance labels, and some recent methods select a set of key instances to assign pseudo labels for model training [3, 10, 30]. Since the true labels of instances are unknown, the pseudo-labels in these methods typically contain a lot of noise, which limits the performance of the trained instance classifier.

2.2. Bag-based MIL Methods

Bag-based methods are currently the mainstream for WSI classification. In these methods, features are first extracted for each instance, and then an aggregation function is used to aggregate the features of each instance to obtain the bag feature, which is used to train a bag classifier. Attention-based aggregation methods [14, 13, 45, 43, 36, 17, 44, 24] have achieved good performance, but they assume that each instance in the bag is independent, which lacks correlation modeling between instances [35]. Recently, some studies [35, 18, 7] have used ViT-based archi-

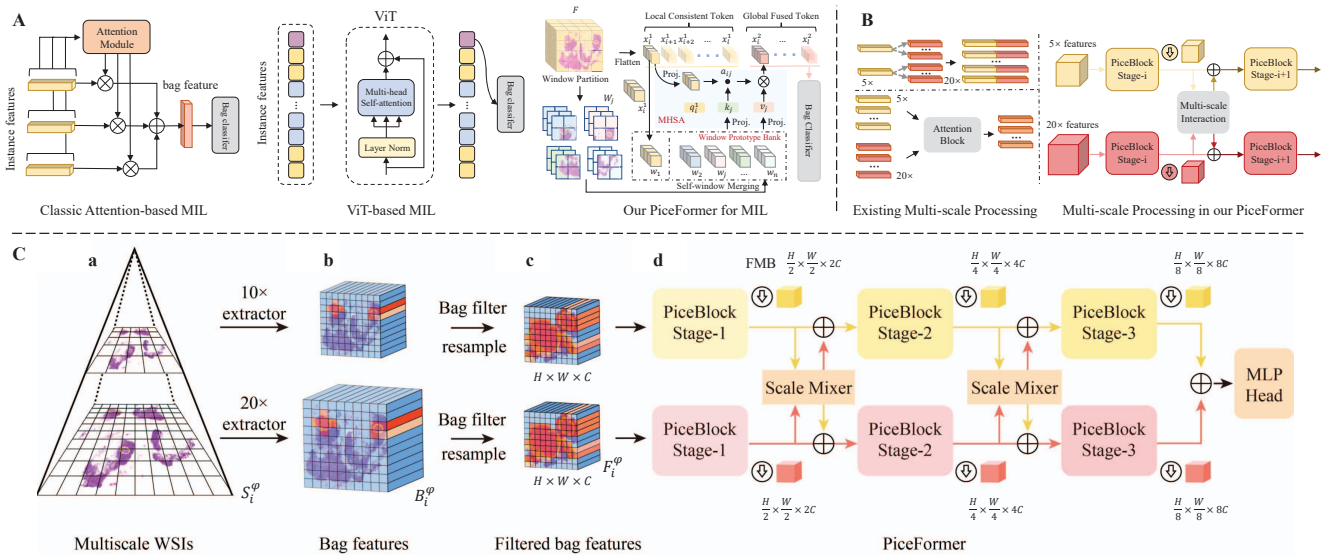

Figure 2. (A) Motivation of our proposed PiceBlock. (B) Motivation of our proposed dual-branch interaction module. (C) Workflow of our MILBooster (taking two branches of 10× and 20× as an example), which consists of two plug-and-play bag filters and a dual-scale, multi-stage bag classifier PiceFormer.

tectures [11] to model the relationship between instances, but they ignore the modeling of local regions and introduce a huge computational cost. In contrast, we propose a window-based PiceBlock to model the correlation between local and global features, which greatly reduces the computational cost of information interaction. In Zhang et al. [44], random splitting of bags is used to construct many small pseudo-bags to increase the number of bags and suppress overfitting. Although random splitting increases the number of bags, it does not effectively increase the overall positive ratio. On the contrary, when the overall positive ratio of the dataset is small, the split small bags may also lead to pseudo-label errors.

2.3. Feature Modeling and Multiscale Processing in WSI Classification

Due to the large number of patches extracted from pathology images, many studies [17, 44, 35, 32, 4, 5] have used pre-trained networks to extract patch features and perform subsequent processing based on these features. Following these studies, we also use a self-supervised method to pre-train an instance feature extractor in this study.

Most current WSI classification methods work on a single magnification level, while some studies [13, 17, 27, 33] have shown that combining features from multiple magnification levels (often two scales) improves the classification performance. However, with simple information fusion strategies such as concatenation [30] and weighted fusion [13], these methods cannot effectively facilitate information interaction between features from different magnifications.

3. Method

3.1. Problem Formulation

Given a dataset $S = \{S_1^\varphi, S_2^\varphi, \ldots, S_{N_\varphi}^\varphi\}$ of N_φ WSIs at a magnification (also called scale) φ ($\varphi = \{5,\ 10,\ 20\}$ in this paper), we cut each WSI S_i^φ into non-overlapping small patches $\{p_{i,j}^\varphi, j = 1, 2, ..., n_i^\varphi\}$, where n_i^φ is the number of patches cut out of S_i^φ. All patches $p_{i,j}^\varphi$ from S_i^φ form a bag, where each patch is an instance of this bag, and the label of this bag is $Y_i^\varphi \in \{0, 1\}$. In the MIL setting, only the labels of the bags from the training set are available, while the labels of the instances are unknown.

Following DSMIL [17], we first utilize the self-supervised method SimCLR [8] to pre-train a feature extractor for all patches $p_{i,j}^\varphi$ at each scale, and then use this feature extractor to map $p_{i,j}^\varphi$ to feature vectors $f_{i,j}^\varphi$. To avoid losing the spatial information of the original slide, we re-arrange and edge-pad $f_{i,j}^\varphi$ according to their positions in the spatial dimension, resulting in a new vector matrix B_i^φ. The subsequent processing is performed on the basis of B_i^φ. Since our method is based on the pre-extracted instance features, we will not differentiate between an instance and its feature hereafter.

3.2. Framework Overview

As shown in Figure 2 (C), MILBooster is a dual-branch MIL framework, and each branch consists of a bag filter and a multi-stage Transformer architecture PiceFormer. Each branch mainly processes WSIs at a certain scale and information exchange between different scales happens in PiceFormer via the Scale Mixer module. In each branch, we first

feed the pre-extracted features B_i^{φ} into the bag filter for pre-processing to obtain the processed features F_i^{φ}. The bag filter detects and filters out a certain percentage of instances in B_i^{φ} with features close to that of negative instances so as to increase the positive instance ratio in positive bags and alleviate the negative impact of a large number of negative instances in positive bags on feature aggregation. Then we feed the filtered features F_i^{φ} of two scales together into PiceFormer. Overall, PiceFormer is a dual-branch bag classifier, consisting of multiple sequentially connected feature extracting modules called PiceBlock in each branch and a dual-branch interaction module called Scale Mixer. Finally, we add the output features of the two branches and input the results to the MLP Head to complete the classification.

Detailed description of Bag Filter is in Section 3.3. The architecture of PiceFormer and its key components are introduced in detail in Section 3.4. For simplification, we will omit the superscript φ when separately describing each branch.

3.3. Bag Filter

The bag filter is implemented on the base of distribution modeling of true negative instances in negative bags. Specifically, we first extract the features of all true negative instances from negative bags in the training set as the true negative feature bank. Then, we use the K-means algorithm to cluster the feature bank into γ clusters, with each cluster denoted as C_γ. On the base of these clusters, we define a positive score $t_{i,j}$ for each instance $f_{i,j}$ as follows,

$$t_{i,j} = \min_{\gamma} D\left(f_{i,j}, C_\gamma\right) = \min_{\gamma} \left(f_{i,j} - \mu_\gamma\right)^T \Sigma_\gamma^{-1} \left(f_{i,j} - \mu_\gamma\right), \tag{1}$$

where $D(\cdot)$ denotes the Mahalanobis distance metric, and μ_γ and Σ_γ are the mean and covariance of cluster C_γ. A low positive score indicates that the instance is close to a cluster in the negative feature bank and it has a low probability of being positive, and vice versa.

The positive score defined in formula 1 is used to filter instances in both training and testing bags. For positive bags in the training set, we filter out $\varepsilon_p\%$ instances with the lowest positive scores, in order to increase the positive instance ratio. For negative bags in the training set, we filter out $\varepsilon_n\%$ instances with the highest positive scores to make the decision boundary clearer. As for the testing sets, the true labels of the bags are unknown, so we filter out $\varepsilon_t\%$ instances with small positive scores for all bags so as to increase the positive instance ratio in positive bags.

To facilitate subsequent window partitioning operations, after the bag filtering, we re-arrange and edge-pad the feature maps based on their relative spatial positions, making them into a regular matrix of size $H \times W \times C$. In addition, the number of instances in high-scale bags is generally greater than that in low-scale bags. In order to effectively balance the number of instances in different scales and facilitate dual-scale information interaction, we perform sequential repeating and edge padding to the low-scale bags and make the filtered bag tensors of different scales have the same shape.

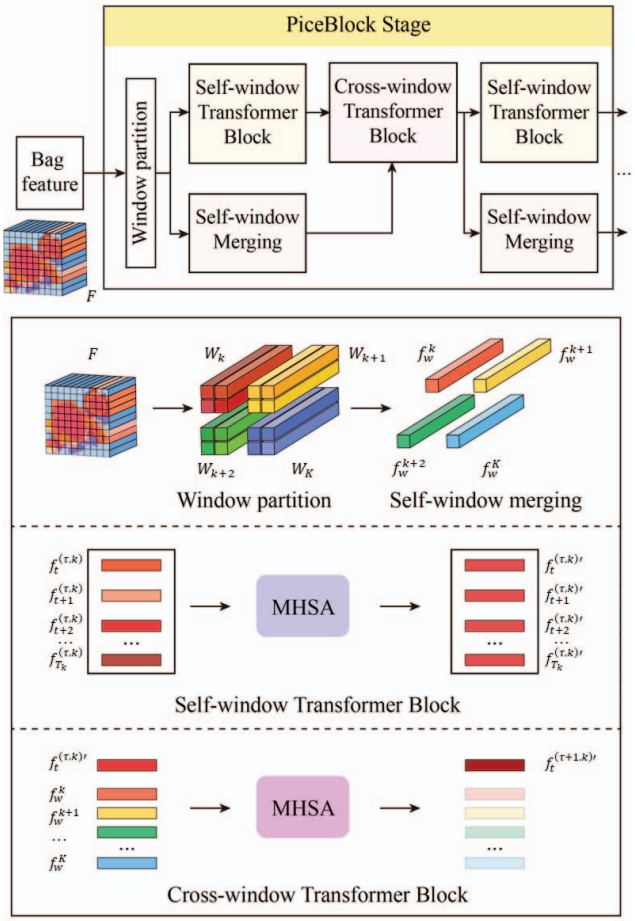

Figure 3. Workflow of PiceBlock, where MHSA represents the Multi-Head Self-Attention mechanism.

3.4. PiceFormer

As shown in Figure 2 (C), PiceFormer is a dual-branch bag classifier, in which we use multiple sequentially connected PiceBlocks as multiple Stages for better feature extraction. After each Stage, we use a Focal Merging Block (FMB) to downsample the feature maps to gradually increase the receptive field of PiceFormer. Finally, we add the features from the two branches and input them into an MLP Head for classification. The following sections will describe each module in detail.

3.4.1 PiceBlock

We propose a window-based PiceBlock to model the correlation between local and global features in pathology im-

ages. The structure of the PiceBlock is shown in Figure 3. For the input feature map F, we first partition it into a set of windows $\{W_k, k = 1, 2, ..., K\}$ of size $P_H \times P_W$, and $W_k \in \mathbb{R}^{P_H \times P_W \times C_w}$, where C_w represents the feature dimension. Then, we input the partitioned feature maps into a Self-Window Transformer Block (SWTB) and a Self-Window Merging Block (SWMB), respectively.

Since instances within a local region of a pathology image often exhibit similar features, these common features can enhance each other. Therefore, we propose to use SWTB to model the correlation among instances within a window. SWTB employs a multi-head attention mechanism to enable full interaction among instances within each window. Taking the instances $f_t^{(\tau,k)}$ within window W_k as an example, where $t = 1, 2, ..., T_k$ is the number of instances within W_k, τ represents the number of times the instance has passed through the Cross-window Transformer Block. The process of inputting $f_t^{(\tau,k)}$ into SWTB and outputting $f_t^{(\tau,k)'}$ is represented by Equation 2.

$$f_t^{(\tau,k)'} = \sum_{r=1}^{T_k} \frac{\exp\left(Q_t^r \left(K_t^r\right)^\top\right)}{\sqrt{C_w} \sum_{r=1}^{T_k} \exp\left(Q_t^r \left(K_t^r\right)^\top\right)} V_t^r, \quad (2)$$

where Q_t, K_t, and V_t are the Query, Key, and Value obtained by linearly mapping $f_t^{(\tau,k)}$, respectively.

SWMB uses average pooling to aggregate the features within each window to obtain the fused feature of the window. SWMB takes all instances within each window W_k as input and outputs their averaged pooled feature $f_w^k \in \mathbb{R}^{1\times1\times C_w}$ as the window feature.

We propose the Cross-window Transformer Block (CWTB) to model the correlation between the instances within a window and all windows, thus achieving efficient information interaction and correlation modeling between local and global features. CWTB takes as input the instance features outputted by SWTB and the window feature outputted by SWMB, and uses multi-head attention mechanism to enhance these two types of information. In CWTB, we construct a new feature set $\mathcal{F}_t = [f_t^{(\tau,k)\prime}, f_w^1, f_w^2, ...f_w^K]$ by combining the instance feature $f_t^{(\tau,k)\prime}$ in window W_k with the window feature f_w^k of each window. Then, we establish a cross-attention mechanism on this feature set, allowing each instance within a window to fully interact with each window feature, and obtain the corresponding output feature $f_t^{(\tau+1,k)}$.

$$f_t^{(\tau+1,k)} = \sum_{r=1}^{K+1} \frac{\exp\left(Q_{\mathcal{F}_t} \left(K_{\mathcal{F}_t}^r\right)^\top\right)}{\sqrt{C_w} \sum_{l=1}^{K+1} \exp\left(Q_{\mathcal{F}_t} \left(K_{\mathcal{F}_t}^r\right)^\top\right)} V_{\mathcal{F}_{t'}}^r, \quad (3)$$

where $Q_{\mathcal{F}_t}$ is the Query linearly mapped from instance $f_t^{(\tau,k)\prime}$, $K_{\mathcal{F}_t}^r$ is the Key linearly mapped from the feature set $\mathcal{F}_t$, and $V_{\mathcal{F}_t}^r$ is the Value linearly mapped from the feature set $\mathcal{F}_t$.

In a PiceBlock, we repeatedly use the combinations of SWTB, SWMB, and CWTB to enhance information interaction.

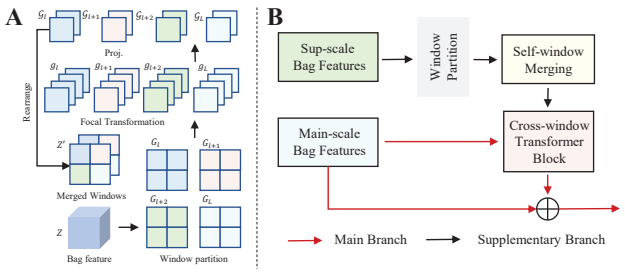

Figure 4. (A) Workflow of the Focal Merging Block. (B) Workflow of the Scale Mixer.

3.4.2 Focal Merging Block

Between each PiceBlock Stage, we use a Focal Merging Block (FMB) to downsample the feature map and form a pyramid structure to further increase the receptive field of PiceFormer, as shown in Figure 4 (A). The input of the FMB is the feature map Z output by the previous PiceBlock, and it is partitioned into windows of size $\xi_h \times \xi_w$, denoted as $\{G_l, l = 1, 2, ..., L\}$, where each window $G_l \in \mathbb{R}^{\xi_h \times \xi_w \times C_G}$. Considering the local similarity of pathology image features, we flatten each spatial scale window G_l with size $\xi_h \times \xi_w$ and stack them along the channel dimension to obtain $\{G_l, l = 1, 2, ..., L\} \rightarrow \{g_l, l = 1, 2, ..., L\}$, where $g_l \in \mathbb{R}^{1\times1\times(\xi_h \times \xi_w \times C_G)}$. Then, we apply a linear mapping matrix to g_l along the channel direction to map it to $\mathcal{G}_l$, where $\mathcal{G}_l \in \mathbb{R}^{1\times1\times\frac{1}{2}(\xi_h \times \xi_w \times C_G)}$. Finally, we rearrange $\mathcal{G}_l$ back to its original window position, completing one Focal Merging operation. After one FMB with a window size of 2×2, the spatial scale of the feature map is halved and the channel dimension is doubled, which greatly enhance the receptive field and feature extraction ability of PiceFormer while reducing computational costs.

3.4.3 Scale Mixer

As shown in Figure 4 (B), the Scale Mixer takes the feature map of either the high or the low scale as the main information, and then uses the feature map of the other scale as supplementary information. The supplementary feature map is partitioned into windows and interacts with the main information using the Self-Window Merging Block and the Cross-Window Transformer Block to obtain processed feature maps, which are then fused into the main information flow. Specifically, Scale Mixer takes the feature maps output by the PiceBlock as input for both branches, and alternately takes one branch's feature map $F_m \in \mathbb{R}^{\eta_h \times \eta_w \times C_\eta}$

as the main information and the other $F_s \in \mathbb{R}^{\eta_h \times \eta_w \times C_\eta}$ as supplementary information. The supplementary feature maps F_s are first partitioned into windows and processed by SWMB to obtain the window features, which are then inputted into CWTB together with main information feature map F_m for information interaction between branches. Finally, the output feature of CWTB is added to the main information feature F_m. We use the Scale Mixer after the PiceBlock in each stage, so that the multi-magnification feature information can fully interact with each other as the network deepens. The Scale Mixer can also be extended to more than two scales.

4. Experiments

4.1. Datasets

We comprehensively evaluated the performance of our MILBooster on three datasets with different tumor types from different centers. The three datasets are Camelyon16 public dataset [2] (breast cancer), TCGA [1]public dataset (lung cancer) and an in-house Cervical Cancer dataset. We experimented on four different WSI classification tasks, including tumor diagnosis, tumor subtyping, prediction of lymph node metastasis from primary lesion WSIs, and patient prognosis prediction. The first two tasks can be directly accomplished by doctors in clinical practice and they are relatively easy. In comparison, the latter two tasks are more difficult and even doctors cannot make the prediction from the given WSI images, where the ground-truth labels for the third task are obtained by directly examining lymph node WSIs after surgery and the ground-truth labels for the fourth task are obtained from long-term follow-up. To demonstrate the powerful performance of our MILBooster at different scales, for each dataset, we provided at least two single-scale evaluation results and one double-scale result. Following DSMIL [17], we cropped each WSI into non-overlapping patches of 224×224 to form a bag. Background blocks with an entropy value of less than 5 were discarded. Detailed descriptions are as follows.

4.1.1 Camelyon16 Public Dataset

The Camelyon16 dataset is a public WSI dataset for detecting breast cancer metastases in lymph nodes. The dataset contains 399 H&E-stained lymph node WSIs (270 for training, and 129 for testing). WSIs with metastases are labeled as positive, while the rest are negative. In addition to providing labels for whether a WSI is positive or negative, the dataset also provides pixel-level labels for positive regions. We only use WSI-level labels for model training and evaluation, while pixel-level labels are not used. We conducted experiments at three magnifications: 20×, 10×, and 5×.

[1]http://www.cancer.gov/tcga

4.1.2 TCGA Lung Cancer Dataset

The TCGA Lung Cancer Dataset includes 1054 H&E-stained WSIs from the Cancer Genome Atlas (TCGA) Data Portal, with the main objective of accurately classifying two subtypes of lung cancer included in the dataset, namely Lung Adenocarcinoma and Lung Squamous Cell Carcinoma. Only the WSI labels are available for this dataset. Following DSMIL [17], we labeled WSIs of Lung Adenocarcinoma as negative and WSIs of Lung Squamous Cell Carcinoma as positive, and conducted experiments at two magnifications, 20× and 5×. We randomly split the WSIs into 840 training slides and 210 testing slides (with 4 low-quality slides discarded).

4.1.3 Cervical Cancer Dataset

The Cervical Cancer Dataset is an in-house clinical dataset, which includes a total of 374 H&E-stained WSIs of primary cervical cancer lesions from different patients after slide selection. All patients underwent abdominal hysterectomy with pelvic lymph node dissection para-aortic lymph node dissection. The lymph node status of all patients was confirmed by professional gynecologic pathologists after surgery. All patients have strict follow-up records of more than five years. We used this dataset to complete two clinical tasks that cannot be directly judged by doctors from H&E slides, namely lymph node tumor metastasis prediction and patient survival prognosis prediction.

Prediction of lymph node metastasis of primary lesion. We labeled the corresponding slides of patients with pelvic lymph node metastasis as positive (209 cases) and the corresponding slides of patients without pelvic lymph node metastasis as negative (165 cases). We conducted experiments at 10× and 5× magnifications. We randomly divided the WSIs into a training set (300 cases) and a test set (74 cases), with an approximate ratio of 4:1.

Prediction of patient survival prognosis. Following Skrede et al. [37], we grouped all patients based on detailed follow-up records according to the median, where those who did not experience cancer-related death within three years were labeled as negative (favorable prognosis), and those who did were labeled as positive (poor prognosis). Then, we randomly divided WSIs into the training set (294 cases) and the test set (80 cases) based on the labels. We conducted experiments at 10× and 5× magnifications.

4.2. Evaluation Metrics, Competitors and Implementation Details

We used the Area Under Curve (AUC) and Accuracy as evaluation metrics. We comprehensively compared MILBooster to six SOTA methods in the field of WSI classification, including ABMIL [14], MILRNN [3], Loss-ABMIL [36], DSMIL [17], TransMIL [35], and DTFD-MIL [44].

Following DSMIL [17], we performed pre-processing to WSI datasets including patch cropping and background removal and we adopted SimCLR [8] as the self-supervised method to pre-extract patch features. The best filter ratios of bag filters vary for each dataset, and we adopted a grid search on the validation set to determine the optimal values. For PiceFormer, the default number of total blocks and stages is set to 12 and 3, respectively. In each stage, SWTB and CWTB are placed alternately and the default hyper-parameter of it is [3, 7, 2]. The window size is set to 6×6 for the window pooling operation in CWTB. The number of heads of MHSA of SWTB and CWTB in stages 1-3 is set as [4, 8, 16], respectively. SGD optimizer with an initial learning rate of 0.002 is used and a linear scheduler is adopted with a 2e-5 minimum learning rate. The warm-up learning rate is set to 2e-4. The total training epoch is 300. For all comparison methods, we reproduced these methods based on the published codes and performed a grid search on the key hyperparameters in our settings.

4.3. Main Results

Tables 1-4 present the experimental results of all methods on the Camelyon16 dataset, the TCGA dataset, and the two different tasks on the Cervical Cancer dataset. For the single-scale experiments, it is evident that MILBooster achieves the best performance in all datasets, scales, and metrics, demonstrating its powerful ability for WSI classification. In the cancer detection task of the Camelyon16 dataset, MILBooster's Accuracy and AUC are 1.2% and 1.8% higher than the second-best method in average across three different scales. In the cancer subtyping task on the TCGA dataset, MILBooster's Accuracy and AUC are 1.5% and 0.6% higher than the second-best method in average across two different scales.

MILBooster shows greater superority in the more difficult tasks on the Cervical Cancer dataset. In the lymph node metastasis task, MILBooster's Accuracy and AUC are 1.8% and 2.2% higher than the second-best method in average across two scales. In the prognosis prediction task, MILBooster's Accuracy and AUC are 1.2% and 0.8% higher than the second-best method in average across two scales.

For the double-scale experiments, first we can see that the double-scale performance of MILBooster is higher than any single-scale performance, which indicates that combining dual-scale information can indeed enhance WSI classification performance. All comparing methods adopted the dual-scale combination approach in DSMIL [17], and MILBooster's Accuracy and AUC are 1.3% and 1.4% higher than the second-best method in average across all four tasks, which demonstrates the effectiveness of MILBooster's dual-scale fusion strategy.

Table 1. The results of Camelyon 16 dataset under single scale and dual-scale (20× and 5×) scenarios.

Model	20× Classification		10× Classification		5× Classification		MS Classification	
	Accuracy	AUC	Accuracy	AUC	Accuracy	AUC	Accuracy	AUC
ABMIL (18'ICML)	0.8450	0.8653	0.8140	0.8379	0.7519	0.7684	0.8760	0.8872
MILRNN (19'Nat. Med)	0.8062	0.8064	0.8140	0.8262	0.7519	0.7603	0.8450	0.8571
Loss-ABMIL (20'AAAI)	0.8605	0.8768	0.8062	0.8299	0.7519	0.7650	0.8837	0.9025
DSMIL(21'CVPR)	0.8682	0.8944	0.8140	0.8401	0.7597	0.7745	**0.8992**	0.9165
TransMIL(21'NeurIPS)	0.8760	0.8987	0.8295	0.8566	0.7674	0.7824	0.8915	0.9127
DTFD-MIL(22'CVPR)	**0.8837**	**0.9008**	**0.8372**	**0.8638**	**0.7907**	**0.8022**	0.8915	**0.9237**
ours	**0.8915**	**0.9187**	**0.8527**	**0.8915**	**0.7984**	**0.8112**	**0.9069**	**0.9427**

Table 2. The results of TCGA-Lung-Cancer dataset under single scale and dual-scale (20× and 5×) scenarios.

Model	20× Classification		5× Classification		MS Classification	
	Accuracy	AUC	Accuracy	AUC	Accuracy	AUC
ABMIL (18'ICML)	0.9000	0.9488	0.8619	0.9269	0.9000	0.9551
MILRNN (19'Nat. Med)	0.8619	0.9107	0.8571	0.9155	0.8905	0.9213
Loss-ABMIL (20'AAAI)	0.9143	0.9517	0.8619	0.9212	0.9286	0.9574
DSMIL(21'CVPR)	0.9190	0.9633	0.8619	0.9373	0.9286	0.9583
TransMIL(21'NeurIPS)	**0.9381**	**0.9830**	0.8667	0.9465	0.9333	0.9599
DTFD-MIL(22'CVPR)	**0.9381**	0.9808	**0.8762**	**0.9483**	**0.9381**	**0.9795**
ours	**0.9476**	**0.9851**	**0.8952**	**0.9600**	**0.9571**	**0.9863**

Table 3. Results on the lymph node metastasis task for the Cervical Cancer dataset under single scale and dual-scale (10× and 5×) scenarios.

Model	10× Classification		5× Classification		MS Classification	
	Accuracy	AUC	Accuracy	AUC	Accuracy	AUC
ABMIL (18'ICML)	0.7973	0.8319	0.7297	0.7716	0.7973	0.8413
MILRNN (19'Nat. Med)	0.7838	0.8111	0.7162	0.7563	0.7973	0.8202
Loss-ABMIL (20'AAAI)	0.7973	0.8324	0.7568	0.7835	0.8108	0.8418
DSMIL(21'CVPR)	0.8243	0.8483	0.7702	0.8022	0.8243	0.8522
TransMIL(21'NeurIPS)	0.8243	0.8501	**0.7838**	**0.8126**	0.8378	0.8634
DTFD-MIL(22'CVPR)	**0.8378**	**0.8533**	0.7703	0.8108	**0.8513**	**0.8678**
ours	**0.8513**	**0.8790**	**0.7973**	**0.8380**	**0.8649**	**0.8820**

Table 4. Results on the prognosis task for the Cervical Cancer dataset under single scale and dual-scale (10× and 5×) scenarios.

Model	10× Classification		5× Classification		MS Classification	
	Accuracy	AUC	Accuracy	AUC	Accuracy	AUC
ABMIL (18'ICML)	0.7000	0.7518	0.6875	0.7439	0.7375	0.7698
MILRNN (19'Nat. Med)	0.7125	0.7333	0.7000	0.7212	0.7125	0.7445
Loss-ABMIL (20'AAAI)	0.7250	0.7617	0.7125	0.7505	0.7375	0.7723
DSMIL(21'CVPR)	0.7375	0.7783	0.7250	0.7622	0.7500	0.7871
TransMIL(21'NeurIPS)	**0.7500**	**0.7835**	**0.7375**	**0.7716**	**0.7750**	**0.7914**
DTFD-MIL(22'CVPR)	0.7375	0.7818	0.7250	0.7650	0.7625	0.7895
ours	**0.7625**	**0.7900**	**0.7500**	**0.7781**	**0.7875**	**0.8036**

5. Ablation Study

MILBooster consists of three main components: bag filter, PiceBlock, and Scale Mixer. The effectiveness of Scale Mixer has already been demonstrated in the results in Tables 1-4. For the other two components, we conducted detailed ablation experiments on the Camelyon16 dataset at 10× magnification. The experiments on bag filter are described in Section 5.1, and those on PiceBlock are described in Section 5.2.

5.1. Bag Filter

Increase of positive ratio after bag filter. Table 5 shows the change of average positive instance ratio of positive bags in the training and testing sets before and after using the bag filter. The original positive instance ratios in

Table 5. Positive instance ratios in the training and testing sets before and after using the bag filter. "w/ bf" represents the use of the bag filter, "Filter Ratio" represents the proportion of instances filtered by the bag filter in each bag relative to the total number of instances, and "w/o bf" represents the original positive ratio of the dataset. Δ represents the increase in positive instance ratio.

Filter Ratio	Pos-ratio of training dataset		Pos-ratio of test dataset	
	w/bf	Δ	**w/bf**	Δ
90%	0.2802	**+16.59%**	0.3801	**+19.45%**
80%	0.2367	**+12.24%**	0.3555	**+16.99%**
70%	0.2115	**+9.72%**	0.3153	**+12.97%**
60%	0.1905	**+7.62%**	0.2912	**+10.56%**
50%	0.1741	**+5.98%**	0.2667	**+8.11%**
40%	0.1597	**+4.54%**	0.2476	**+6.20%**
30%	0.1455	**+3.12%**	0.2314	**+4.58%**
20%	0.1351	**+2.08%**	0.2127	**+2.71%**
10%	0.1242	**+0.99%**	0.2018	**+1.62%**
w/o bf	0.1143	-	0.1856	-

Table 6. Plug-and-play experiment of bag filter.

Method	bag filter	AUC	Δ
ABMIL	✗	0.8379	**+1.11%**
	✓	0.8490	
Loss-ABMIL	✗	0.8299	**+1.65%**
	✓	0.8464	
DSMIL	✗	0.8401	**+1.05%**
	✓	0.8506	
TransMIL	✗	0.8566	**+1.21%**
	✓	0.8687	
DTFD-MIL	✗	0.8638	**+0.95%**
	✓	0.8733	
ours	✗	0.8764	**+1.51%**
	✓	**0.8915**	

the training and testing sets of this dataset were only 11.4% and 18.6%, respectively. This indicates that there are a large number of negative instances in positive bags during training and testing, which could have a potential negative impact on the aggregation of positive bag features. After using the bag filter, there was a significant increase in the positive instance ratios in both the training and the testing sets. This indicates that the bag filter can indeed significantly increase the proportion of positive instances in the bags.

Plug-and-play capability of bag filter. We conducted a plug-and-play experiment of the proposed bag filter in comparing methods with a filter ratio of 0.6, and the results are shown in Table 6. It can be seen that after using the bag filter, the performance of all methods has been significantly improved, with an average increase of 1.2% the AUC metric, which demonstrates the high efficiency and plug-and-play ability of the bag filter.

Sensitivity to the filter ratio. We conducted a sensitivity test on the filter ratio, and the results are shown in Table 7. It can be seen that on this dataset, the best performance

Table 7. Sensitivity test on the filter ratio of the bag filter.

Filter Ratio	90%	80%	70%	**60%**	50%	40%	30%	20%	10%	w/o bf
AUC	0.8608	0.8712	0.8839	**0.8915**	0.8901	0.8868	0.8847	0.8823	0.8815	0.8764

was achieved by filtering out 60% of instances. Though, in Table 5, the higher proportion of instances are filtered, the higher the positive instance ratio will be, the model's performance does not always increase as the positive ratio increases according to Table 7. This is due to the destruction of feature distribution caused by removing too many instances. To further demonstrate this phenomenon, we selected several typical positive and negative bags and visualized their feature distributions before and after bag filter using the t-SNE method in Figure 5.

In the first row of Figure 5, the original positive bag has a positive instance ratio of only 6.4%. After filtering with 90% and 60% filter ratio, the positive instance ratio increases to 42.6% and 14.9%, respectively, while the overall distribution of the original positive and negative instances is well maintained. However, in the second row, although the positive instance ratio of this positive bag is significantly increased from 14.0% to 64.4% after filtering with 90% filter ratio, it loses all positive instances in the lower right corner and a large number of negative instances, which destruct the original feature distribution and harm the performance. In contrast, the 60% filter ratio not only improves the positive ratio but also maintains the main distribution of positive and negative instances fairly well, thus achieving better performance. The situation is the same for negative bags. As shown in the third row, we use 90% and 60% filter ratios to filter negative bags separately. It can be seen that the 90% filter ratio filters out too many instances, causing the negative bags to lose their overall feature distribution. The 60% filter ratio maintains the original feature distribution well. This also explains why the 60% filter ratio achieved the best performance in our ablation experiments.

Effects on the training and testing sets. In another experiment on the $10\times$ Camelyon16 dataset, the AUC was 0.8764 without using the bag filter for both the training and testing sets. When the bag filter was applied to the training set, the AUC improved to 0.8868. The best performance, with an AUC of 0.8915, was achieved when both the training and testing sets utilized the bag filter. Incorporating the bag filter in both sets resulted in the highest performance.

5.2. PiceBlock

To demonstrate the effectiveness of the proposed PiceBlock, we conducted experiments on a variant of the proposed method without using Scale Mixer between the two branches and compared it with current state-of-the-art methods. Experiments are conducted under two settings, with and without using the bag filter, and the results are shown

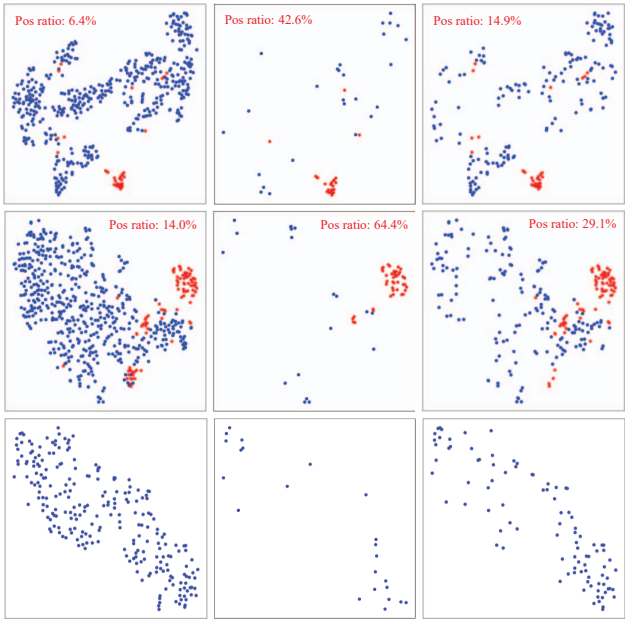

Figure 5. Visualization of typical t-SNE features before and after filtering, where the first two rows are two typical positive bags and their processing results; the third row is a typical negative bag and its processing results. In the t-SNE visualization, red represents positive instances and blue represents negative instances.

in Table 8. We can see that: (1) Without using the bag filter, the performance of the variant is better than all comparing methods. (2) With all methods using the bag filter, the performance of the variant is still better than all comparing methods and the margin is enlarged. In particular, we notice that Swin-Transformer[20] uses a sliding window approach to enhance local and global feature aggregation. Although there is currently no research on using Swin-Transformer in WSI classification, we design a baseline model based on Swin-Transformer as a competitor and denote it as Swin-baseline. It can be seen that the variant without Scale Mixer significantly outperforms the Swin-baseline.

We compared the computational efficiency of the PiceBlock and ViT-B/16 [10] under the same input and network depth. Specifically, we set the input as a matrix $F \in \mathbb{R}^{N\times C}$ with the size of 576×768, where $N = 576$ is the number of feature vectors and $C = 768$ is the dimension of features. The parameter setting of our PiceFormer is as follows: the total number of blocks and stages is set to 12 and 3, respectively. In each stage, SWTB and CWTB are placed alternately. The window size is set to 6×6 for the window pooling operation in CWTB. For the competitor ViT-B/16 [10], the parameter of the architecture is set as follows: the total number of layers is set to 12, the hidden size is 768, the MLP dimension is set to 3072 and the number of heads in each MHSA is 12. In order to keep the input the same as

Table 8. AUC results of the ablation experiment of PiceBlock. The filter ratio used for the bag filter was 0.6.

Model	w/o bf	w/ bf
ABMIL	0.8379	0.8490
Loss-ABMIL	0.8299	0.8464
DSMIL	0.8401	0.8506
TransMIL	0.8566	0.8687
DTFD-MIL	**0.8638**	**0.8733**
Swin-baseline	0.8612	0.8716
ours	**0.8764**	**0.8915**

that of our PiceFormer for fairness, we transform the input image with size $\mathbb{R}^{3\times384\times384}$ to $\mathbb{R}^{576\times768}$. Given the same inputs, our PiceBlock (48.7G FLOPs, 31.8 seconds) can achieve better results with fewer FLOPs and shorter time compared to ViT-B/16 (55.4G FLOPs, 33.2 seconds).

5.3. Focal Merging Block

The Focal Merging Block (FMB) aims to reduce computational costs and increase the receptive field. We also conducted an ablation study on the 20× Camelyon 16 dataset. The AUC with the FMB module (0.9187) outperformed the AUC without it (0.9035), highlighting the positive impact of incorporating the FMB module.

6. Conclusion

Our paper introduces MILBooster, a robust dual-scale and multi-stage MIL framework that significantly enhances WSI classification performance across three critical dimensions: distribution, correlation, and magnification. Firstly, we introduce a plug-and-play bag filter which effectively elevates the positive instance ratio within positive bags by focusing on feature distribution modeling. Secondly, we present a pioneering Transformer architecture, named PiceBlock, which leverages window-based techniques to meticulously capture the interplay between local and global features within pathology images. Additionally, we put forth the Scale Mixer, a dual-branch information interaction module, which enhances overall interaction among diverse magnification levels. Empirical validation on three datasets encompassing four clinical tasks underscores the superiority of MILBooster in both single-scale and dual-scale classification scenarios. Our innovative PiceBlock carries the potential for broader applicability in domains spanning natural image and medical image processing. Moreover, our method offers substantial clinical benefits. In tasks like lymph node metastasis prediction and prognosis estimation, we achieve a notable 1.8% and 1.2% accuracy gain over the second-ranked option. This means around 18 to 12 more accurate identifications within every 1000-patient group, enabling more precise and impactful treatment approaches.

References

[1] Babak Ehteshami Bejnordi, Geert Litjens, Meyke Hermsen, Nico Karssemeijer, and Jeroen AWM van der Laak. A multi-scale superpixel classification approach to the detection of regions of interest in whole slide histopathology images. In *Medical Imaging 2015: Digital Pathology*, volume 9420, pages 99–104. SPIE, 2015. 2

[2] Babak Ehteshami Bejnordi, Mitko Veta, Paul Johannes Van Diest, Bram Van Ginneken, Nico Karssemeijer, Geert Litjens, Jeroen AWM Van Der Laak, Meyke Hermsen, Quirine F Manson, Maschenka Balkenhol, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA*, 318(22):2199–2210, 2017. 6

[3] 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, 6

[4] Richard J Chen, Chengkuan Chen, Yicong Li, Tiffany Y Chen, Andrew D Trister, Rahul G Krishnan, and Faisal Mahmood. Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16144–16155, 2022. 3

[5] Richard J Chen and Rahul G Krishnan. Self-supervised vision transformers learn visual concepts in histopathology. *arXiv preprint arXiv:2203.00585*, 2022. 3

[6] Richard J Chen, Ming Y Lu, Jingwen Wang, Drew FK Williamson, Scott J Rodig, Neal I Lindeman, and Faisal Mahmood. Pathomic fusion: an integrated framework for fusing histopathology and genomic features for cancer diagnosis and prognosis. *IEEE Transactions on Medical Imaging*, 41(4):757–770, 2020. 1

[7] Richard J Chen, Ming Y Lu, Wei-Hung Weng, Tiffany Y Chen, Drew FK Williamson, Trevor Manz, Maha Shady, and Faisal Mahmood. Multimodal co-attention transformer for survival prediction in gigapixel whole slide images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4015–4025, 2021. 1, 2

[8] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning (ICML)*, pages 1597–1607. PMLR, 2020. 3, 7

[9] Veronika Cheplygina, Marleen de Bruijne, and Josien PW Pluim. Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. *Medical Image Analysis*, 54:280–296, 2019. 1

[10] Philip Chikontwe, Meejeong Kim, Soo Jeong Nam, Heounjeong Go, and Sang Hyun Park. Multiple instance learning with center embeddings for histopathology classification. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pages 519–528. Springer, 2020. 1, 2, 9

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2, 3

[12] Yi Gao, William Liu, Shipra Arjun, Liangjia Zhu, Vadim Ratner, Tahsin Kurc, Joel Saltz, and Allen Tannenbaum. Multi-scale learning based segmentation of glands in digital colonrectal pathology images. In *Medical Imaging 2016: Digital Pathology*, volume 9791, pages 175–180. SPIE, 2016. 2

[13] Noriaki Hashimoto, Daisuke Fukushima, Ryoichi Koga, Yusuke Takagi, Kaho Ko, Kei Kohno, Masato Nakaguro, Shigeo Nakamura, Hidekata Hontani, and Ichiro Takeuchi. Multi-scale domain-adversarial multiple-instance cnn for cancer subtype classification with unannotated histopathological images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3852–3861, 2020. 1, 2, 3

[14] Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *International Conference on Machine Learning (ICML)*, pages 2127–2136. PMLR, 2018. 1, 2, 6

[15] Fahdi Kanavati, Gouji Toyokawa, Seiya Momosaki, Michael Rambeau, Yuka Kozuma, Fumihiro Shoji, Koji Yamazaki, Sadanori Takeo, Osamu Iizuka, and Masayuki Tsuneki. Weakly-supervised learning for lung carcinoma classification using deep learning. *Scientific Reports*, 10(1):9297, 2020. 1

[16] Oren Z Kraus, Jimmy Lei Ba, and Brendan J Frey. Classifying and segmenting microscopy images with deep multiple instance learning. *Bioinformatics*, 32(12):i52–i59, 2016. 2

[17] 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 (CVPR)*, pages 14318–14328, 2021. 1, 2, 3, 6, 7

[18] Hang Li, Fan Yang, Yu Zhao, Xiaohan Xing, Jun Zhang, Mingxuan Gao, Junzhou Huang, Liansheng Wang, and Jianhua Yao. Dt-mil: deformable transformer for multi-instance learning on histopathological image. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pages 206–216. Springer, 2021. 1, 2

[19] Tiancheng Lin, Zhimiao Yu, Hongyu Hu, Yi Xu, and Chang-Wen Chen. Interventional bag multi-instance learning on whole-slide pathological images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19830–19839, 2023. 1

[20] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10012–10022, 2021. 9

[21] Ming Y Lu, Richard J Chen, Dehan Kong, Jana Lipkova, Rajendra Singh, Drew FK Williamson, Tiffany Y Chen, and Faisal Mahmood. Federated learning for computational pathology on gigapixel whole slide images. *Medical Image Analysis*, 76:102298, 2022. 1

[22] Ming Y Lu, Richard J Chen, Jingwen Wang, Debora Dillon, and Faisal Mahmood. Semi-supervised histology classification using deep multiple instance learning and contrastive predictive coding. *arXiv preprint arXiv:1910.10825*, 2019. 1

[23] Ming Y Lu, Tiffany Y Chen, Drew FK Williamson, Melissa Zhao, Maha Shady, Jana Lipkova, and Faisal Mahmood. Ai-based pathology predicts origins for cancers of unknown primary. *Nature*, 594(7861):106–110, 2021. 1

[24] 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

[25] Xiaoyuan Luo, Linhao Qu, Qinhao Guo, Zhijian Song, and Manning Wang. Negative instance guided self-distillation framework for whole slide image analysis. *IEEE Journal of Biomedical and Health Informatics*, 2023. 2

[26] Faisal Mahmood, Daniel Borders, Richard J Chen, Gregory N McKay, Kevan J Salimian, Alexander Baras, and Nicholas J Durr. Deep adversarial training for multi-organ nuclei segmentation in histopathology images. *IEEE Transactions on Medical Imaging*, 39(11):3257–3267, 2019. 1

[27] Sam Maksoud, Kun Zhao, Peter Hobson, Anthony Jennings, and Brian C Lovell. Sos: Selective objective switch for rapid immunofluorescence whole slide image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3862–3871, 2020. 3

[28] Oded Maron and Tomás Lozano-Pérez. A framework for multiple-instance learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 10, 1997. 2

[29] Linhao Qu, Siyu Liu, Xiaoyu Liu, Manning Wang, and Zhijian Song. Towards label-efficient automatic diagnosis and analysis: a comprehensive survey of advanced deep learning-based weakly-supervised, semi-supervised and self-supervised techniques in histopathological image analysis. *Physics in Medicine & Biology*, 2022. 1

[30] Linhao Qu, Xiaoyuan Luo, Shaolei Liu, Manning Wang, and Zhijian Song. Dgmil: Distribution guided multiple instance learning for whole slide image classification. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pages 24–34. Springer, 2022. 1, 2, 3

[31] Linhao Qu, Yingfan Ma, Xiaoyuan Luo, Manning Wang, and Zhijian Song. Rethinking multiple instance learning for whole slide image classification: A good instance classifier is all you need. *arXiv preprint arXiv:2307.02249*, 2023. 2

[32] Linhao Qu, Manning Wang, Zhijian Song, et al. Bi-directional weakly supervised knowledge distillation for whole slide image classification. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:15368–15381, 2022. 2, 3

[33] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 234–241. Springer, 2015. 3

[34] Jérôme Rony, Soufiane Belharbi, Jose Dolz, Ismail Ben Ayed, Luke McCaffrey, and Eric Granger. Deep weakly-supervised learning methods for classification and localization in histology images: a survey. *arXiv preprint arXiv:1909.03354*, 2019. 1

[35] 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 (NeurIPS)*, 34:2136–2147, 2021. 1, 2, 3, 6

[36] Xiaoshuang Shi, Fuyong Xing, Yuanpu Xie, Zizhao Zhang, Lei Cui, and Lin Yang. Loss-based attention for deep multiple instance learning. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 34, pages 5742–5749, 2020. 1, 2, 6

[37] Ole-Johan Skrede, Sepp De Raedt, Andreas Kleppe, Tarjei S Hveem, Knut Liestøl, John Maddison, Hanne A Askautrud, Manohar Pradhan, John Arne Nesheim, Fritz Albregtsen, et al. Deep learning for prediction of colorectal cancer outcome: a discovery and validation study. *The Lancet*, 395(10221):350–360, 2020. 6

[38] Chetan L Srinidhi, Ozan Ciga, and Anne L Martel. Deep neural network models for computational histopathology: A survey. *Medical Image Analysis*, 67:101813, 2021. 1

[39] Hiroki Tokunaga, Yuki Teramoto, Akihiko Yoshizawa, and Ryoma Bise. Adaptive weighting multi-field-of-view cnn for semantic segmentation in pathology. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12597–12606, 2019. 2

[40] Xinggang Wang, Yongluan Yan, Peng Tang, Xiang Bai, and Wenyu Liu. Revisiting multiple instance neural networks. *Pattern Recognition*, 74:15–24, 2018. 2

[41] Gang Xu, Zhigang Song, Zhuo Sun, Calvin Ku, Zhe Yang, Cancheng Liu, Shuhao Wang, Jianpeng Ma, and Wei Xu. Camel: A weakly supervised learning framework for histopathology image segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10682–10691, 2019. 1

[42] Yongluan Yan, Xinggang Wang, Xiaojie Guo, Jiemin Fang, Wenyu Liu, and Junzhou Huang. Deep multi-instance learning with dynamic pooling. In *Asian Conference on Machine Learning (ACML)*, pages 662–677. PMLR, 2018. 2

[43] Jiawen Yao, Xinliang Zhu, Jitendra Jonnagaddala, Nicholas Hawkins, and Junzhou Huang. Whole slide images based cancer survival prediction using attention guided deep multiple instance learning networks. *Medical Image Analysis*, 65:101789, 2020. 1, 2

[44] Hongrun Zhang, Yanda Meng, Yitian Zhao, Yihong Qiao, Xiaoyun Yang, Sarah E Coupland, and Yalin Zheng. 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 (CVPR)*, pages 18802–18812, 2022. 1, 2, 3, 6

[45] Xinliang Zhu, Jiawen Yao, Feiyun Zhu, and Junzhou Huang. Wsisa: Making survival prediction from whole slide histopathological images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7234–7242, 2017. 1, 2