A Multiple Instance Learning Framework for Breast Cancer Based on Pseudo-Bag Augmentation and Double-Layer Masking

1st Shuaichao Zhang
Southwest University of Science and Technology
Mianyang, China
919371623@qq.com

2nd Minxian Liu
Southwest University of Science and Technology
Mianyang, China
3466685@qq.com

3rd Bo Su
Southwest University of Science and Technology
Mianyang, China
bosu@foxmail.com

4th Zhengwei Du
Southwest University of Science and Technology
Mianyang, China
2242919550@qq.com

***Abstract*—Breast cancer histopathological image analysis is crucial for diagnosis and treatment planning. Although multiple instance learning (MIL) methods based on whole slide images (WSIs) have shown great promise, existing approaches often suffer from two limitations: overfitting caused by insufficient training samples and a training bias dominated by easily classified instances, which leads to the neglect of hard instances. To address these issues, we propose a multiple instance learning framework for breast cancer based on pseudo-bag augmentation and double-layer masking (PADM-MIL). Specifically, PADM-MIL first expands the training set via pseudo-bag partitioning and mixing. It then integrates a hierarchical Transformer and introduces a teacher network to construct slide-level prototypes. Interpretable instance importance (importance score) is computed using the cosine similarity between instance representations and prototypes. These importance scores further guide instance-level masking to mine hard instances. Subsequently, the student network performs additional group-level masking at the sub-bag level on hard-instance bags, enabling the model to stably focus on key tissue regions. Experiments on two public breast cancer datasets demonstrate that PADM-MIL consistently outperforms existing methods in terms of AUC, ACC, and F1-score. Results on the TCGA-LUNG lung cancer dataset further indicate that the proposed framework exhibits promising generalization.**

***Index Terms*—breast cancer, multiple instance learning, pseudo-bag augmentation, hard instance mining, double-layer masking**

I. Introduction

Breast cancer is one of the most prevalent malignant cancers among women worldwide, and early detection and accurate subtyping are critical for improving patient survival and prognosis [1]. Conventional pathological diagnosis relies on experienced pathologists to manually examine histological slides, a process that is not only time-consuming and labor-intensive but also prone to subjectivity and inter-observer variability [2]. With advances in digital pathology, the acquisition of whole slide images (WSIs) has become increasingly convenient, providing a data basis for developing computer-aided diagnosis systems [3]. However, WSIs typically contain gigapixel-scale ultra-high-resolution information, which poses substantial computational and storage challenges when directly applying conventional deep learning approaches [4].

Multiple instance learning (MIL), as a weakly supervised learning paradigm [5], treats a WSI as a bag and image patches as instances, making it well-suited for WSI analysis. However, conventional MIL methods often suffer from overfitting due to limited training samples, resulting in limited generalization. To mitigate this issue, one approach is to apply patch-level augmentations such as geometric transformations and color perturbations [6], while another approach is to augment the data by exploiting the structural properties of bags within the MIL framework.

Moreover, existing MIL methods mainly rely on attention mechanisms to assign weights to instance contributions [7]. Although attention-based aggregation has proven effective, it suffers from two limitations. First, most approaches emphasize instances with high attention scores while overlooking hard instances that may also be informative [8]. Second, the ultra-long instance sequences in WSIs constrain the use of self-attention in Transformer-based architectures [9]. Recent studies have attempted to address these challenges separately: hard-instance mining strategies have been explored to emphasize difficult samples during training [10], and hierarchical MIL frameworks mitigate long-sequence issues by grouping instances [9]. However, the potential synergy between instance-level and group-level processing remains underexplored.

Motivated by these observations, we propose a multiple instance learning framework for breast cancer based on pseudo-bag augmentation and double-layer masking (PADM-MIL). PADM-MIL first enlarges the training set via pseudo-bag partitioning and mixing. It then integrates a teacher–student paradigm with a hierarchical architecture: the teacher network performs instance-level masking to retain hard instances, while

the student network conducts selective masking at the second hierarchical stage based on group importance. In this way, PADM-MIL enables double-layer instance mining for breast cancer WSI classification.

We conduct experiments on two public breast cancer datasets and one lung cancer dataset. The results demonstrate that PADM-MIL achieves strong overall performance on WSI classification and exhibits promising generalization.

II. RELATED WORK

A. Multiple Instance Learning for WSI

Multiple instance learning (MIL) was first introduced by Dietterich et al. to address drug activity prediction [11]. In MIL, training data are organized into bags of instances. In computational pathology, a bag typically corresponds to an entire whole slide image (WSI), whereas instances are patch images obtained from the WSI through a tiling or segmentation procedure. Under the standard MIL assumption, if a bag contains at least one positive instance, the bag is labeled as positive [12].

Early MIL methods, such as max-pooling and mean-pooling [13], construct bag representations by aggregating instance features with simple pooling functions. In recent years, attention-based MIL methods have become dominant. AB-MIL, proposed by Ilse et al. [7], significantly improves classification performance by learning instance-level weights. CLAM, introduced by Lu et al. [14], adopts a multi-branch architecture to enhance representation capacity. DSMIL, proposed by Li et al. [15], leverages a dual-stream architecture and contrastive learning to improve feature discriminability.

B. Transformers in MIL

Shao et al. first introduced Transformers into MIL and proposed TransMIL [16], which captures correlations among instances via multi-head self-attention and achieves state-of-the-art performance on multiple WSI classification tasks. However, the quadratic complexity of standard Transformers limits their applicability to ultra-long sequences.

To address this issue, a variety of approximate attention mechanisms have been proposed, such as Nyström attention [17] and sparse attention variants [18], although these approaches typically sacrifice modeling capacity. MSG-Transformer [19] introduces a messenger-token mechanism that enables local–global information interaction by propagating information across fixed windows. Inspired by this design, MMIL-Transformer [9] incorporates messenger tokens into MIL to build a hierarchical bag structure, substantially reducing computational cost while preserving full self-attention. Building upon this framework, we further introduce a group-level importance scoring mechanism. Specifically, instance attention produced by the teacher model is used to compute an aggregated importance score for each pseudo-bag, and attention-guided group selection is performed in the second stage to prioritize pseudo-bags containing key information. This mechanism enables the model to focus on diagnostically relevant pathological regions.

C. Augmentation and Hard-Instance Mining

Data augmentation [20] is an important strategy for alleviating overfitting in deep learning [21]. Mixup [22], a simple yet effective augmentation technique, has been widely used in image classification. However, directly applying Mixup to MIL is challenging due to issues such as varying instance numbers and semantic misalignment across bags. Pseb-Mixup [23] addresses these challenges by introducing a systematic bag-level augmentation strategy.

Hard instance mining is a key technique for improving a model's discriminative capability. In MIL, existing methods have primarily emphasized the utilization of salient instances. MHIM-MIL [10] is among the first to propose a masked hard-instance mining strategy, which occludes salient instances to force the model to attend to hard instances. PADM-MIL integrates hard-instance mining with a hierarchical architecture and employs a double-layer masking mechanism to mine informative instances for breast cancer WSI classification.

III. METHOD

A. Overview

We introduce a pseudo-bag mixing augmentation strategy to effectively expand the training set. In addition, by combining a teacher–student paradigm with a hierarchical MIL architecture, the proposed double-layer masking mechanism can mine more informative hard instances, thereby improving the model's discriminative capability. The overall design of our MIL-based classification model is illustrated in Fig. 1.

Specifically, we first follow the preprocessing pipeline of CLAM by treating each WSI as a bag, partitioning it into patches, and extracting patch features using a CNN. We then cluster the instance features within each bag and perform hierarchical sampling to construct smaller pseudo-bags. According to the masking scheme, selected pseudo-bags are merged to form mixed bags.

In the teacher network, each input mixed bag (or original bag) is randomly partitioned into groups and fed into a hierarchical Transformer. A slide-level prototype representation is obtained from the final global token, and importance score is derived from the discrepancy between the prototype and instance representations to perform the first-stage hard-instance masking.

In the student network, when regrouping the masked bag, a second-stage masking is conducted based on group importance; the remaining groups are then input to the hierarchical Transformer.

Finally, we compute a consistency loss by aligning the student's global representation with the teacher's global representation, and optimize the model using the combination of this consistency loss and the student's classification loss, enabling teacher-guided learning and yielding improved representations.

B. Pseudo-bag Partitioning and Mixing

We augment training data by partitioning each slide into pseudo-bags and mixing pseudo-bags across slides to enlarge the effective bag set while preserving semantics.

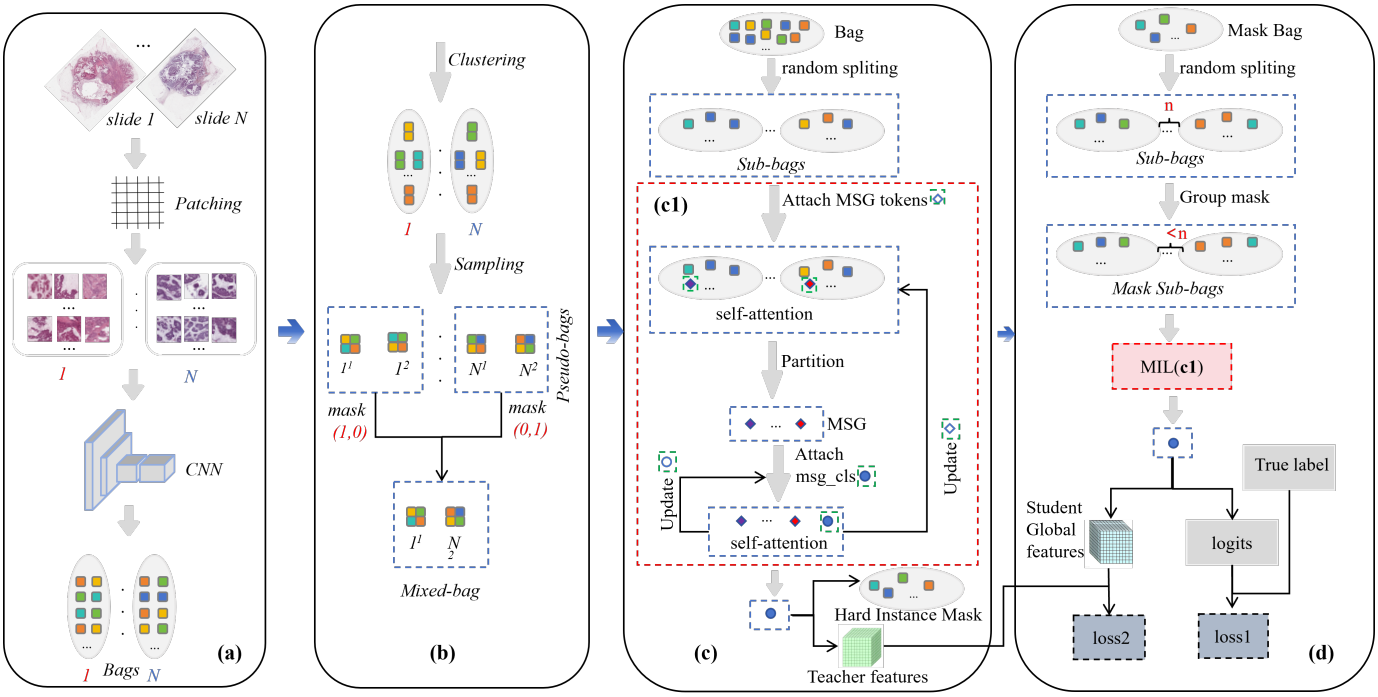

Fig. 1. PADM-MIL overview: (a) patch extraction and CNN features. (b) pseudo-bag mixing. (c) teacher: hierarchical Transformer + prototype-based importance + instance masking. (c1) Hierarchical Transformer Architecture. (d) student: sub-bag selection + consistency learning.

1) Pseudo-bag partitioning. Inspired by the related work on pseudo-bag mixup augmentation [23], we perform mixture enhancement at the pseudo-bag level. Given a bag S containing n instances $\{p_1, p_2, \ldots, p_n\}$, we first perform clustering analysis on the instance features, and the sampling follows the principle of without replacement. Specifically, we apply the K-means algorithm to partition instances into l phenotype categories, where each category corresponds to a distinct tissue morphology pattern. Let the clustering result be $\{C_1, C_2, \ldots, C_l\}$, where C_j denotes the set of instances belonging to the j-th phenotype category.

To construct pseudo-bags, we perform stratified sampling from each phenotype category. That is, for each pseudo-bag B_i^{pseudo}, instances are randomly sampled from each C_j according to a predefined proportion, ensuring that every pseudo-bag contains instances from multiple phenotype categories and thus maintains the semantic diversity of the original bag. This process can be formalized as follows:

$$B_i^{\text{pseudo}} = \bigcup_{j=1}^{l} \text{Sample}\left(C_j, \left\lfloor \frac{|C_j|}{n_p} \right\rfloor\right) \tag{1}$$

where n_p denotes the predefined number of pseudo-bags, and $\text{Sample}(\cdot, k)$ represents randomly sampling k elements from a set. During sampling, we evenly divide each cluster's samples into n_p portions for non-replacement sampling, the remainders are randomly distributed. We set a fixed random seed for the sampling to ensure repeatability.

2) Pseudo-bag mixing augmentation. We perform mixing augmentation at the pseudo-bag level. Given two pseudo-bags B_a^{pseudo} and B_b^{pseudo} sampled from different original bags, with corresponding bag-level labels y_a and y_b, the mixed pseudo-bag and its label are defined as:

$$\tilde{B} = Concat\left(B_a^{pseudo}, B_b^{pseudo}\right) \tag{2}$$

$$\tilde{y} = \lambda_{mix} y_a + (1 - \lambda_{mix})\, y_b \tag{3}$$

where $\lambda_{mix} \sim \text{Beta}(\alpha, \alpha)$ is the mixing coefficient and $\alpha \in (0, 1)$ is a hyperparameter. $\text{Concat}(\cdot, \cdot)$ denotes the concatenation (merging) operation. The mixing operation is applied with probability $p = 0.4$; when the sampled random number exceeds p, the original pseudo-bag is kept without mixing.

C. Teacher Network and Prototype-Based Importance

We adopt a momentum teacher–student framework to estimate stable instance importance and provide representation targets for the student.

1) Momentum update. Let θ_s denote the parameters of the student network and θ_t denote the parameters of the teacher network. The teacher network is updated as follows:

$$\theta_t \leftarrow \lambda_{mom}\theta_t + (1 - \lambda_{mom})\theta_s \tag{4}$$

where λ_{mom} is the momentum coefficient, typically set close to 1; in our implementation, we set $\lambda_{mom} = 0.999$.

2) Hierarchical Transformer and grouping. As illustrated in Fig. 1(c1), Stage 1 performs self-attention within each sub-bag and summarizes it with a messenger token. Stage 2 applies self-attention over messenger tokens to exchange information across sub-bags and produce a slide embedding.

For the grouping rule, given a bag containing n instances, we randomly divide it into m groups of approximately equal size (here we fix the window size w = 1000, and $m = \lceil \frac{n}{w} \rceil$). The last group will be filled to match the group size. During the evaluation, we fix the random seed of the grouping to ensure the results are deterministic.

3) Class prototypes. After each epoch, the teacher computes a prototype per class by averaging slide embeddings:

$$\mu_c = \frac{1}{|S_c|} \sum_{S_i \in S_c} f_t(S_i) \tag{5}$$

where S_c is the set of slides in class c and $f_t(\cdot)$ outputs the teacher slide embedding. The class prototypes are computed solely from original (non-mixed) bags.

4) Instance importance estimation. Based on the constructed slide prototypes, we evaluate the importance of each instance by computing the cosine similarity between the instance representation and the prototypes. For an instance p_i, its importance score is defined as:

$$a_i = \frac{\exp(\cos(h(p_i), \mu_y)/\tau)}{\sum_c \exp(\cos(h(p_i), \mu_c)/\tau)} \tag{6}$$

Here,$\cos(\cdot,\cdot)$ denotes the cosine similarity function, τ is a temperature parameter and we set $\tau = 0.1$. μ_y is the prototype of the class to which the current bag belongs, if it is a mixed bag, each instance within the bag needs to calculate its importance using the actual label of its source. $h(p_i)$ is the output instance embedding from Stage-1 (before messenger-token aggregation) of the teacher network. This prototype-based score is interpretable and does not require explicit attention weights from the aggregator. It should be noted that this process is only used for training the hard instance mining, and it is not necessary during the inference stage.

D. Double-Layer Masking and Objective

1) Instance-level masking. In the first layer, follow the strategy of masking hard instance mining [10], the teacher network processes the instance sequence with the hierarchical Transformer and masks instances according to their importance scores. Specifically, instances are ranked by a_i, and the top-β_h fraction of instances with the highest importance scores is masked (removing these instances from the sequence), thereby forcing the model to attend to other regions that still contain diagnostically relevant information.

To achieve a curriculum learning effect, the high-attention masking ratio β_h is dynamically adjusted as follows:

$$\beta_h^{(t)} = \beta_h^{\text{init}} \cdot \left(1 - \frac{t}{T}\right) \tag{7}$$

where t denotes the current epoch, T is the total number of training epochs, and β_h^{init} is the initial masking ratio. This schedule masks more salient instances at early stages to emphasize hard example learning, and gradually reduces the masking ratio in later stages to leverage complete information for fine-tuning.

2) Sub-bag-level masking. In the second stage, the student network (as shown in Fig. 1(d)) randomly partitions the masked instances into multiple sub-bags $\{G_1, G_2, \ldots, G_m\}$. The importance score of each sub-bag is obtained by aggregating the importance scores of the instances it contains:

$$A_{G_k} = \frac{1}{|G_k|} \sum_{p_i \in G_k} a_i \tag{8}$$

Based on the sub-bag importance scores, we perform attention-guided group selection at the second stage of the hierarchical Transformer. Specifically, only the top-K sub-bags with the highest importance scores are retained for subsequent self-attention computation:

$$\{G'_1, G'_2, \ldots, G'_K\} = \text{TopK}(\{G_k\}_{k=1}^m, \{A_{G_k}\}_{k=1}^m) \tag{9}$$

where $K = m(1-\gamma)$, and γ denotes the group masking ratio. Together with instance-level masking, this selection mechanism facilitates hard-instance mining while making it easier for the model to learn discriminative instance representations, thereby improving its ability to focus on pathological regions containing critical diagnostic information.

3) Loss function. The overall objective of PADM-MIL consists of a classification loss and a consistency loss. For the classification loss, we employ the cross-entropy loss:

$$\mathcal{L}_{\text{cls}} = -\tilde{y} \log(\hat{y}) - (1-\tilde{y}) \log(1-\hat{y}) \tag{10}$$

where $\hat{y}$ denotes the predicted probability and $\tilde{y}$ is the mixed soft label.

For consistency loss, let $(z_s, h_s) = f_\theta(X)$ be the bag-level logits and representation produced by the student (online) encoder, and $h_t = f_{\bar{\theta}}(X)$ be the representation of the teacher (EMA) encoder. We enforce representation-level consistency by minimizing the cosine distance between h_s and h_t:

$$\mathcal{L}_{con} = 2 - 2 \left\langle \frac{h_s}{\|h_s\|_2}, \text{stopgrad}\left(\frac{h_t}{\|h_t\|_2}\right) \right\rangle, \tag{11}$$

where gradients are stopped on the teacher branch since it is updated only by EMA. The final objective is $\mathcal{L} = \mathcal{L}_{cls} + \alpha_{con}\mathcal{L}_{con}$ with $\alpha_{con} = 0.05$.

IV. EXPERIMENTS

A. Experiment Settings

Datasets and evaluation metrics. We conduct a comprehensive evaluation of PADM-MIL on three public datasets: Camelyon16, TCGA-BRCA, and TCGA-LUNG. Camelyon16 is a benchmark dataset for breast cancer lymph node metastasis, consisting of two classes (normal and tumor). From the 270 officially defined training slides, we split them with a 3:1 ratio, using 202 slides for training and the remaining 68 slides for validation, while keeping the official test set unchanged. TCGA-BRCA is a breast cancer pathology dataset containing two subtypes, invasive ductal carcinoma (IDC) and invasive lobular carcinoma (ILC), with a total of 953 WSIs. TCGA-LUNG is a lung cancer cohort including lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC), comprising 1,044 WSIs. For TCGA-BRCA and TCGA-LUNG, we

perform patient-level splitting and partition each dataset into training/validation/test sets with a ratio of 60:15:25. During the preprocessing stage, we employed the pre-trained ResNet50 network and utilized the preprocessing method of CLAM to crop each slice into non-overlapping patches of 256×256 size at a 20× magnification. We also removed the background areas and set the output dimension to 1024. We report WSI-level test performance in terms of accuracy (ACC), area under the ROC curve (AUC), and F1-score (F1).

Implementation details and baselines. All experiments are conducted under CUDA 11.8 and PyTorch 2.0.1. During training, we optimize model weights using the Adam optimizer with a learning rate of 0.0002 and a weight decay of 0.00001. The batch size is set to 1, and all models are trained for 100 epochs with early stopping (patience = 20). We compare our method with a range of MIL approaches, including classical baselines such as max-pooling, mean-pooling, and AB-MIL, as well as more recent methods such as DSMIL, TransMIL, and MHIM-MIL. The model calculates the total loss value using $\mathcal{L}_{cls}+\alpha_{con}\mathcal{L}_{con}$. For the sake of fairness, on the same dataset, the features extracted using the same CLAM method were employed across all models.

B. Experimental Results

The experimental results are shown in Tables I and II. Table I reports the performance of PADM-MIL and existing MIL methods on the two datasets. As shown, compared with two strong baselines, TransMIL and MHIM-MIL, PADM-MIL achieves consistent improvements on Camelyon16 and TCGA-BRCA in terms of AUC, ACC, and F1-score to varying degrees. In addition, we evaluate our model on the TCGA-LUNG dataset. As indicated in Table II, PADM-MIL also ranks among the top-performing methods, suggesting that the proposed framework exhibits a certain degree of generalizability.

TABLE I
THE CLASSIFICATION PERFORMANCE OF THE MODEL ON THE BREAST CANCER DATASET.

Dataset	Method	AUC(%)	ACC(%)	F1(%)
Camelyon16	Max-pooling	80.89	81.40	69.23
	Mean-pooling	72.87	59.80	47.76
	ABMIL [7]	88.29	84.50	78.72
	DSMIL [15]	87.28	84.50	75.61
	TransMIL [16]	87.40	85.27	77.12
	MHIM-MIL [10]	89.31	86.04	80.85
	PADM-MIL	**90.10**	**87.60**	**81.40**
TCGA-BRCA	Max-pooling	87.14	83.88	89.32
	Mean-pooling	88.56	81.82	87.28
	ABMIL [7]	87.08	85.95	90.66
	DSMIL [15]	89.84	85.95	90.56
	TransMIL [16]	90.75	85.95	90.66
	MHIM-MIL [10]	88.37	83.88	89.01
	PADM-MIL	**91.79**	**88.43**	**92.43**

TABLE II
THE CLASSIFICATION PERFORMANCE OF THE MODEL ON THE LUNG CANCER DATASET

Dataset	Method	AUC(%)	ACC(%)	F1(%)
TCGA-LUNG	Max-pooling	90.01	83.97	84.62
	Mean-pooling	90.17	83.96	84.41
	ABMIL [7]	88.49	82.44	83.45
	DSMIL [15]	90.58	83.97	84.89
	TransMIL [16]	89.36	82.44	83.21
	MHIM-MIL [10]	89.93	83.29	**85.23**
	PADM-MIL	**91.04**	**83.97**	85.21

TABLE III
THE EFFECTIVENESS OF EACH METHOD ON THE TWO DATASETS

Dataset	Module	AUC(%)	ACC(%)	F1(%)
Camelyon16	baseline	89.31	85.73	80.42
	group-level masking	89.33	85.77	80.43
	instance-level masking	89.78	86.23	81.10
	Pseb-mixed	90.71	86.82	81.02
	Pseb-mixed and double masking	90.10	87.60	81.40
TCGA-BRCA	baseline	88.43	83.06	88.39
	group-level masking	88.47	83.13	88.44
	instance-level masking	90.01	86.23	89.79
	Pseb-mixed	89.35	85.21	89.11
	Pseb-mixed and double masking	91.79	88.43	92.43

C. Ablation Studies

1) We adopt the hierarchical Transformer architecture as the baseline, with the number of layers fixed at 2. We conducted ablation experiments on two datasets, TCGA-BRCA and Camelyon16, to verify the effectiveness of each module in our model.

As shown in Table III, adding only group-level selection yields negligible gains on both Camelyon16 and TCGA-BRCA, suggesting that group-level modeling alone is insufficient. In contrast, Pseb-mixed consistently improves AUC, ACC, and F1 on both datasets, confirming its effectiveness as a robust augmentation strategy. Further introducing double masking exhibits dataset-dependent behavior: on TCGA-BRCA, it brings a substantial and consistent boost (AUC/ACC/F1 all increase), indicating the benefit of hard instance/group mining under teacher guidance for fine-grained subtype classification. On Camelyon16, we observe a metric divergence—ACC and F1 increase while AUC slightly decreases—implying improved threshold-dependent classification performance but mildly reduced ranking quality, likely due to the sparse and uneven tumor distribution where masking may occasionally remove critical positive evidence.

2) Effects of the number of pseudo-bags and masking ratios. On TCGA-BRCA, we investigate whether the number of pseudo-bags n_p, the initial instance-level masking ratio β_h^{init}, and the group-level masking ratio γ affect the final performance of the model.

As shown in Fig. 2, when the number of pseudo-bags increases from 16 to 18, the AUC improves while the ACC

decreases, which may indicate overfitting. As we further increase the number of pseudo-bags, the overfitting effect is alleviated and the model achieves its best performance at $n_p = 22$. However, when n_p is increased beyond 22, overfitting reappears. Although the overfitting tendency later diminishes, the performance no longer surpasses that obtained at $n_p = 22$. Therefore, we conclude that $n_p = 22$ yields the optimal performance.

In Fig. 3, the variations across different values of l are relatively small, suggesting that the model is generally insensitive to the choice of l. Consequently, when applying K-means clustering, it is not necessary to overly pursue an optimal number of clusters.

From Fig. 4, we observe consistent trends between AUC and ACC, suggesting that the initial instance-level masking ratio β_h^{init} does not induce overfitting. The model achieves the best performance at $\beta_h^{init} = 0.2$. As the ratio continues to increase, performance exhibits a monotonic decline, indicating that excessively large instance masking ratios can impair generalization and prediction accuracy.

As for Fig. 5, the overall trend shows a slight performance drop followed by a recovery and a relatively stable fluctuation, implying that the group masking ratio γ has a weaker impact on performance than the instance masking ratio. Nevertheless, the model attains its best performance at $\gamma = 0.3$. In practice, overly small values of γ should be avoided to better exploit the model's capacity.

3) Heatmap analysis. We select a WSI containing lesion regions for qualitative analysis. As shown in Fig. 6, the figure visualizes the attention scores on the histopathological slide, where blue indicates low attention and red indicates high attention. The left panel shows the whole-slide heatmap, and the right panel provides a magnified view of the region outlined by the red box in the left panel.

From Fig. 6, we observe that the boundary between the lesional and normal tissues is often highlighted in blue (these areas of low-attention may correspond to the lesional or normal tissue), whereas the red regions tend to concentrate on the most conspicuous parts of the lesion. If the model focuses only on the red regions, it may incorrectly classify certain blue boundary areas as normal tissue, thereby reducing diagnostic accuracy. By masking a portion of high-attention regions, our method forces the model to learn from these boundary areas, which improves discriminative capability and leads to more reliable classification results.

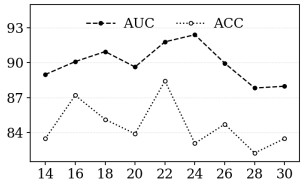

Fig. 2. Ablation studies on n_p.

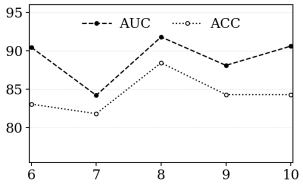

Fig. 3. Ablation studies on l.

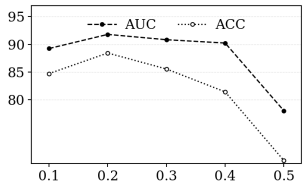

Fig. 4. Ablation studies on β_h^{init}.

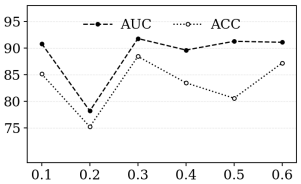

Fig. 5. Ablation studies on γ.

D. Further Analysis

Based on the comparative results, PADM-MIL achieves strong overall performance on Camelyon16 and TCGA-BRCA, and remains competitive on the cross-cancer TCGA-LUNG dataset, indicating that the combination of "pseudo-bag mixing augmentation + double-layer masking" not only improves model performance but also benefits generalization.

Furthermore, ablation studies show that pseudo-bag mixing augmentation yields stable gains on both breast cancer datasets, and incorporating the double-layer masking mechanism leads to additional overall improvements; however, a slight AUC drop is observed on Camelyon16. Considering the task characteristics, this discrepancy is likely related to the highly sparse and uneven distribution of tumor regions in Camelyon16: when high-attention instances are masked, if a slide contains only a few critical tumor instances, overly aggressive masking increases the probability of inadvertently occluding key evidence, thereby degrading bag-level discrimination.

This suggests that double-layer masking does not necessarily benefit from a larger fixed ratio. A more appropriate approach is to adopt dataset-adaptive masking strength and scheduling under different datasets and positive-rate regimes, so as to balance hard-instance mining with the preservation of critical diagnostic cues.

Regarding hyperparameter sensitivity, experiments indicate that the number of pseudo-bags and the initial instance-level masking ratio have a more pronounced impact on performance, whereas the number of phenotype clusters has a relatively minor effect and the group masking ratio exhibits a moderate influence. This trend is also consistent with intuition. The

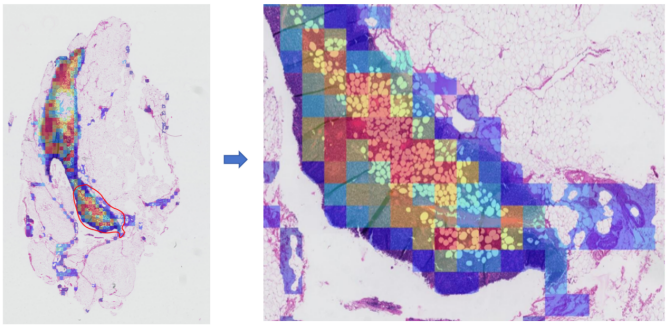

Fig. 6. WSI Heatmap

pseudo-bag number n_p determines the extent to which pseudo-bags cover morphological diversity: too small a value leads to insufficient augmentation, while too large a value may introduce excessive randomness and even exacerbate overfitting.

The instance-level masking ratio directly modulates whether training signals are dominated by salient instances: an overly high ratio may remove critical evidence required for discrimination, whereas an overly low ratio may be insufficient to highlight hard instances. In contrast, the number of phenotype clusters tends to be more robust within a reasonable range because coarse-grained clustering remains stable, making the model less sensitive to this choice.

Based on these observations, practical deployment of PADM-MIL should prioritize tuning n_p and the instance-level masking ratio, while treating group masking as an auxiliary mechanism for stable focusing and redundancy suppression. By retaining only important sub-bags for second-stage interaction, group masking reduces ineffective computation and mitigates the interference of noisy sub-bags in global aggregation, thereby achieving a better trade-off between performance and efficiency.

Although PADM-MIL delivers promising results, several limitations and extensions warrant further investigation. First, pseudo-bag partitioning currently relies on K-means, and the numbers of pseudo-bags/clusters must be predefined; future work may explore adaptive clustering or density-/hierarchy-based pseudo-bag construction, and incorporate cross-slide global clustering to improve phenotype consistency and controllability of augmentation. Second, the optimal configuration of double-layer masking is data-dependent; introducing uncertainty- or confidence-aware dynamic masking may reduce the risk of inadvertently masking critical evidence in scenarios where discriminative cues are scarce.

V. CONCLUSION

We propose a pseudo-bag augmented double-layer masking MIL framework for breast cancer, which improves weakly supervised WSI classification through the principle of "pseudo-bag augmentation + double-layer masking." Specifically, phenotype-aware pseudo-bag partitioning and mixing augmentation are employed to expand the training set and alleviate overfitting caused by limited samples. In addition, by integrating a teacher–student paradigm with a hierarchical Transformer, PADM-MIL performs two-stage masking at both the instance level and the sub-bag level, guiding the model to focus on diagnostically critical yet hard pathological regions. Comparative experiments against representative baselines demonstrate that the proposed framework achieves overall superior classification performance. In future work, we will investigate adaptive clustering strategies and extend the framework to other computational pathology tasks such as segmentation and detection.

REFERENCES

[1] R. Ding, Y. Xiao, M. Mo, et al., "Breast cancer screening and early diagnosis in Chinese women," Cancer Biology & Medicine, vol. 19, no. 4, pp. 450–467, 2022.

[2] B. M. Turner and D. G. Hicks, "Pathologic diagnosis of breast cancer patients: evolution of the traditional clinical-pathologic paradigm toward 'precision' cancer therapy," Biotechnic & Histochemistry, vol. 92, no. 3, pp. 175–200, 2017.

[3] N. Kiran, F. N. U. Sapna, F. N. U. Kiran, et al., "Digital pathology: transforming diagnosis in the digital age," Cureus, vol. 15, no. 9, 2023.

[4] A. Zuraw and F. Aeffner, "Whole-slide imaging, tissue image analysis, and artificial intelligence in veterinary pathology: An updated introduction and review," Veterinary Pathology, vol. 59, no. 1, pp. 6–25, 2022.

[5] Z. Luo, D. Guillory, B. Shi, et al., "Weakly-supervised action localization with expectation-maximization multi-instance learning," in Proc. European Conf. Computer Vision (ECCV), 2020, pp. 729–745.

[6] K. Maharana, S. Mondal, and B. Nemade, "A review: Data preprocessing and data augmentation techniques," Global Transitions Proceedings, vol. 3, no. 1, pp. 91–99, 2022.

[7] M. Ilse, J. Tomczak, and M. Welling, "Attention-based deep multiple instance learning," in Proc. Int. Conf. Mach. Learn. (ICML), 2018, pp. 2127–2136.

[8] S. A. Javed, D. Juyal, H. Padigela, et al., "Additive MIL: Intrinsically interpretable multiple instance learning for pathology," Advances in Neural Information Processing Systems, vol. 35, pp. 20689–20702, 2022.

[9] R. Zhang, Q. Zhang, Y. Liu, et al., "Multi-level multiple instance learning with transformer for whole slide image classification," arXiv:2306.05029, 2023.

[10] W. Tang, S. Huang, X. Zhang, et al., "Multiple instance learning framework with masked hard instance mining for whole slide image classification," in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), 2023, pp. 4078–4087.

[11] T. Dietterich, R. Lathrop, and T. Lozano-Perez, "Solving the multiple instance problem with axis-parallel rectangles," Artificial Intelligence, vol. 89, pp. 31–71, 1997.

[12] J. Amores, "Multiple instance classification: Review, taxonomy and comparative study," Artificial Intelligence, vol. 201, pp. 81–105, 2013.

[13] X. Wang, Y. Yan, P. Tang, et al., "Revisiting multiple instance neural networks," Pattern Recognition, vol. 74, pp. 15–24, 2018.

[14] M. Y. Lu, D. F. K. Williamson, T. Y. Chen, et al., "Data-efficient and weakly supervised computational pathology on whole-slide images," Nature Biomedical Engineering, vol. 5, no. 6, pp. 555–570, 2021.

[15] 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 Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2021, pp. 14318–14328.

[16] Z. Shao, H. Bian, Y. Chen, et al., "TransMIL: Transformer based correlated multiple instance learning for whole slide image classification," Advances in Neural Information Processing Systems, vol. 34, pp. 2136–2147, 2021.

[17] Y. Xiong, Z. Zeng, R. Chakraborty, et al., "Nyströmformer: A Nyström-based algorithm for approximating self-attention," in Proc. AAAI Conf. Artif. Intell., vol. 35, no. 16, 2021, pp. 14138–14148.

[18] Y. Kim, T. Wang, D. Xiong, et al., "Multiple instance neural networks based on sparse attention for cancer detection using T-cell receptor sequences," BMC Bioinformatics, vol. 23, no. 1, p. 469, 2022.

[19] J. Fang, L. Xie, X. Wang, et al., "MSG-Transformer: Exchanging local spatial information by manipulating messenger tokens," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2022, pp. 12063–12072.

[20] P. Alimisis, I. Mademlis, P. Radoglou-Grammatikis et al., "Advances in diffusion models for image data augmentation: A review of methods, models, evaluation metrics and future research directions," Artificial Intelligence Review, vol. 58, no. 4, p. 112, 2025.

[21] A. Mumuni and F. Mumuni, "Data augmentation: A comprehensive survey of modern approaches," Array, vol. 16, Art. no. 100258, 2022.

[22] M. Gadermayr, L. Koller, M. Tschuchnig, et al., "Mixup-MIL: Novel data augmentation for multiple instance learning and a study on thyroid cancer diagnosis," in Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI), 2023, pp. 477–486.

[23] P. Liu, L. Ji, X. Zhang, et al., "Pseudo-bag mixup augmentation for multiple instance learning-based whole slide image classification," IEEE Transactions on Medical Imaging, vol. 43, no. 5, pp. 1841–1852, 2024.