Attention-Challenging Multiple Instance Learning for Whole Slide Image Classification

Yunlong Zhang[1,2], Honglin Li[1,2], Yunxuan Sun[1,2], Sunyi Zheng[2], Chenglu Zhu[2], and Lin Yang[2]

[1] Zhejiang University
[2] Westlake University

Abstract. In the application of Multiple Instance Learning (MIL) methods for Whole Slide Image (WSI) classification, attention mechanisms often focus on a subset of discriminative instances, which are closely linked to overfitting. To mitigate overfitting, we present Attention-Challenging MIL (ACMIL). ACMIL combines two techniques based on separate analyses for attention value concentration. Firstly, UMAP of instance features reveals various patterns among discriminative instances, with existing attention mechanisms capturing only some of them. To remedy this, we introduce Multiple Branch Attention (MBA) to capture more discriminative instances using multiple attention branches. Secondly, the examination of the cumulative value of Top-K attention scores indicates that a tiny number of instances dominate the majority of attention. In response, we present Stochastic Top-K Instance Masking (STKIM), which masks out a portion of instances with Top-K attention values and allocates their attention values to the remaining instances. The extensive experimental results on three WSI datasets with two pre-trained backbones reveal that our ACMIL outperforms state-of-the-art methods. Additionally, through heatmap visualization and UMAP visualization, this paper extensively illustrates ACMIL's effectiveness in suppressing attention value concentration and overcoming the overfitting challenge. The source code is available at https://github.com/dazhangyu123/ACMIL.

Keywords: Computational pathology · Whole slide image · Multiple instance learning · Overfitting

1 Introduction

Whole slide image (WSI) classification is a critical undertaking in digital pathology, aiming to extract valuable information from high-resolution scanned images for precise diagnosis [23,33,51,54], prognosis [11,32,57,63], and treatment planning [14, 35, 37, 40, 41] of diseases. In recent years, multiple instance learning (MIL) [1, 18, 38] has emerged as a promising approach for WSI classification, treating each WSI as a "bag" and its extracted small patches as "instances" within the bag, thus enabling efficient classification of WSIs through assigning a single label to the entire slide.

Overfitting is a significant challenge in utilizing MIL methods for WSI classification [34, 47, 59]. Common WSI datasets exhibit intrinsic characteristics of
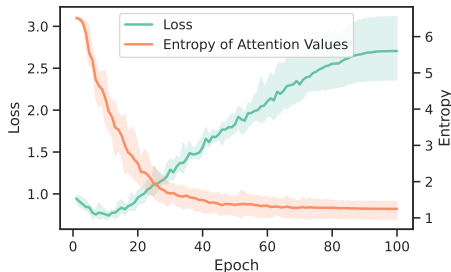
Fig. 1: The change of validation loss and entropy of attention values throughout the training of ABMIL. The results are reported on LBC with SSL pretrained features. There exists the strong negative correlation between loss and entropy.
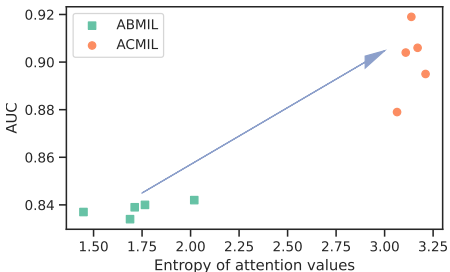
Fig. 2: Comparison of AUC and entropy of attention values between ABMIL and ACMIL. One point denotes the result of a seed on LBC with SSL pretrained features. ACMIL achieves the higher AUC and entropy than ABMIL.

limited data scale, ultra-high resolutions, and staining bias, which makes overfitting more likely [2]. Specifically, these datasets often consist of a relatively small number of slides, typically in the hundreds, with a resolution ranging from $50,000 \times 50,000$ to $10,000 \times 10,000$ [36]. Moreover, pathology images are susceptible to staining bias caused by variations in tissue preparations, staining protocols, and digital scanning methods [60], leading models to learn spurious correlations [34].

In the attention mechanism, attention values/heatmaps provide insights into the model's decision-making process. Multiple existing works [36,45,47,58] alongside our own experiments (as indicated in Sec. 4.3) have pointed out the excessive concentration of attention values in current MIL methods. Specifically, MIL's attention mechanisms often concentrate on a subset of discriminative instances (i.e., instances relevant to the bag label) while disregarding the remaining ones. We investigate the correlation between attention value concentration and overfitting, utilizing the entropy of attention values and validation loss. Fig. 1 depicts a strong negative correlation between loss and entropy throughout the training process, illustrating that over-concentration of attention values (indicated by lower entropy) significantly compromises the model's generalization ability (indicated by higher loss values). Moreover, in the field of natural image classification, recent studies [17,26,49] have demonstrated that models solely relying on a portion of discriminative features could be susceptible to overfitting. Transitioning to WSI classification, fixating on a subset of discriminative instances similarly impedes the model's ability to generalize. These findings highlight the tight connection between attention value concentration and overfitting.

Recently, numerous efforts have been made to address the overfitting challenge by enhancing representation quality [16, 24, 31, 36, 53], building spatial instance correlations [21, 30, 45] and developing data augmentation methods [13, 42, 47, 56, 59]. Additionally, some of these studies [30, 36] suggest that reducing the concentration of attention values can enhance model interpretability.

However, the investigation of attention values concentration for alleviating overfitting remains under-explored.

To mitigate overfitting, we present two analyses for attention value concentration using UMAP and Top-K value statistics. Then, we introduce Attention-Challenging MIL (ACMIL), which combines two novel techniques based on these two analyses. First, by observing UMAP of instance features, we find that *there are various patterns among discriminative instances, and attention mechanisms tend to capture some of them.* To solve this, we introduce Multiple Branch Attention (MBA). MBA utilizes multiple attention branches, each focusing on capturing instances with a specific pattern, thereby ensuring that more discriminative instances contribute to the final prediction. Second, by analyzing the cumulative value of Top-K attention scores, we find that *a tiny number of instances (e.g., K=10) occupy majority attention*, resulting in overlooking sophisticated discriminative instances. To suppress these instances, we propose Stochastic Top-K Instance Masking (STKIM). STKIM randomly masks out a portion of instances with Top-K attention values and assigns their attention values to the remaining instances. Combining MBA and STKIM, our ACMIL effectively alleviates the attention value concentration and suppresses overfitting (see Fig. 2).

We conduct experiments on three WSI datasets (i.e., CAMELYON16, BRACS, and our in-house LBC dataset) with two backbones (i.e., ImageNet pre-trained ResNet18 and SSL pre-trained ViT/S-16). Experimental results demonstrate the superiority of our ACMIL over existing state-of-the-art methods. We also present substantial experimental results, including heatmap visualization and UMAP visualization, to comprehensively demonstrate the effectiveness of ACMIL in suppressing attention value concentration and combatting overfitting.

2   Related Work

2.1   Combating Overfitting in WSI Analysis

In the domain of WSI classification, combating the challenge of overfitting has received substantial attention. Next three paragraphs detail methods from three different aspects.

Some efforts have concentrated on enhancing the quality of instance representations. Early studies (e.g., [7, 27, 45]) rely on backbones pre-trained on the ImageNet dataset. However, the substantial domain gap between natural and pathological images hinders representation quality. Recent works (e.g., [16, 24, 36, 48, 53]) address this by emphasizing Self-Supervised Learning (SSL) to learn patch-level feature representations. In addition, efforts such as the work by Chen et al. [10] leverage hierarchical SSL for high-resolution image representations. Further, studies by Li et al. [31] and Wang et al. [52] demonstrate that fine-tuning the pre-trained encoder is essential for acquiring task-specific information.

Another line of research has focused on establishing spatial instance correlations. DSMIL [30], H²MIL [25], and DAS-MIL [5] consider the hierarchical

structure of patches and aggregate multi-scale representations in attention mechanisms. Furthermore, some studies introduce self-attention layers [43, 45, 55] and graph neural networks [9, 21, 61] to model correlations between different areas.

Further strategies have concentrated on data augmentation. Examples include DTFD-MIL [59], which introduces pseudo-bags for expanding bag counts and employs a double-tier MIL framework. IPS [4], Zoom-In Network [29], and Top-K MIL [13] generate bag representations by aggregating the representations of salient patches. Remix [56] and RankMix [12] introduce instance representation mixup for MIL. MHIM-MIL [47] and WENO [42] augments bags by randomly masking salient instances.

Although our ACMIL shares a similar spirit with some of these works, the proposed ACMIL excels by further building on detailed analysis for attention value concentration. As a result, our ACMIL exhibits stronger interpretability against existing solutions.

2.2   Over-Concentration of Attention Values

In the realm of natural image classification, research has shown that an excessive focus on certain parts of an object can impede the overall effectiveness of model generalization [17, 26, 62]. To tackle this issue, various heuristic techniques have been proposed. For instance, Cutout [17, 62] is a valuable data augmentation method that randomly masks square regions of input during training. Another approach, RSC [26], involves regularization that eliminates salient features activated during training. This paper investigates the issue of attention value concentration in WSI classification tasks. We identify two specific phenomena related to attention value concentration existing in WSI classification and propose two techniques to address them respectively.

3   Method

Based on the ABMIL (detailed in Sec. 3.1), we present ACMIL to alleviate the overfitting problem, which is built on two components: Multiple Branch Attention (MBA) and Stochastic Top-K Instance Masking (STKIM). We describe the details of two components in Sec. 3.2 and 3.3, respectively.

3.1   ABMIL for WSI Classification

In the binary MIL classification problem [7], a bag of instance, $\boldsymbol{X} = \{\boldsymbol{x}_n\}_{n=1}^N$, is associated with a single bag label, $\boldsymbol{Y}$. Each instance, x_n, is associated with a single binary label, y_n, which remains unknown during training. The assumption behind the MIL can be written as:

$$\boldsymbol{Y} = \begin{cases} 0, & \text{iff } \sum_{n=1}^N y_n = 0 \\ 1, & \text{otherwise} \end{cases} \tag{1}$$

In the ABMIL [27], the multiple instance learning is modeled by a three-step process. i) Instance transformation into a low-dimensional embedding through neural networks: $\boldsymbol{h}_n = f(\boldsymbol{x}_n)$. ii) Aggregation of all instance embeddings into the bag-level representation using an attention operator. Specifically, this operation is defined as:

$$\boldsymbol{z} = \sum_{n=1}^{N} a_n \boldsymbol{h}_n \tag{2}$$

Here, $a_n = \sigma(\boldsymbol{h}_n)$ represents the attention values for n-th instance, $\boldsymbol{h}_n$. In the case of ABMIL, a gated attention (GA) mechanism [15] is adopted:

$$\sigma(\boldsymbol{h}_n) = \frac{\exp\{\boldsymbol{w}^{\mathrm{T}}(\tanh(\boldsymbol{V}_1 \boldsymbol{h}_n) \odot \mathrm{sigm}(\boldsymbol{V}_2 \boldsymbol{h}_n))\}}{\sum_{j=1}^{N} \exp\{\boldsymbol{w}^{\mathrm{T}}(\tanh(\boldsymbol{V}_1 \boldsymbol{h}_j) \odot \mathrm{sigm}(\boldsymbol{V}_2 \boldsymbol{h}_j))\}} \tag{3}$$

where $\boldsymbol{V}_1, \boldsymbol{V}_2 \in \mathbb{R}^{L \times M}$, $\boldsymbol{w} \in \mathbb{R}^{L \times 1}$ are parameters, $\odot$ is an element-wise multiplication and $\mathrm{sigm}(\cdot)$ is the sigmoid non-linearity. iii) The bag prediction is generated based on the aggregated bag embedding: $\hat{\boldsymbol{Y}} = g(\boldsymbol{z})$.

3.2 Mutiple Branch Attention

Motivation. It is challenging to capture all discriminative instances using a single attention branch (see Fig. 3). This challenge arises due to variations in patterns among discriminative patches, stemming from differences in texture and morphology. Additionally, DNNs tend to exhibit a form of "laziness" where they prioritize capturing simpler patterns to minimize training loss, neglecting more intricate and challenging patterns [19, 20]. To tackle this issue, we design the MBA that captures more discriminative instances by multiple attention branches.

Fig. 3: Motivation of MBA. UMAP visualization [39] of tumor instance features from CAMELYON16 'test_113' case. *There are various patterns/clusters among tumor instances, and relying on one single branch tends to capture a part of clusters. Three instances are selected to exhibit their texture differences.*

As depicted in Fig. 4 top view, the MBA firstly captures M patterns and then aggregates their embeddings to make predictions. Each pattern is captured by an attention branch. To maintain both the discriminative nature of patterns and semantic diversity between them, we introduce two regularization techniques: semantic regularization and diversity regularization. Firstly, to ensure capturing discriminative patterns, the semantic regularization is accomplished by hanging a MLP layer behind each pattern embedding, equipping with the cross entropy loss function:
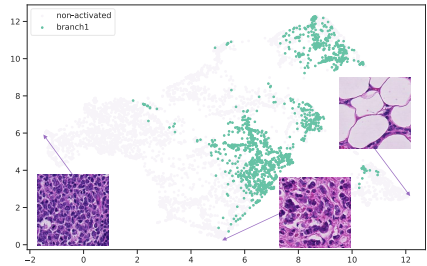
$$\mathcal{L}_p = -\frac{1}{M} \sum_{i=1}^{M} \boldsymbol{Y} \log \hat{\boldsymbol{Y}}_i + (1 - \boldsymbol{Y}) \log\left(1 - \hat{\boldsymbol{Y}}_i\right) \tag{4}$$
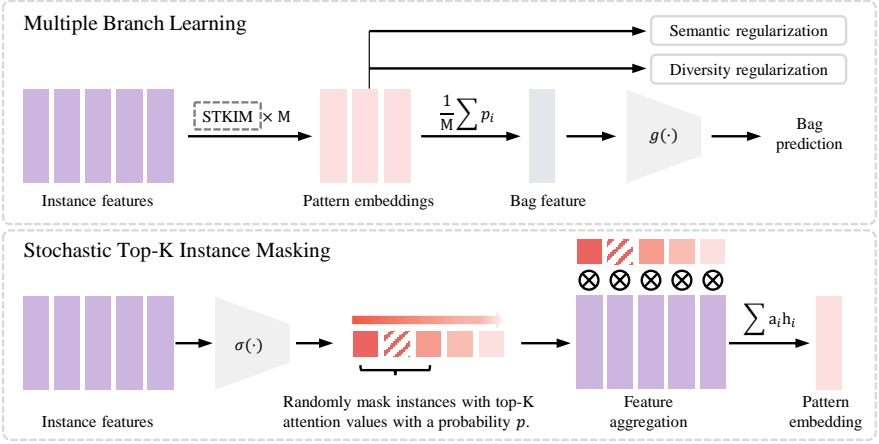
Fig. 4: Overview of the proposed MBA (top view) and STKIM (bottom view). In the MBA, M discriminative patterns are extracted from patch features using the attention operator regularized by semantic and diversity regularization terms. Then, the mean operator is applied to these M pattern features to produce the bag feature, which is utilized for bag-level prediction. In the STKIM, instances with Top-K attention values are randomly masked with a probability p.

where $\hat{Y}_i = g_i(z_i)$ is the prediction based on i-th pattern embedding, z_i. However, only equipping with cross-entropy loss may learn similar patterns and cannot dig out more discriminative information. To tackle this issue, we further introduce a diversity loss as follows:

$$\mathcal{L}_d = \frac{2}{M(M-1)} \sum_{i=1}^{M} \sum_{j=i+1}^{M} \cos(\boldsymbol{a}_i, \boldsymbol{a}_j) \tag{5}$$

where $\boldsymbol{a}_i$ consists of all attention values of i-th pattern, $\boldsymbol{a}_i = \{a_{i1}, \cdots, a_{iN}\}$, also named heatmap as custom. The $\cos(\cdot)$ function is used to measure the similarity of the heatmaps between branches. By diversifying the heatmaps, the embedding of each branch can concentrate on different patterns.

To aggregate the captured patterns to make predictions, the average of heatmaps is utilized as the heatmap of the whole bag:

$$\boldsymbol{a} = \frac{1}{M} \sum_{i=1}^{M} \boldsymbol{a}_i \tag{6}$$

where $\boldsymbol{a}$ is the heatmap of the whole bag, with a dimension of N. Then, the bag embedding can be obtained by aggregating the instance features using averaged heatmap $\boldsymbol{a}$. Moreover, since $\sum_{n=1}^{N}(\frac{1}{M}\sum_{i=1}^{M} a_{in})\boldsymbol{h}_n = \frac{1}{M}\sum_{i=1}^{M}(\sum_{n=1}^{N} a_{in}\boldsymbol{h}_n)$, the bag embedding also can be formulated by applying mean pooling operator to pattern embeddings. The top view of Fig. 4 adopts the latter formulation for

brevity. The loss function for the bag classifier is defined as:

$$\mathcal{L}_b = -\boldsymbol{Y} \log \hat{\boldsymbol{Y}} + (1 - \boldsymbol{Y}) \log \left(1 - \hat{\boldsymbol{Y}}\right) \tag{7}$$

Finally, the overall loss function for the ACMIL can be written as the combination of three loss terms defined in Eq. 4, 5 and 7,

$$\mathcal{L} = \mathcal{L}_b + \mathcal{L}_p + \mathcal{L}_d \tag{8}$$

Discussion. It's important to highlight that when the parameter M is set to 1 in MBA, it essentially mirrors the feature aggregation process of ABMIL, allowing for the discernment of a single pattern. In this sense, MBA serves as an extension of ABMIL, specifically designed to capture a more diverse set of patterns. We further discuss the connection between MBA and Multiple-Head Attention (MHA). HIPT [10] has unveiled that distinct heads in MHA can effectively capture different visual concepts, akin to the role played by our MBA. However, these two techniques can be easily distinguished by: 1) MBA has diversity regularization, ensuring that different branches can learn different concepts. This is absent in MHA, resulting in different heads learning the same concept [10]. We demonstrate the critical role of $\mathcal{L}_d$ for performance in Tab. 3b. 2) MHA is a type of attention formulation, while MBA operates independently of the attention formulation, accommodating MHA within its framework. Appendix Sec. 9.1 reports the results of combining MHA and ACMIL.

3.3   Stochastic Top-K Instance Masking

Motivation. A tiny number of instances will occupy the majority of attention in ABMIL while ignoring sophisticated discriminative instances. As depicted in Fig. 5, the sum of top-10 attention values is larger than 0.85 on all three datasets. However, the WSI typically involves more than 10 discriminative instances. For instance, in the CAMELYON16 dataset, 129 out of 155 tumor slides contain 10 to 20,000 cancerous instances. In essence, numerous discriminative instances are overlooked. To deal with this issue, the proposed STKIM aims to suppress the salient instances and assign more attention to the remaining instances.

Fig. 5: Motivation of STKIM. Accumulation of Top-K attention values. *Instances with Top-K attention values occupy majority attention.* Results are derived from features extracted through supervised pre-training.

   As depicted in Fig. 4 bottom view, STKIM introduces a masking operation into the attention mechanism, before feature aggregation and after attention values generation. The primary objective is to suppress Top-K salient instances. A
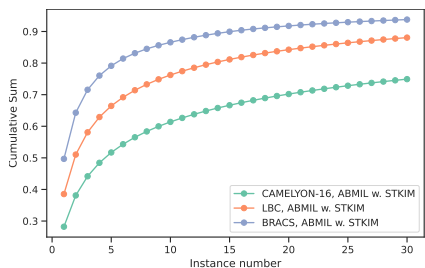
straightforward solution to achieve this is to mask out all of the Top-K salient instances. However, this method poses certain challenges. It can result in the loss of information associated with key instances, which are crucial for discrimination. Furthermore, it might lead to a statistical mismatch between the feature representations before and after discarding these key instances. To address these issues, we draw inspiration from dropout [46] and cutout [17,62] commonly used in computer vision. Our proposed solution employs stochastic masking for instance features with Top-K attention values. Specifically, we begin by sorting all attention values from highest to lowest. Subsequently, we randomly set the attention values of the Top-K instances to 0, with a probability of p. This process can be formulated as:

$$a_n = \begin{cases} 0, & \text{with probability } p \text{ and within Top-K values} \\ a_n, & \text{otherwise} \end{cases} \tag{9}$$

where p and K are two hyperparameters that control the intensity of masking. Following Eq. 9, we will assign the attention values of masked instances to the remaining instances by $a_n \to \frac{1}{\sum_{n=1}^{N} a_n} a_n$. Notably, drawing inspiration from dropout and cutout, we remove STKIM at the inference time.

Discussion. While STKIM, MHIM-MIL [47], and WENO [42] all adopt the technique that masks salient instances, there are notable technical distinctions between them. Firstly, our STKIM has the different masking strategy compared with WENO and MHIM-MIL. STKIM only masks a minority of instances (i.e., $K = 10$) with a probability of p. As a comparison, the other two methods mask out a larger number of instances. WENO masks out 95 instances. MHIM-MIL masks 1% instances. In our framework, our scheme performs best in three strategies (see Appendix Sec. 9.5). Secondly, both MHIM-MIL and WENO necessitate a well-trained model for masking out salient instances, utilizing the remaining instances for model training. They both employ a teacher-student framework, wherein the teacher model needs to be pre-trained beforehand (the warm-up process in WENO and the pre-training stage in MHIM-MIL). In contrast, STKIM requires neither a teacher-student framework nor a pre-training process, thus highlighting simplicity and efficiency.

4   Experiments

4.1   Experimental Details

Datasets and Evaluation Metrics. The performance of ACMIL is evaluated on two public WSI datasets, i.e., CAMELYON16 [3] and BRACS [6], and one private benchmark, LBC. CAMELYON16 dataset consists of 400 WSIs in total, including 270 for training and 130 for testing. Following [30, 59], we further randomly split the training and validation sets from the official training set with a ratio of 9:1. We do not resplit BRACS dataset as it has been officially split to 395 of training set, 65 of validation set, and 87 of test set. We follow the challenge for a 3-class WSI classification: benign tumor, atypical tumor, and malignant. The

Table 1: The performance of different MIL approaches across three datasets, two pre-trained methods, and two evaluation metrics. The most superior performance is highlighted in **bold**, while the second-best performance is indicated by underlining.

Performance Method	CAMELYON-16		BRACS		LBC	
	F1-score	AUC	F1-score	AUC	F1-score	AUC
ResNet-18 ImageNet pretrained						
Max-pooling	0.582±0.170	0.620±0.155	0.489±0.047	0.738±0.014	0.476±0.033	0.775±0.010
Mean-pooling	0.592±0.026	0.597±0.033	0.484±0.029	0.685±0.011	0.511±0.022	0.797±0.011
Clam-SB	0.742±0.024	0.763±0.049	0.521±0.046	0.750±0.039	0.514±0.024	0.805±0.017
TransMIL	0.643±0.088	0.706±0.076	0.444±0.040	0.732±0.043	0.385±0.013	0.693±0.027
DSMIL	0.736±0.028	0.773±0.034	0.511±0.052	0.751±0.028	0.458±0.029	0.766±0.023
DTFD-MIL	0.758±0.051	0.815±0.063	0.469±0.016	0.717±0.032	0.473±0.021	0.776±0.021
IBMIL	0.777±0.009	0.799±0.050	0.510±0.043	0.726±0.034	0.489±0.017	0.791±0.021
MHIM-MIL	0.752±0.034	0.772±0.026	0.511±0.022	**0.774±0.021**	0.543±0.037	0.816±0.009
ABMIL	0.757±0.020	0.790±0.027	0.523±0.028	0.723±0.035	0.465±0.040	0.798±0.013
ACMIL(ours)	**0.798±0.029**	**0.841±0.030**	**0.552±0.048**	0.754±0.008	**0.546±0.028**	**0.821±0.015**
ViT-S/16 SSL pretrained						
Max-pooling	0.903±0.054	0.956±0.029	0.596±0.029	0.823±0.033	0.590±0.043	0.829±0.023
Mean-pooling	0.577±0.057	0.569±0.081	0.522±0.038	0.739±0.007	0.559±0.024	0.827±0.012
Clam-SB	0.925±0.035	0.969±0.024	0.631±0.034	0.863±0.005	0.617±0.022	0.865±0.018
TransMIL	0.922±0.019	0.943±0.024	0.631±0.030	0.841±0.006	0.539±0.028	0.805±0.010
DSMIL	0.943±0.007	0.966±0.009	0.577±0.028	0.816±0.028	0.562±0.028	0.820±0.033
DTFD-MIL	0.948±0.007	**0.980±0.011**	0.612±0.080	0.870±0.022	0.612±0.034	0.842±0.010
IBMIL	0.912±0.034	0.954±0.022	0.645±0.041	0.871±0.014	0.604±0.032	0.834±0.014
MHIM-MIL	0.932±0.024	0.970±0.037	0.625±0.060	0.865±0.017	0.658±0.041	0.872±0.022
ABMIL	0.914±0.031	0.945±0.027	0.680±0.051	0.866±0.029	0.595±0.036	0.831±0.022
ACMIL(ours)	**0.954±0.012**	0.974±0.012	**0.722±0.030**	**0.888±0.010**	**0.662±0.043**	**0.901±0.011**

liquid-based cytology (LBC) dataset collected 1,989 WSIs and included 4 classes, i.e., Negative, ASC-US, LSIL, and ASC-H/HSIL. We randomly split the whole dataset into training, validation, and test sets with the ratio of 6:2:2. Following [31], macro-AUC and macro-F1 scores are reported since all three datasets are class imbalanced. Each of the main experiments is performed five times with random parameter initializations, and the average classification performance and standard deviation are reported. Besides, following [36,59], the test performance is reported in epochs with the best validation performance.

Baselines. We systematically assess the efficacy of our ACMIL approach by benchmarking it against conventional MIL pooling strategies, Max-pooling and Mean-pooling, as well as contemporary attention-based techniques such as AB-MIL [27], DSMIL [30], TransMIL [45], CLAM-SB [36], DTFD-MIL [59], MHIM-MIL [47], and IBMIL [34]. In pursuit of a comprehensive comparison across diverse aggregation operators, we utilize two distinct sets of features derived from ResNet-18 pre-trained on the ImageNet dataset [22] and ViT-S/16 pretrained using DINO [8] on a substantial collection of 36,666 WSIs [28]. The results of all other methods are reproduced using the official code they provide under the same settings.

Implementation Details. Implementation Details are in Appendix Sec. 8.

4.2 WSI Classification Results

Tab. 1 provides a thorough comparison of performance between ACMIL and existing MIL methods. This evaluation spans three diverse datasets, involves two different choices for pretraining methods, and employs two crucial evaluation metrics, resulting in a comprehensive assessment with a total of 12 terms.
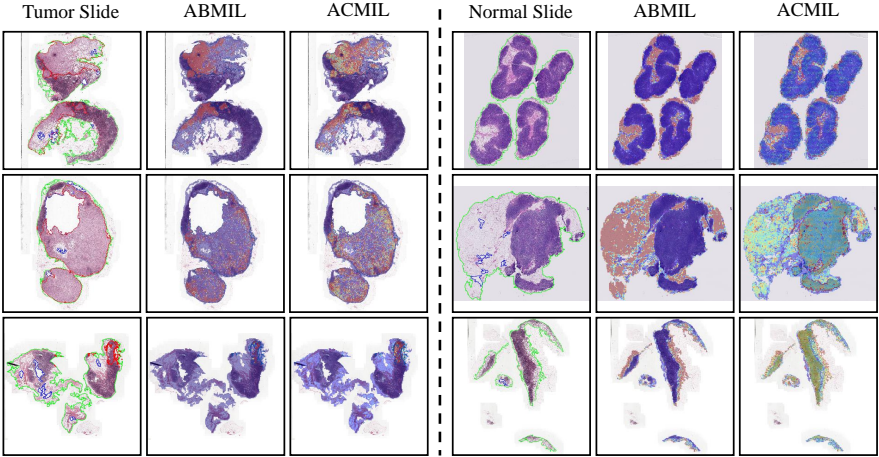
Fig. 6: Heatmap visualization of WSI examples produced by ABMIL [27] (baseline) and our ACMIL. The left part shows three tumor WSIs that come from CAMELYON16 dataset, and their tumor regions are delineated by red lines. ACMIL generates attention values that cover a more extensive portion of the tumor region compared to ABMIL. The right part shows three normal WSIs that come from CAMELYON16 dataset. ABMIL primarily focuses on a part of tissue such as adipose, while ACMIL extends its attention to the more normal tissues.

Considering the overall performance, ACMIL consistently outshines existing methods. It secures the top position in 10 out of the 12 metrics and holds the second position in the remaining 2 metrics. Specifically, for the CAMELYON16, ACMIL achieves outstanding results using ResNet-18 pre-trained on ImageNet embeddings, surpassing the runner-up by 2.1% and 2.6% in terms of F1-score and AUC, respectively. On the other hand, with ViT-S/16 SSL pretrained embeddings, existing attention-based MIL methods exhibit remarkable performance, boosting F1-scores and AUC values exceeding 0.9. Notably, ACMIL achieves comparable performance with the former best-performing method, DTFD-MIL, in this setup. For the BRACS, ACMIL demonstrates a substantial lead when utilizing ViT-S/16 SSL pre-trained embeddings, surpassing the second-best performance by margins of 4.2% and 1.7% in F1-score and AUC, respectively. Moreover, when employing ResNet-18 pre-trained on ImageNet embeddings, ACMIL achieves comparable performance with the previously top-performing method, MHIM-MIL. For the LBC, ACMIL stands out significantly among the other methods across all four metrics.

4.3 Localization Results

Heatmap visualization. Fig. 6 presents heatmap visualizations illustrating examples of our approach's performance in comparison to the baseline method, ABMIL [27]. Three tumor slides (left part) and three normal slides (right part) are selected to showcase the heatmap differences. Due to the space limitation, we present more visualization in Appendix Sec. 9.3 for further insights.

For the tumor slides, ABMIL tends to concentrate its attention on only a fraction of the tumor regions, potentially overlooking other significant areas. In contrast, ACMIL allocates attention across a wider spectrum of tumor regions, resulting in better alignment with expert annotations. For the normal slides, ABMIL predominantly focuses on specific tissue types, such as adipose tissue. This will lead to misinterpretation that only the adipose tissue is the normal tissue and other normal regions are uncorrelated to the WSI label. On the other hand, ACMIL effectively distributes attention values to encompass all normal regions, ensuring all regions are correlated for the WSI label. This approach closely mimics human intuition and satisfies the definition of the MIL formulation.

FROC results. We employ the FROC metric suggested by CAMELYON16 challenge to evaluate the localization of tumor region quantitatively. As shown in Tab. 2, the proposed ACMIL achieves higher FROC than ABMIL.

Table 2: Comparison of FROC between ABMIL and ACMIL

	ABMIL	ACMIL
FROC	0.3987	0.4233

4.4   Ablation Study

Fig. 7 illustrates the AUC scores of ACMIL across three datasets when utilizing a ViT/B-16 feature extractor and varying hyperparameter settings. Several key observations emerge from these experiments:

Effect of branches number M in MBA. As shown in the first column, we find that the choice of M affects performance significantly. Combining three datasets, setting $M = 5$ achieves the best performance.

Effect of masking probability p in STKIM. As shown in the second column, we find that the choice of M also affects performance significantly. Notably, setting $p = 1.0$ (masking all of Top-K instances) leads to performance deterioration across all three datasets. For LBC and CAMELYON, a $p = 1.0$ setting even results in performance lower than the blue dotted lines. Otherwise, $p = 0.6$ achieves the best performance on the BRACS dataset, whereas $p = 0.8$ achieves the best performance on the other two datasets.

Effect of number of masking instances K in STKIM. The third column shows that hyperparameter K exhibits minimal sensitivity, where different K values result in a performance difference of less than 1.0% AUC. In practice, setting K to 10 is generally sufficient for achieving near-optimal performance.

Implementing either MBA or STKIM individually leads to significant performance improvements. The blue dotted lines represent ACMIL's AUC performance without MBA or STKIM, outperforming the orange dotted lines (ABMIL's performance) across all subfigures. Particularly noteworthy is the observation that MBA achieves better improvement than STKIM on all three datasets, with blue dotted lines in the last two columns surpassing those in the first column, especially on the CAMELYON and LBC datasets.

Combining MBA with STKIM yields greater performance improvements compared to using either MBA or STKIM alone. The green dots represent ACMIL's performance under different hyperparameter combinations, with 39 out of 45 green dots exceeding blue dotted lines.
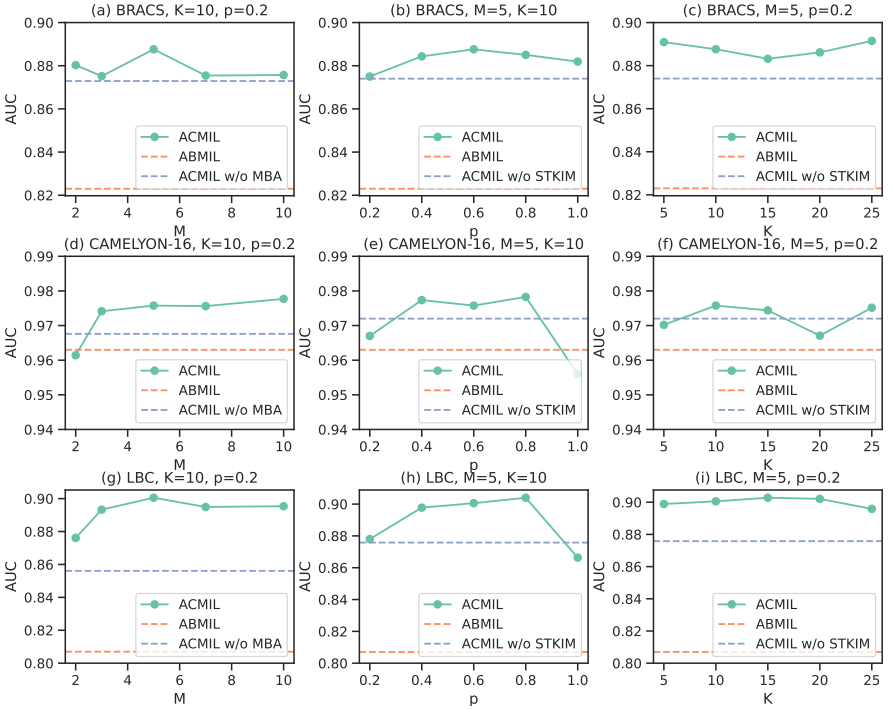
Fig. 7: Ablation study on features extracted through the SSL pre-training. The effect of three hyperparameters, K, p, M, is investigated. Note that the orange dotted line denotes the performance of baseline, ABMIL, and the blue dotted line denotes the performance of ACMIL w/o MBA or STKIM. Five conclusions derived from the figure can be found in Sec. 4.4.

4.5   Further Analysis

MBA can capture diverse patterns. We employ UMAP [39] to visualize instance features within the tumor region of the CAMELYON16 'test_090' case. In Fig. 8a, it's evident that the tumor instances exhibit two primary patterns. However, ABMIL primarily activates the left pattern (colored orange) and neglects the right one. On the other hand, as demonstrated in Fig. 8b, MBA's various branches (branch1, branch2, branch3, and branch5) collectively capture the substructures of the left pattern, while branch4 specifically captures the right pattern. Combining all branches can capture more comprehensive patterns.

ACMIL can learn more discriminative bag features. We employ UMAP [39] to visualize bag features from the LBC test set, as illustrated in Fig. 9a and 9b. This visualization illustrates that our ACMIL is capable of learning more discriminative features compared to ABMIL. Specifically, ACMIL successfully separates the LSIL and ASC-H/HSIL clusters from the Negative cluster. To quantitatively assess the clustering performance, we employ V-measure [44]. ACMIL achieves a V-measure score of 0.316, a significant improvement over ABMIL, which scores 0.224.
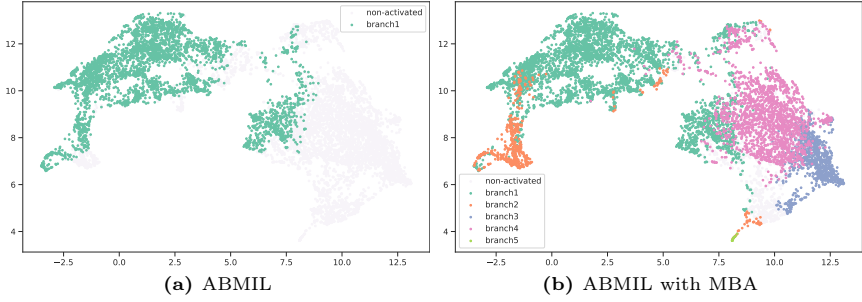
(a) ABMIL   **(b)** ABMIL with MBA

Fig. 8: UMAP visualization of tumor instance features from the CAMELYON16 'test_090' case.The tumor instances display distinct patterns, posing a challenge for a single branch to capture all of them. As a result, ABMIL overlooks the right pattern/cluster. In contrast, multiple branches in MBA capture different patterns separately, and their combination enables the activation of more tumor instances. An instance is considered active when its attention value surpasses $\frac{1}{N}$.
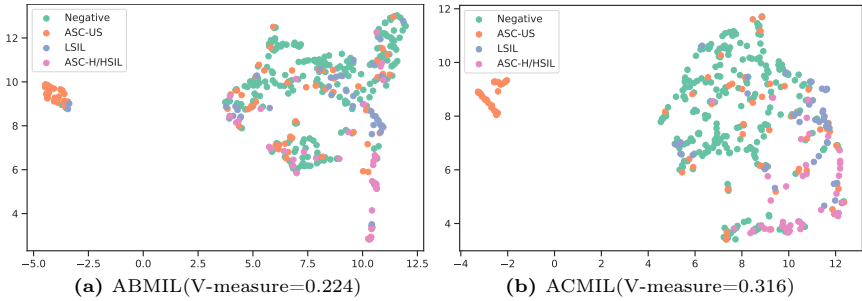

(a) ABMIL(V-measure=0.224)   **(b)** ACMIL(V-measure=0.316)

Fig. 9: UMAP visualization [39] of bag features for LBC test set. ACMIL effectively learns more discriminative features than ABMIL by improving the separation of 'LSIL' and 'ASC-H/HSIL' features from the 'Negative' class. Improving feature separation is also corroborated by the V-measure score [44], a clustering metric that considers both the homogeneity and completeness of the clusters.

STKIM can suppress the concentration of Top-K attention values. Fig. 10 illustrates a comparison of the cumulative sum of Top-K attention values with and without STKIM. The plot clearly shows that the use of STKIM helps mitigate the scenario where Top-K attention values excessively dominate in the attention mechanism. This effect is particularly pronounced for CAMELYON16 dataset, where the cumulative sum of the top-10 values decreases from 0.87 to 0.6.
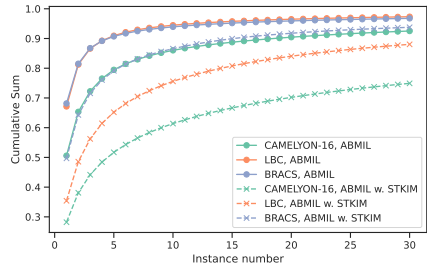

Fig. 10: Comparison of the accumulative sum of Top-K attention values with and without STKIM. The use of STKIM helps alleviate the issue of excessive concentration of attention values within the Top-K range.

Table 3: (a): Performance comparison between ACMIL with (w.) and without (w/o.) T-STKIM. T-STKIM means using STKIM at the test phase. The Gap column reports the performance difference between with and without T-STKIM. Using STKIM at the test phase slightly reduces performance. (b): Performance comparison between ACMIL with (w.) and without (w/o.) $\mathcal{L}_d$. The Gap column reports the performance difference between without and with $\mathcal{L}_d$. ACMIL without $\mathcal{L}_d$ drastically reduces its performance.

(a) w. T-STKIM v.s. w/o. T-STKIM

		ViT-S/16 SSL pre-trained		
Dataset	Metric	w. T-STKIM	w/o. T-STKIM	Gap(%)
Camelyon	F1-score	0.927±0.057	0.954±0.012	+2.7
	AUC	0.967±0.017	0.974±0.012	+0.7
BRACS	F1-score	0.697±0.033	0.722±0.030	+2.5
	AUC	0.875±0.012	0.888±0.010	+1.3
LBC	F1-score	0.637±0.034	0.662±0.043	+2.5
	AUC	0.878±0.012	0.901±0.011	+2.3

		ResNet-18 Imagenet pre-trained		
Dataset	Metric	w. T-STKIM	w/o. T-STKIM	Gap(%)
Camelyon	F1-score	0.780±0.026	0.798±0.029	+1.8
	AUC	0.837±0.028	0.841±0.030	+0.4
BRACS	F1-score	0.566±0.054	0.552±0.048	-1.4
	AUC	0.750±0.021	0.754±0.008	+0.4
LBC	F1-score	0.535±0.027	0.546±0.028	+1.1
	AUC	0.808±0.019	0.821±0.015	+1.3

(b) w. $\mathcal{L}_d$ v.s. w/o. $\mathcal{L}_d$

		ViT-S/16 SSL pre-trained		
Dataset	Metric	w/o. $\mathcal{L}_d$	w. $\mathcal{L}_d$	Gap(%)
Camelyon	F1-score	0.901±0.037	0.954±0.012	+5.3
	AUC	0.943±0.027	0.974±0.012	+3.1
BRACS	F1-score	0.642±0.046	0.722±0.030	+8.0
	AUC	0.859±0.020	0.888±0.010	+2.9
LBC	F1-score	0.603±0.023	0.662±0.043	+5.9
	AUC	0.837±0.009	0.901±0.011	+6.4

		ResNet-18 Imagenet pre-trained		
Dataset	Metric	w/o. $\mathcal{L}_d$	w. $\mathcal{L}_d$	Gap(%)
Camelyon	F1-score	0.747±0.022	0.798±0.029	+5.1
	AUC	0.796±0.032	0.841±0.030	+5.5
BRACS	F1-score	0.500±0.031	0.552±0.048	+5.2
	AUC	0.760±0.026	0.754±0.008	-0.6
LBC	F1-score	0.532±0.019	0.546±0.028	+1.4
	AUC	0.809±0.018	0.821±0.015	+1.2

Do we need STKIM at the test phase? The answer is No. In Tab. 3a, we present the outcomes of ACMIL with and without STKIM during the test phase, along with the performance differences between these settings. Across 11 out of 12 evaluation metrics, the version of ACMIL without STKIM during testing outperforms the version with STKIM slightly. This suggests that STKIM is not necessary during the test phase.

Do we need diversity loss in MBA? The answer is Yes. In Tab. 3b, we present the outcomes of ACMIL with and without $\mathcal{L}_d$, along with the performance differences between these settings. Notably, the last column clearly indicates a significant performance drop for ACMIL without $\mathcal{L}_d$. This emphasizes the crucial role of $\mathcal{L}_d$ in encouraging different branches to acquire distinctive discriminative knowledge within the MBA technique.

5   Conclusion

Due to the intrinsic properties of WSI, MIL methods have often led to overfitting, limiting their applications. This paper reveals that the overly concentrated attention values in the heatmap are closely related to overfitting. To address this, we propose ACMIL, which is underpinned by two novel techniques: MBA and STKIM. Our experimental results on three datasets demonstrate that ACMIL significantly surpasses SOTA methods. Moreover, this paper provides comprehensive experiments confirming the effectiveness of ACMIL in suppressing the attention value concentration and alleviating overfitting. We hope that our work can inspire future exploration into leveraging attention values for a comprehensive analysis of attention mechanisms. We also hope that our ACMIL can be applied to a broader spectrum of WSI analysis tasks.

References

1. Amores, J.: Multiple instance classification: Review, taxonomy and comparative study. Artificial intelligence **201**, 81–105 (2013) 1

2. Bejani, M.M., Ghatee, M.: A systematic review on overfitting control in shallow and deep neural networks. Artificial Intelligence Review pp. 1–48 (2021) 2

3. Bejnordi, B.E., Veta, M., Van Diest, P.J., Van Ginneken, B., Karssemeijer, N., Litjens, G., Van Der Laak, J.A., Hermsen, M., Manson, Q.F., Balkenhol, M., 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) 8

4. Bergner, B., Lippert, C., Mahendran, A.: Iterative patch selection for high-resolution image recognition. arXiv preprint arXiv:2210.13007 (2022) 4, 20

5. Bontempo, G., Bolelli, F., Porrello, A., Calderara, S., Ficarra, E.: A graph-based multi-scale approach with knowledge distillation for wsi classification. TMI (2023) 3

6. Brancati, N., Anniciello, A.M., Pati, P., Riccio, D., Scognamiglio, G., Jaume, G., De Pietro, G., Di Bonito, M., Foncubierta, A., Botti, G., et al.: Bracs: A dataset for breast carcinoma subtyping in h&e histology images. Database **2022**, baac093 (2022) 8

7. Campanella, G., Hanna, M.G., Geneslaw, L., Miraflor, A., Werneck Krauss Silva, V., Busam, K.J., Brogi, E., Reuter, V.E., Klimstra, D.S., Fuchs, T.J.: Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nature medicine **25**(8), 1301–1309 (2019) 3, 4

8. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9650–9660 (2021) 9

9. Chan, T.H., Cendra, F.J., Ma, L., Yin, G., Yu, L.: Histopathology whole slide image analysis with heterogeneous graph representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15661–15670 (2023) 4

10. Chen, R.J., Chen, C., Li, Y., Chen, T.Y., Trister, A.D., Krishnan, R.G., Mahmood, F.: Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16144–16155 (2022) 3, 7, 20

11. Chen, R.J., Lu, M.Y., Weng, W.H., Chen, T.Y., Williamson, D.F., Manz, T., Shady, M., Mahmood, F.: Multimodal co-attention transformer for survival prediction in gigapixel whole slide images. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4015–4025 (2021) 1

12. Chen, Y.C., Lu, C.S.: Rankmix: Data augmentation for weakly supervised learning of classifying whole slide images with diverse sizes and imbalanced categories. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23936–23945 (2023) 4

13. Chikontwe, P., Kim, M., Nam, S.J., Go, H., Park, S.H.: Multiple instance learning with center embeddings for histopathology classification. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part V 23. pp. 519–528. Springer (2020) 2, 4

14. Cornish, T.C., Swapp, R.E., Kaplan, K.J.: Whole-slide imaging: routine pathologic diagnosis. Advances in anatomic pathology **19**(3), 152–159 (2012) 1

15. Dauphin, Y.N., Fan, A., Auli, M., Grangier, D.: Language modeling with gated convolutional networks. In: International conference on machine learning. pp. 933–941. PMLR (2017) 5, 20

16. Dehaene, O., Camara, A., Moindrot, O., de Lavergne, A., Courtiol, P.: Self-supervision closes the gap between weak and strong supervision in histology. arXiv preprint arXiv:2012.03583 (2020) 2, 3

17. DeVries, T., Taylor, G.W.: Improved regularization of convolutional neural networks with cutout. arXiv preprint arXiv:1708.04552 (2017) 2, 4, 8

18. Dietterich, T.G., Lathrop, R.H., Lozano-Pérez, T.: Solving the multiple instance problem with axis-parallel rectangles. Artificial intelligence **89**(1-2), 31–71 (1997) 1

19. Geirhos, R., Jacobsen, J.H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., Wichmann, F.A.: Shortcut learning in deep neural networks. Nature Machine Intelligence **2**(11), 665–673 (2020) 5

20. Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F.A., Brendel, W.: Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. arXiv preprint arXiv:1811.12231 (2018) 5

21. Guan, Y., Zhang, J., Tian, K., Yang, S., Dong, P., Xiang, J., Yang, W., Huang, J., Zhang, Y., Han, X.: Node-aligned graph convolutional network for whole-slide image representation and classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18813–18823 (2022) 2, 4

22. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016) 9

23. He, L., Long, L.R., Antani, S., Thoma, G.R.: Histology image analysis for carcinoma detection and grading. Computer methods and programs in biomedicine **107**(3), 538–556 (2012) 1

24. Holdenried-Krafft, S., Somers, P., Montes-Majarro, I.A., Silimon, D., Tarín, C., Fend, F., Lensch, H.: Dual-query multiple instance learning for dynamic meta-embedding based tumor classification. arXiv preprint arXiv:2307.07482 (2023) 2, 3

25. Hou, W., Yu, L., Lin, C., Huang, H., Yu, R., Qin, J., Wang, L.: Hˆ 2-mil: Exploring hierarchical representation with heterogeneous multiple instance learning for whole slide image analysis. In: AAAI. vol. 36, pp. 933–941 (2022) 3

26. Huang, Z., Wang, H., Xing, E.P., Huang, D.: Self-challenging improves cross-domain generalization. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16. pp. 124–140. Springer (2020) 2, 4

27. Ilse, M., Tomczak, J., Welling, M.: Attention-based deep multiple instance learning. In: International conference on machine learning. pp. 2127–2136. PMLR (2018) 3, 5, 9, 10, 19, 20, 21

28. Kang, M., Song, H., Park, S., Yoo, D., Pereira, S.: Benchmarking self-supervised learning on diverse pathology datasets. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3344–3354 (2023) 9

29. Kong, F., Henao, R.: Efficient classification of very large images with tiny objects. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2384–2394 (2022) 4

30. Li, B., Li, Y., Eliceiri, K.W.: Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In: Pro-

ceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14318–14328 (2021) 2, 3, 8, 9

31. Li, H., Zhu, C., Zhang, Y., Sun, Y., Shui, Z., Kuang, W., Zheng, S., Yang, L.: Task-specific fine-tuning via variational information bottleneck for weakly-supervised pathology whole slide image classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7454–7463 (2023) 2, 3, 9

32. Li, R., Yao, J., Zhu, X., Li, Y., Huang, J.: Graph cnn for survival analysis on whole slide pathological images. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 174–182. Springer (2018) 1

33. Li, Y., Ping, W.: Cancer metastasis detection with neural conditional random field. arXiv preprint arXiv:1806.07064 (2018) 1

34. Lin, T., Yu, Z., Hu, H., Xu, Y., Chen, C.W.: Interventional bag multi-instance learning on whole-slide pathological images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19830–19839 (2023) 1, 2, 9

35. Litjens, G., Sánchez, C.I., Timofeeva, N., Hermsen, M., Nagtegaal, I., Kovacs, I., Hulsbergen-Van De Kaa, C., Bult, P., Van Ginneken, B., Van Der Laak, J.: Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis. Scientific reports **6**(1), 26286 (2016) 1

36. Lu, M.Y., Williamson, D.F., Chen, T.Y., Chen, R.J., Barbieri, M., Mahmood, F.: Data-efficient and weakly supervised computational pathology on whole-slide images. Nature biomedical engineering **5**(6), 555–570 (2021) 2, 3, 9, 19, 20

37. Madabhushi, A.: Digital pathology image analysis: opportunities and challenges. Imaging in medicine **1**(1), 7 (2009) 1

38. Maron, O., Lozano-Pérez, T.: A framework for multiple-instance learning. Advances in neural information processing systems **10** (1997) 1

39. McInnes, L., Healy, J., Melville, J.: Umap: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:1802.03426 (2018) 5, 12, 13, 24, 25

40. Pantanowitz, L., Valenstein, P.N., Evans, A.J., Kaplan, K.J., Pfeifer, J.D., Wilbur, D.C., Collins, L.C., Colgan, T.J.: Review of the current state of whole slide imaging in pathology. Journal of pathology informatics **2**(1), 36 (2011) 1

41. Pinckaers, H., Van Ginneken, B., Litjens, G.: Streaming convolutional neural networks for end-to-end learning with multi-megapixel images. IEEE transactions on pattern analysis and machine intelligence **44**(3), 1581–1590 (2020) 1

42. Qu, L., Wang, M., Song, Z., et al.: Bi-directional weakly supervised knowledge distillation for whole slide image classification. Neurips **35**, 15368–15381 (2022) 2, 4, 8, 24, 25

43. Qu, L., Yang, Z., Duan, M., Ma, Y., Wang, S., Wang, M., Song, Z.: Boosting whole slide image classification from the perspectives of distribution, correlation and magnification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 21463–21473 (2023) 4

44. Rosenberg, A., Hirschberg, J.: V-measure: A conditional entropy-based external cluster evaluation measure. In: Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL). pp. 410–420 (2007) 12, 13

45. Shao, Z., Bian, H., Chen, Y., Wang, Y., Zhang, J., Ji, X., et al.: Transmil: Transformer based correlated multiple instance learning for whole slide image classification. Advances in neural information processing systems **34**, 2136–2147 (2021) 2, 3, 4, 9, 20

46. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. The journal of machine learning research **15**(1), 1929–1958 (2014) 8

47. Tang, W., Huang, S., Zhang, X., Zhou, F., Zhang, Y., Liu, B.: Multiple instance learning framework with masked hard instance mining for whole slide image classification. arXiv preprint arXiv:2307.15254 (2023) 1, 2, 4, 8, 9, 23, 24, 25

48. Tellez, D., Litjens, G., van der Laak, J., Ciompi, F.: Neural image compression for gigapixel histopathology image analysis. IEEE transactions on pattern analysis and machine intelligence **43**(2), 567–578 (2019) 3

49. Tiwari, R., Shenoy, P.: Overcoming simplicity bias in deep networks using a feature sieve (2023) 2

50. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017) 20

51. Wang, D., Khosla, A., Gargeya, R., Irshad, H., Beck, A.H.: Deep learning for identifying metastatic breast cancer. arXiv preprint arXiv:1606.05718 (2016) 1

52. Wang, H., Luo, L., Wang, F., Tong, R., Chen, Y.W., Hu, H., Lin, L., Chen, H.: Iteratively coupled multiple instance learning from instance to bag classifier for whole slide image classification. arXiv preprint arXiv:2303.15749 (2023) 3

53. Wang, X., Xiang, J., Zhang, J., Yang, S., Yang, Z., Wang, M.H., Zhang, J., Yang, W., Huang, J., Han, X.: Scl-wc: Cross-slide contrastive learning for weakly-supervised whole-slide image classification. Advances in neural information processing systems **35**, 18009–18021 (2022) 2, 3

54. Wang, Y., Kartasalo, K., Weitz, P., Acs, B., Valkonen, M., Larsson, C., Ruusuvuori, P., Hartman, J., Rantalainen, M.: Predicting molecular phenotypes from histopathology images: A transcriptome-wide expression–morphology analysis in breast cancer. Cancer research **81**(19), 5115–5126 (2021) 1

55. Xiong, C., Chen, H., Sung, J., King, I.: Diagnose like a pathologist: Transformer-enabled hierarchical attention-guided multiple instance learning for whole slide image classification. arXiv preprint arXiv:2301.08125 (2023) 4

56. Yang, J., Chen, H., Zhao, Y., Yang, F., Zhang, Y., He, L., Yao, J.: Remix: A general and efficient framework for multiple instance learning based whole slide image classification. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 35–45. Springer (2022) 2, 4

57. Yao, J., Zhu, X., Jonnagaddala, J., Hawkins, N., Huang, J.: Whole slide images based cancer survival prediction using attention guided deep multiple instance learning networks. Medical Image Analysis **65**, 101789 (2020) 1

58. Yufei, C., Liu, Z., Liu, X., Liu, X., Wang, C., Kuo, T.W., Xue, C.J., Chan, A.B.: Bayes-mil: A new probabilistic perspective on attention-based multiple instance learning for whole slide images. In: The Eleventh International Conference on Learning Representations (2022) 2

59. Zhang, H., Meng, Y., Zhao, Y., Qiao, Y., Yang, X., Coupland, S.E., Zheng, Y.: 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. pp. 18802–18812 (2022) 1, 2, 4, 8, 9, 20

60. Zhang, Y., Sun, Y., Li, H., Zheng, S., Zhu, C., Yang, L.: Benchmarking the robustness of deep neural networks to common corruptions in digital pathology. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 242–252. Springer (2022) 2

61. Zhao, Y., Yang, F., Fang, Y., Liu, H., Zhou, N., Zhang, J., Sun, J., Yang, S., Menze, B., Fan, X., et al.: Predicting lymph node metastasis using histopathological images based on multiple instance learning with deep graph convolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4837–4846 (2020) 4

62. Zhong, Z., Zheng, L., Kang, G., Li, S., Yang, Y.: Random erasing data augmentation. In: Proceedings of the AAAI conference on artificial intelligence. vol. 34, pp. 13001–13008 (2020) 4, 8

63. Zhu, X., Yao, J., Zhu, F., Huang, J.: Wsisa: Making survival prediction from whole slide histopathological images. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7234–7242 (2017) 1

6   Overview

In this appendix, we provide valuable resources and insights, including the source code (Sec. 7), implementation details (Sec. 8), additional experimental results (Sec. 9), and a discussion on limitations (Sec. 10).

7   Source Code

The source code of ACMIL is available at https://github.com/dazhangyu123/ACMIL. For further information on the environment setup and experiment execution, please refer to README.md. The implementation of ACMIL is based on the source code of ABMIL [27] and CLAM [36].

8   Implementation Details

Data Pre-processing. We adopt the data pre-processing method from CLAM [36], which involves threshold segmentation and filtering to locate tissue regions in each whole-slide image (WSI). From these regions, we extract non-overlapping patches of size 256×256 at a magnification of $\times 20$ for Camelyon16 and LBC datasets, and at a magnification of $\times 10$ for BRACS.

Feature Extraction. Given that ACMIL freezes the feature extractor during training, we extract and save features with 512 dimensions for ResNet-18 and 384 dimensions for ViT-S/16 to conserve space and expedite computation.

Model Architecture. The learnable components of the model include one fully-connected layer to reduce features to 256 dimensions for ResNet-18 and 128 dimensions for ViT-S/16, a gated attention network, and a fully-connected layer for making predictions.

Training. All models are trained for 100 epochs using a cosine learning rate decay starting at 0.0001 for ViT-S/16 and 0.0002 for ResNet-18. We employ an Adam optimizer with a weight decay of 0.0001, and the batch size is set to 1.

Hyperparameters. For the setting of Camelyon16 and natural supervised pre-training, we set hyperparameters as $M = 2, K = 10, p = 0.6$. For the other situation, we set hyperparameters as $M = 5, K = 10, p = 0.6$.

Table 4: The performance comparison between the baseline and our ACMIL across two attention mechanisms (i.e., gated attention (GA) and multiple head attention (MHA)), three datasets, and two pretrained methods.

Performance / Method	CAMELYON-16		BRACS		LBC		Average
	F1-score	AUC	F1-score	AUC	F1-score	AUC	
ResNet18 ImageNet pretrained							
GA	0.757±0.020	0.790±0.027	0.523±0.028	0.723±0.035	0.465±0.040	0.798±0.013	0.676
+ACMIL	0.798±0.029	0.841±0.030	0.552±0.048	0.754±0.008	0.546±0.028	0.821±0.015	0.719
Δ(%)	+4.1	+5.1	+2.9	+3.1	+8.1	+2.3	+4.3
MHA	0.752±0.030	0.775±0.027	0.502±0.039	0.738±0.019	0.531±0.025	0.817±0.011	0.686
+ACMIL	0.799±0.018	0.875±0.017	0.541±0.063	0.723±0.028	0.555±0.038	0.818±0.012	0.719
Δ(%)	+4.7	+10.0	+3.9	-1.5	+2.4	+0.1	+3.3
ViT-S/16 SSL pretrained							
GA	0.914±0.031	0.945±0.027	0.680±0.051	0.866±0.029	0.595±0.036	0.831±0.022	0.805
+ACMIL	0.954±0.012	0.974±0.012	0.722±0.030	0.888±0.010	0.662±0.043	0.901±0.011	0.850
Δ(%)	+4.0	+2.9	+4.2	+2.2	+6.7	+7.0	+4.5
MHA	0.931±0.032	0.961±0.017	0.656±0.030	0.850±0.030	0.619±0.032	0.864±0.013	0.813
+ACMIL	0.936±0.027	0.973±0.014	0.667±0.059	0.879±0.028	0.649±0.024	0.876±0.012	0.830
Δ(%)	+0.5	+1.2	+1.1	+2.9	+3.0	+1.2	+1.7

9   More Experimental Results

The additional experimental results include performance comparison against baselines (Sec. 9.1), visualization of validation metrics (Sec. 9.2), additional heatmap visualizations (Sec. 9.3), UMAP visualization of normal instances (Sec. 9.4), performance comparison between different masking strategies (Sec. 9.5), and discussion about computational cost (Sec. 9.6).

9.1   Performance Evaluation against Baselines

To assess the adaptability of our ACMIL to different attention mechanisms, we selected two prominent attention mechanisms as our baselines. The first is the gated attention (GA) mechanism [15], employed in approaches like ABMIL [27], CLAM [36], and DTFD-MIL [59]. The second is the multiple head attention (MHA) mechanism [50], utilized in methods such as TransMIL [45] and IPS transformer [4]. The results are presented in Table 4.

With GA as the baseline, ACMIL exhibits a substantial and comprehensive improvement in performance. All 12 performance metrics show enhancements, with an average gain of 4.4 points, a minimum increase of 2.2 points, and a maximum improvement of 8.1 points.

With MHA as the baseline, ACMIL also demonstrates performance improvements in the majority of terms (i.e., 11 out of 12 terms), achieving an average improvement of 2.5 points. In comparison to GA, MHA introduces parallel processing (i.e., heads). This modification enables the learning of different visual concepts across heads [10], contributing to a slight attenuation in the improvements brought by ACMIL.
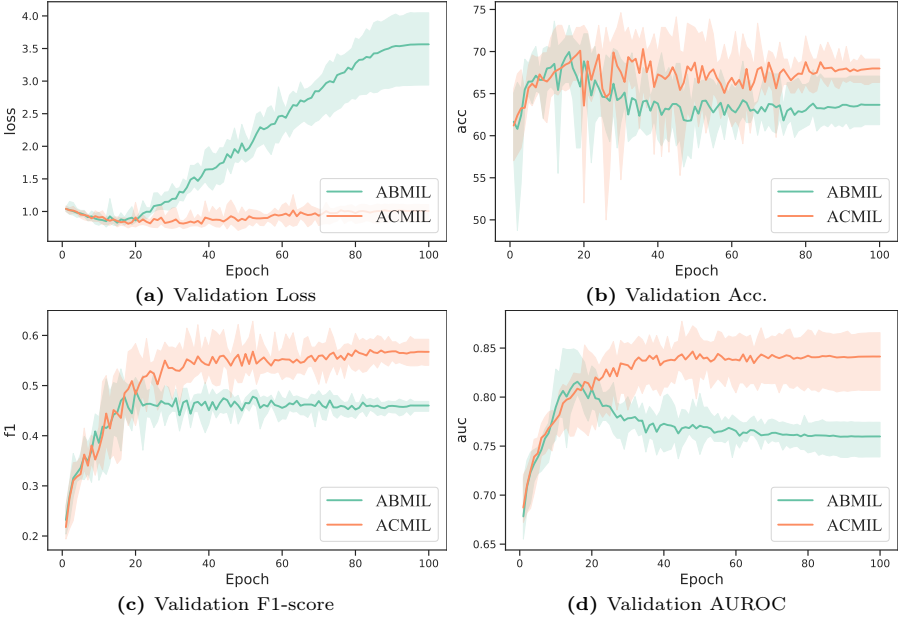
Fig. 11: Performance comparison between ABMIL [27] and our ACMIL on LBC validation set throughout the training process. ABMIL displays pronounced signs of overfitting, as indicated by a significant increase in validation loss and a decline in the other three evaluation metrics. Conversely, ACMIL effectively mitigates the overfitting issue.

9.2   Visualization of validation metrics across training epochs

Fig. 11 depicts validation metrics across training epochs. ABMIL [27], one of the most commonly-used MIL methods, shows significant overfitting since loss drastically increases and validation metrics significantly decrease as the training processes. As a comparison, ACMIL suppresses the increase of validation loss throughout the training process and the decrease of the other three evaluation metrics. As a result, ACMIL indeed effectively mitigates the overfitting issue.

9.3   More heatmap visualization

Heatmap visualizations of five attention branches in MBA. In Fig. 12, we present the heatmap visualizations for five attention branches and delve into the effects of these distinct branches. We've chosen two test slides in Camelyon16 for this analysis, including one tumor slide and one normal slide. For the tumor slide, we observe that all five branches capture the cancerous instances. Notably, the third and fifth branches successfully capture the entirety of the tumor regions, while the remaining three branches only manage to capture a subset of the tumor regions. Additionally, the third branch activates the adipose, and the fifth branch activates the lymphocyte regions. Overall, the averaged heatmap captures the
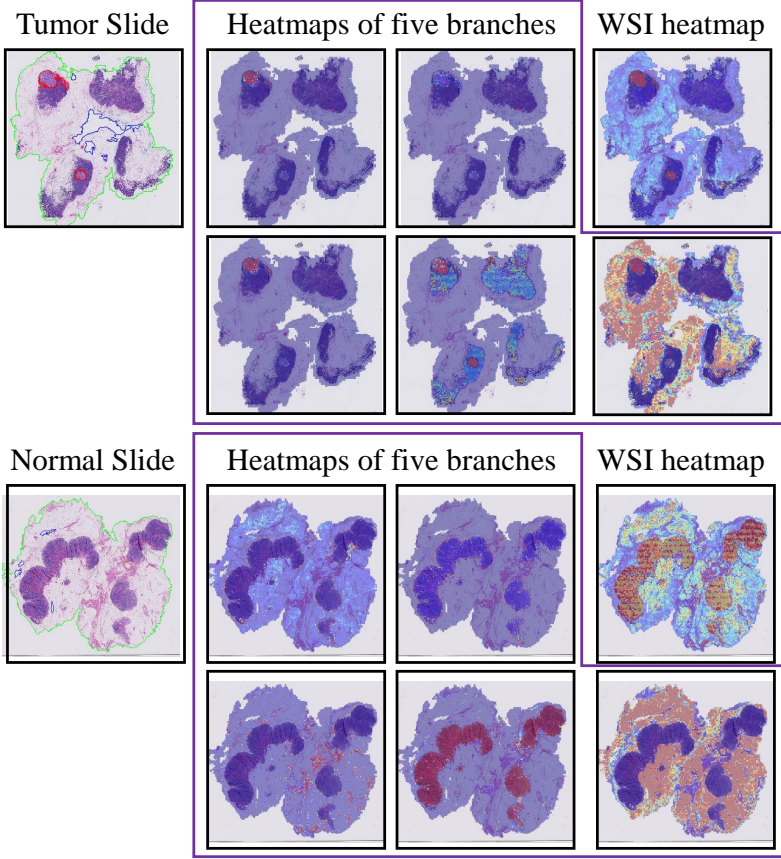
Fig. 12: Heatmap visualizations for five attention branches. Different branches specialize in capturing specific features, contributing to better interpretability for the bag (final) heatmap.

whole tumor regions, along with slightly activating some normal regions. For the normal slide, the first two branches activate instances lying between adipose and lymphocyte regions. The third branch predominantly activates adipose tissue, the fourth branch emphasizes muscle regions, and the fifth branch highlights lymphocyte regions. Overall, the averaged heatmap activates all normal regions. This analysis illustrates how the different branches specialize in capturing specific features, contributing to a more comprehensive understanding of the data.

Heatmap visualizations with bad interpretability. In Fig. 13, we present three cases (i.e., two tumor slides and one normal slide) with heatmap visualizations that exhibit poor interpretability. The first slide is a tumor slide. While ACMIL activates a greater number of cancerous instances than ABMIL (as indicated by the yellow box), it also activates some normal instances (visible in the green box). This mixed activation can potentially mislead experts during practi-
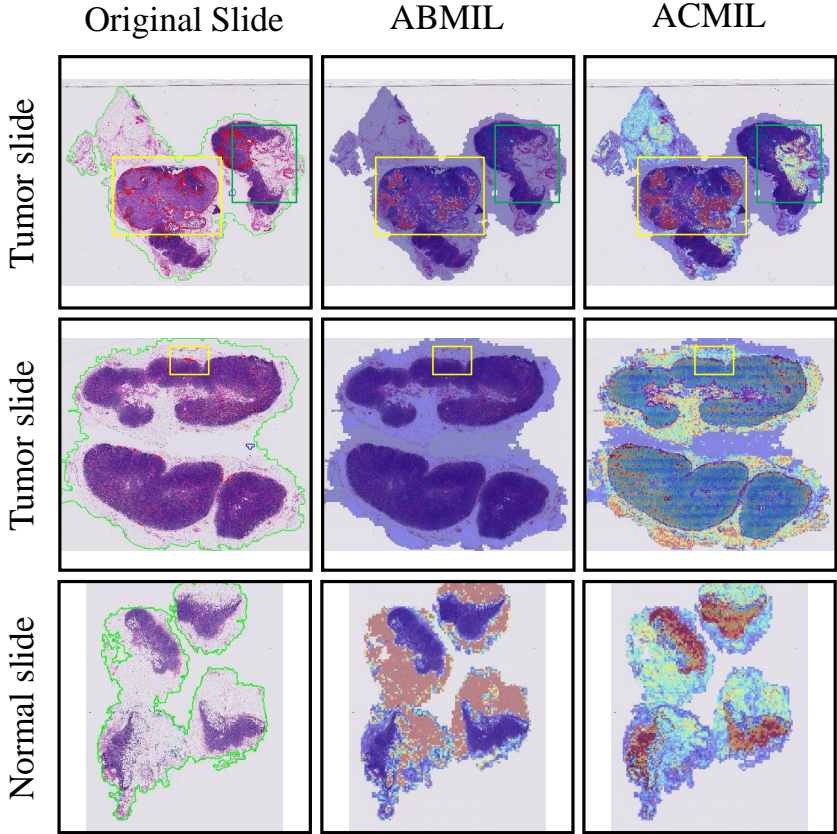
Fig. 13: Heatmap visualizations with bad interpretability. Three cases indicate that ACMIL's approach of assigning broader attention values to a wide range of predictive instances doesn't consistently enhance interpretability.

cal interpretability analysis. The second instance also concerns a tumor case but with small tumor regions. ABMIL accurately localizes the tumor regions (see yellow box). In contrast, ACMIL allocates more attention values to a broader range of predictive instances, which results in an inability to precisely locate the tumor regions. The third case pertains to a normal slide. In contrast to AB-MIL, which provides misleading interpretability by predominantly focusing on adipose tissue, ACMIL assigns excessive attention values to lymphocyte regions. Consequently, the heatmap primarily highlights lymphocyte tissue instead of the expected comprehensive representation of normal tissue.

Comparison of heatmap visualization between MHIM-MIL [47] and ACMIL. In Fig. 14, we present the heatmap visualizations of MHIM-MIL and ACMIL. For the tumor slide (first row), MHIM-MIL and ACMIL both capture all cancerous instances in the tumor region, but MHIM-MIL activates more normal instances than ACMIL. For the normal slide (second row), MHIM-MIL
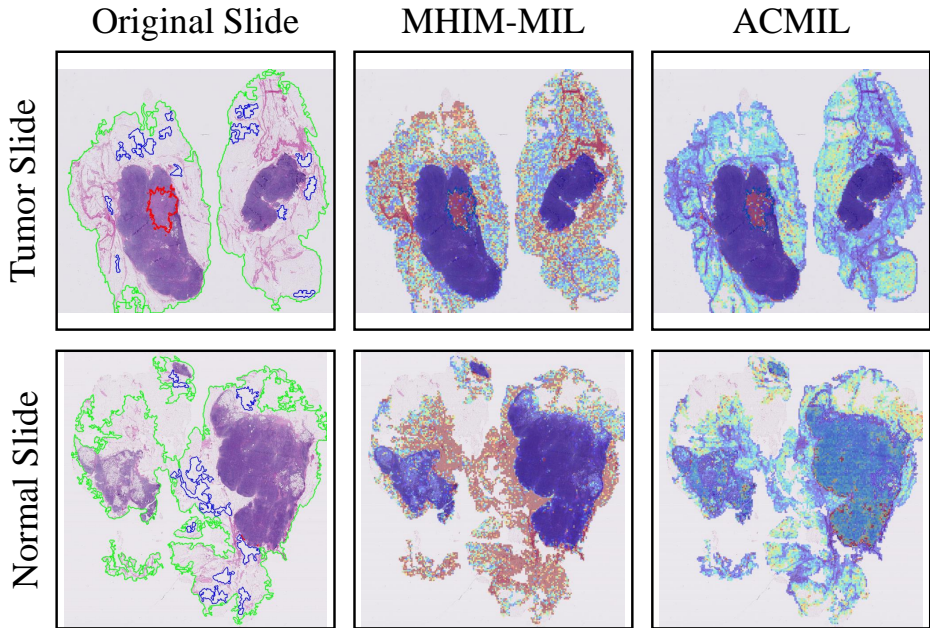
Fig. 14: Comparison of heatmap visualizations between MHIM-MIL [47] and ACMIL (Zoom in for best view). ACMIL performs better in capturing comprehensive predictive instances.

predominately activates adipose, whereas ACMIL activates all normal instances more uniformly.

9.4   Instance Feature Analysis for Normal Slide

In Fig. 15, we present the UMAP visualization [39] of normal instance features in a typical Camelyon case, 'test_016'. The comparison between ABMIL and ACMIL is quite evident. ABMIL, with a single attention branch, activates only a fraction of normal instances. Conversely, ACMIL utilizes five branches, with each branch specializing in capturing specific patterns, resulting in the activation of nearly all normal instances. This observation demonstrates the superior ability of ACMIL to encompass a broader range of patterns in the data.

9.5   Performance Comparison between Different Masking Strategies

In Tab. 5, we present the performance comparison between three strategies used in STKIM, WENO, and MHIM-MIL. Specifically, our STKIM masks 10 instances with a probability of 0.6 by default. WENO [42] masks 95 instances with a probability of 1.0 by default. MHIM-MIL [47] masks 1% instances with a probability
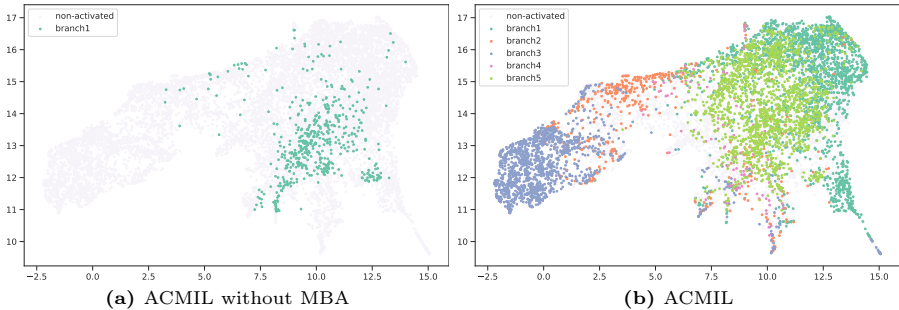
(a) ACMIL without MBA   **(b)** ACMIL

Fig. 15: UMAP visualization [39] of instance features in a normal case, Camelyon16 'test_016'. The normal instances exhibit distinct patterns, making it challenging for a single-branch model like ABMIL to capture them comprehensively. Consequently, ABMIL may overlook certain instances. In contrast, our ACMIL leverages multiple branches, each adept at capturing specific patterns, enabling ACMIL to activate a greater number of normal instances. Note that the instance is considered active when its attention value surpasses $\frac{1}{N}$.

Table 5: The performance comparison between different masking strategies on SSL pre-trained embedding. Strategy1 denotes the default masking strategy in STKIM, i.e., masking 10 instances with a probability of 0.6. Strategy2 denotes the default masking strategy in WENO [42], i.e., masking 95 instances with a probability of 1.0. Strategy3 denotes the default masking strategy in MHIM-MIL [47], i.e., masking 1% instances with a probability of 0.5. Strategy 1 outperforms the other two strategies.

Performance Method	CAMELYON-16		BRACS		LBC	
	F1-score	AUC	F1-score	AUC	F1-score	AUC
Strategy1	0.954	0.974	0.722	0.888	0.662	0.901
Strategy2	0.741(-0.213)	0.843(-0.131)	0.521(-0.201)	0.844(-0.044)	0.502(-0.160)	0.784(-0.117)
Strategy3	0.916(-0.038)	0.967(-0.007)	0.694(-0.028)	0.881(-0.007)	0.659(-0.003)	0.892(-0.009)

of 0.5. We find that the masking strategy of our STKIM performs best, significantly suppressing the strategy in WENO and slightly surpassing the strategy in MHIM-MIL.

9.6   Computational Cost

STKIM and MHIM-MIL [47]. We conducted a comprehensive comparison between STKIM and MHIM-MIL, focusing on computational cost and performance, as detailed in Tab. 6. For the computational cost, STKIM demonstrates nearly identical training time consumption and GPU memory usage as the baseline, ABMIL. This similarity arises because STKIM primarily integrates a sorting algorithm, which does not substantially increase resource requirements. On the other hand, MHIM-MIL introduces a teacher model while requiring two forward propagations, leading to significantly higher GPU memory usage and training time consumption. Due to the masking operator being discarded in the

Table 6: Comparison of performance and computational cost requirements between MHIM-MIL and STKIM. We report the AUC, FLOPs, training time per epoch (Time), and peak memory usage (Mem.) on the CAMELYON-16 (C16) dataset. The flops are measured with the number of instances of a bag being 1024.

Model	C16	BRACS	LBC	FLOPs	Time	Mem.
ResNet18 ImageNet pretrained						
ABMIL	**0.790**	0.723	0.798	**201M**	**8.0s**	**0.3G**
MHIM-MIL	0.772	0.774	0.816	**201M**	20.8s	1.9G
STKIM	0.779	**0.789**	**0.820**	201M	**8.0s**	**0.3G**
ViT-S/16 SSL pretrained						
ABMIL	0.945	0.866	0.831	**84M**	**6.4s**	**0.2G**
MHIM-MIL	**0.970**	0.865	**0.872**	84M	16.8s	1.0G
STKIM	0.968	**0.873**	0.856	**84M**	6.5s	**0.2G**

Table 7: Comparison of performance, time, and memory requirements between ABMIL and MBA. We report the auc, the FLOPs, the training time per epoch (Time), and the peak memory usage (Mem.) on the CAMELYON-16 dataset (C16). The flops are measured with the number of instances of a bag being 1024.

Model	C16	BRACS	LBC	FLOPs	Time	Mem.
ResNet18 ImageNet pretrained						
ABMIL	0.790	0.723	0.798	**201M**	**8.0s**	**0.3G**
+MBA	**0.850**	**0.797**	**0.818**	202M	11.6s	**0.3G**
ViT-S/16 SSL pretrained						
ABMIL	0.945	0.866	0.831	**84M**	**6.4s**	**0.2G**
+MBA	**0.973**	**0.878**	**0.875**	85M	9.3s	**0.2G**

evaluation, STKIM and MHIM-MIL keep the same evaluation cost (FLOPs) as ABMIL. For the performance, STKIM delivers comparable results to MHIM-MIL across three datasets and with two pretrained backbone models. Notably, STKIM outperforms MHIM-MIL in four out of six performance metrics while lagging behind in the remaining two.

MBA. In Tab. 7, we present the comparison of performance and computational cost between ABMIL and MBA. Notably, MBA demonstrates a substantial performance improvement over ABMIL. Meanwhile, due to introducing a small number of parameters, the FLOPs, and Memory cost increases marginally. Otherwise, the inclusion of the newly introduced diversity loss leads to a notable increase in time cost.

10   Limitations

Although ACMIL enhances the generalization ability and interpretability of MIL methods in WSI analysis, certain limitations necessitate further exploration. Firstly, the selection of hyperparameters M and K significantly impacts performance, and the optimal choice depends on the dataset, requiring practitioners to determine the best value through trial and error. In the future, how to simplify the framework should be considered. Secondly, our paper does not account for the correlation between instances, which is crucial for understanding the complex tumor structure. This aspect will be a focus of future investigations.