Focus Your Attention: Multiple Instance Learning with Attention Modification for Whole Slide Pathological Image Classification

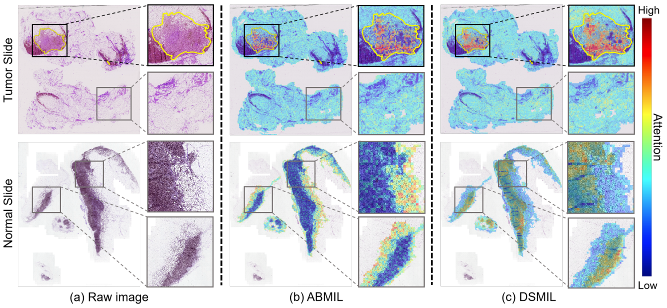
Hailun Cheng, Shenjin Huang, Linghan Cai, Yangfan Xu, Runming Wang, and Yongbing Zhang

***Abstract*—Computer-aided pathology diagnosis based on whole slide images, which is often formulated as a weakly supervised multiple instance learning (MIL) paradigm. Current approaches generally employ attention mechanisms to aggregate instance-level features. However, the weakly supervised signal and the imbalanced instance distribution often lead to inaccurate attention localization, compromising the performance and generalization capability of the MIL framework. To address these problems, this paper presents a novel MIL framework called FAMIL that focuses on inaccurate attention and refines them. FAMIL adopts a dual-branch structure and incorporates two innovative online data augmentation strategies: attention-based Mixup (ABMix) and attention-based Masking (ABMask). ABMix emphasizes the significance of positive instances, generalizing Mixup in the MIL scenarios, while ABMask flexibly identifies challenging positive instances to optimize the feature representation. Moreover, these two methods are plug-and-play and can be easily embedded into attention-based MIL methods. Extensive experiments on three public benchmarks demonstrate the superiority of our FAMIL, outperforming current state-of-the-art methods. The test AUC for the binary tumor classification can be up to 92.61% over CAMELYON16. And the AUC over the cancer subtype classification can be up to 93.81% and 98.41% on TCGA-NSCLC and TCGA-RCC datasets, respectively.**

***Index Terms*—Pathological image classification, multiple instance learning, data augmentation, attention-based Mixup, attention-based masking**

I. Introduction

PATHOLOGICAL image analysis is regarded as the gold standard for therapy decision and cancer prognosis [1]–[4]. With the advancement of scanning technology, traditional tissue specimens are increasingly transformed into digital whole slide images (WSIs), which enable computer-assisted diagnosis. However, the huge number of pixels per WSI (e.g., $40,000 \times 40,000$ pixels) and the lack of fine-grained annotations (patch-level labels) pose significant challenges for the direct application of deep learning in pathological image analysis. To address these challenges, multiple instance learning (MIL) [5]–[11] has been widely adopted. In MIL, each WSI is treated as a bag that contains thousands of instances (tile patches) extracted from the WSI. A bag is labeled as positive if at least one instance is positive, otherwise it is negative. The MIL-based method requires only WSI-level labels, thus significantly reducing the data annotation burden.



(a) Raw image (b) ABMIL (c) DSMIL

Fig. 1. Two types of inaccurate attention localization in attention-based MIL. The yellow curve is the ground truth of the tumor region. Black boxes denote that low attention erroneously locates the tumor area, and gray boxes indicate that high attention incorrectly locates the non-tumor area.

The application of MIL in pathology analysis aims to solve two primary problems, namely WSI classification and tumor localization [12]–[15]. Among MIL methods, a popular solution is ABMIL [6], which obtains the contribution of each instance in bag aggregation through attention scoring. Based on the attention mechanism, ABMIL can identify areas with tumors, thus better executing the classification task. Currently, pathological image analysis has flourished by the variants of ABMIL. DSMIL [7] utilizes attention to adjust multi-scale features, improving the accuracy of tumor localization. TransMIL [9] introduces the correlation among instances for better bag classification. These studies generally believe that the attention mechanism provides interpretability in WSI classification. Ideally, for positive WSIs, instances with high attention scores are positive instances, and instances with low attention scores are negative instances; while for negative WSIs, the attention scores should be evenly distributed and consistently low across instances.

However, the attention-based methods often make mistakes in tumor localization, as illustrated in Fig. 1. The mistakes can be divided into two categories: (1) high attention locates

This work was supported in part by the National Natural Science Foundation of China under Grant 62331011 and the School Interdisciplinary Research Fund, Tsinghua Shenzhen International Graduate School, under Grant 01020200063 & 20225820246. (*Corresponding authors: Yongbing Zhang and Runming Wang.*)

HL Cheng, YF Xu, and RM Wang are with the Institute of Biopharmaceutical and Health Engineering, Shenzhen International Graduate School, Tsinghua University, Shenzhen, 518055, China. (e-mails: chl22@mails.tsinghua.edu.cn; xuyf22@mails.tsinghua.edu.cn; runmingwang@sz.tsinghua.edu.cn).

SJ Huang is with the Faculty of Computing, Harbin Institute of Technology, Harbin, 150001, China. (e-mails: shenjinhuang@stu.hit.edu.cn).

LH Cai and YB Zhang are with the School of Computer Science and Technology, Harbin Institute of Technology, Shenzhen, 518055, China. (e-mails: cailh@stu.hit.edu.cn; ybzhang08@hit.edu.cn).

HL Cheng and SJ Huang contribute equally to this paper.

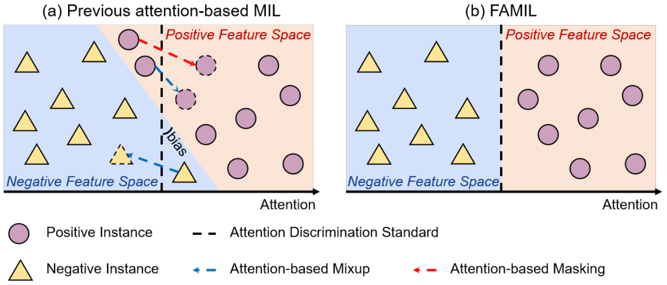


Fig. 2. The relation between instance decision boundary and attention. In (a), a bias exists in current attention-based MIL frameworks. In (b), FAMIL improves the attention distribution, achieving consistency between attention discrimination standard and instance decision boundary.

non-tumor areas, and (2) low attention locates the tumor area. These inhibit the performance potential of attention-based MIL methods and harm the model's interpretability. In response to the above problems, Lin et al. [16] argue that these phenomena are caused by color distribution, and introduce causal inference into the MIL to improve feature representation. To improve the localization accuracy of attention, Tourniaire et al. [17] propose a subtle loss function to regulate attention distribution. However, this method requires instance-level annotations for positive WSIs, making it challenging to implement for weakly supervised WSI classification. Owing to the imbalance in the number of positive and negative instances, we argue that these two mistakes of attention localization arise from insufficient tumor feature learning, leading to the mismatching between attention discrimination standard and instance decision boundary, as shown in Fig. 2 (a).

As an effective solution for enhancing the feature learning of the model, data augmentation has been widely studied in natural scenarios [18]–[21]. In pathological image analysis, owing to the large number of instances in a bag, the image-based operations inevitably bring about large computational costs [22]. Therefore, existing works usually adopt feature-level augmentation such as Mixup to improve MIL performance [23]–[25]. For example, Gadermayr et al. [24] propose a multilinear intra-slide interpolation Mixup to improve the accuracy of the MIL model; Chen et al. [23] introduce instance-level pseudo-labeling and ranking into Mixup to solve the problem of insufficient training data and imbalanced classification. However, these methods treat positive and negative instances equally, failing to emphasize the significance of positive instances in the MIL paradigm. Moreover, they necessitate alignment operations, which require that the number of instances in the two bags to be mixed is the same. These oversights render them ineffective in addressing the issue of inaccurate attention caused by the weak supervision signal.

To this end, this paper focuses on inaccurate attention and presents a novel multiple instance learning framework named FAMIL for WSI classification. Specifically, FAMIL includes two online data augmentation strategies to modify inaccurate attention distribution, as shown in Fig. 2. To decrease the high attention to non-tumor regions, we propose an attention-based Mixup which improves the variability of the bag for enhancing the instance discriminative power of the model. Additionally, an attention-based masking is developed to precisely locate tumor areas. The main idea of attention-based masking is to discard salient (easy-to-distinguish) positive instances, thus forcing the model to learn hard ones. Overall, the contributions of this paper can be summarized as follows:

- This paper focuses on unexpected attention in the MIL framework and proposes a FAMIL to effectively enhance the feature representation of instances, modifying inaccurate attention.
- We propose ABMix, a novel data augmentation technique built upon the principles of Mixup. Unlike traditional Mixup methods, ABMix does not require any size and semantic alignment. By highlighting the importance of positive examples, ABMix achieves a more accurate attention localization and model performance.
- To further enhance the network's ability to identify critical features, we introduce ABMask, a data augmentation approach tailored to encourage hard positive instance mining flexibly. By guiding the model to actively discover and emphasize challenging positive samples, ABMask achieves more accurate instance localization capabilities.
- Extensive experiments on three datasets demonstrate the superiority of our FAMIL, with state-of-the-art results. Furthermore, our data augmentation strategies can be easily applied in attention-based MIL for better performance.

The rest of this paper is organized as follows. In Section II, we review the MIL for WSI classification and related data augmentation techniques. Next, we present FAMIL in Section III. Extensive experiments and analyses are illustrated in Section IV. Section V and Section VI show the discussion and conclusion of this work.

II. Related Work

A. MIL for WSI Classification

MIL [26] has been extensively explored in the WSI classification task, which is a weakly supervised learning paradigm that utilizes bag-level labels rather than instance-level labels for training. Previous algorithms can be mainly classified into two categories: The first one is the instance-based MIL frameworks [26]–[31], utilizing instance-level pseudo-labels to train an instance classifier and then aggregate instance prediction into bag prediction. However, the instance-level pseudo-labels derived from bag-level labels usually contain a lot of noise [32], which impairs the final classification performance. Consequently, the performance of instance-based MIL methods is generally inferior to bag-based MIL methods.

The second type is the bag-based MIL frameworks [33]–[37], which aggregate the instance features into bag features by certain aggregation methods and utilize bag labels for training. Max-pooling [38] and Mean-pooling [39] are two traditional aggregation methods, but their simple mechanisms usually lead to sub-optimal performance. To improve the performance, ABMIL [6] is proposed, introducing a learnable aggregator that generates bag-level representations by utilizing attention scores assigned to the instance representations. Building upon this work, Lu et al. proposed a CLAM [40], which selects the top-k salient instances based on attention scores and computes instance-level loss for better instance representation. Some studies also reconstruct the attention aggregator. For example, TransMIL [9] is proposed to employ a self-attention mechanism [41] to model the relationship among instances, while DSMIL [7] takes the distance between instances and the most salient one as attention scores, introducing a comprehensive multi-scale embedding fusion technology to enhance patch representation. In addition, feature clustering methods [34], [42], [43] compute cluster centroids of all feature embeddings, and then the representative feature embeddings are used for the final prediction. These approaches aim to enhance the interpretability and performance of models by leveraging the inherent structure of the data. On this basis, prototype-based methods [44]–[46] explore various ways to compute representative features by defining prototypes as typical components of images. These methods often involve encoding image patches, constructing clusters, and then decoding these clusters to obtain interpretable prototypes. The weighted combinations of prototype occurrences are used for image-level classification, enhancing both interpretability and classification performance.

B. Data Augmentation for MIL

Data augmentation can enhance the robustness of the model and has been broadly applied in the training of neural networks [47]–[49]. Traditional image-level data enhancement methods [50]–[52], such as flipping, rotating, and blurring, often consume a large amount of computing resources due to the huge size of WSIs. At present, most of the data augmentation methods used for WSIs are feature-level Mixup [22]–[25], [53]. The basic concepts of Mixup are as follows:

$$\widetilde{x} = \lambda x_i + (1-\lambda)x_j, \quad \widetilde{y} = \lambda y_i + (1-\lambda)y_j, \tag{1}$$

where the two input samples x_i and x_j are drawn from the training dataset, the labels corresponding to the input samples are y_i and y_j, and $\lambda \in [0,1]$ is sampled from $\sim$ Beta(α, α). Following the above formulation, there are two alignments [22]: (1) Size alignment: for WSIs, the number of instances in each bag and the feature dimensions of the instances are required to be aligned. (2) Semantic alignment: the sample and its corresponding label should be determined by the same λ. Based on these two alignments, many Mixup method variants have been proposed for WSI diagnosis. For example, ReMix [54] reduces the number of instances needed for alignment by replacing them with clustered prototypes. RankMix [23] sorts the instances of each bag according to their attention scores, and then removes lower-scoring instances from the bag with a larger number of instances to align two bags. PseMix achieves alignment by sampling pseudo bags following prototype clustering.

However, owing to the significant difference in the number of instances between WSIs, alignment loses a large amount of instance information, and existing Mixup methods mix bags as a whole, ignoring the importance of positive instances. Furthermore, the sub-bags generated during the mixing process often inherit the labels of the parent bags, which can lead to errors, especially when the bag contains few positive instances. This paper proposes a Mixup method based on an attention mechanism without requiring alignment. This paper proposes a Mixup method based on an attention mechanism without requiring alignment, effectively addressing these issues. Meanwhile, we introduce a flexible masking technique based on the attention mechanism to enhance the model's feature representation ability of instances.

III. Methodology

Fig. 3 shows the overview of our FAMIL. In this section, we elaborate on the application of ABMix and ABMask in the FAMIL framework. In the design of ABMix, we consider the multiple instance learning theory, generalizing the Mixup to pathological image classification with varying dataset characteristics and conditions. Meanwhile, ABMask selectively abandons salient positive instances for encouraging the model to mine hard samples.

A. Preliminaries

In MIL, any input WSI X is considered as a bag with multiple instances, which can be represented as $X = \{x_i\}_{i=1}^n$. x_i is a patch cropped from the WSI and considered as the i-th instance of X, and n is the number of instances. The bag label $Y \in \{0,1\}$ and instance labels $\{y_i\}_{i=1}^n$ follow:

$$Y = \begin{cases} 0, & \text{iff } \sum_i y_i = 0, \\ 1, & \text{others.} \end{cases} \tag{2}$$

Eq. 2 reflects that a negative bag includes only negative instances, whereas a positive bag contains at least one positive instance.

In MIL, we can only obtain the bag label, while the labels of each instance in a positive bag are not available. In the WSI classification task, the bag-based MIL is a popular solution, which derives a bag representation $\mathbf{F} \in \mathbb{R}^{1\times d}$ from the instance features $\mathbf{Z} = \left\{\mathbf{z}_i \in \mathbb{R}^{1\times d}\right\}_{i=1}^n$, where d is the dimension of the feature. The above operation is referred to as instance aggregation. With the bag-level representation, a bag classifier $\mathcal{C}_b(\cdot)$ is trained to classify the bag. Among existing methods, the mainstream instance aggregation strategy is the attention-based aggregation [6], which is formulated as:

$$\mathbf{F} = \sum_{i=1}^n a_i \mathbf{z}_i. \tag{3}$$

Here a_i is the attention score for the i-th instance, which is obtained by:

$$a_i = \frac{\exp\{\mathbf{W}^{\mathrm{T}}(\tanh(\mathbf{V}\mathbf{z}_{\mathrm{i}}^{\mathrm{T}}) \odot \mathrm{sigm}\ (\mathbf{U}\mathbf{z}_{\mathrm{i}}^{\mathrm{T}}))\}}{\sum_{k=1}^n \exp\{\mathbf{W}^{\mathrm{T}}(\tanh(\mathbf{V}\mathbf{z}_{\mathrm{k}}^{\mathrm{T}}) \odot \mathrm{sigm}\ (\mathbf{U}\mathbf{z}_{\mathrm{k}}^{\mathrm{T}}))\}}, \tag{4}$$

where $\mathbf{W} \in \mathbb{R}^{r\times 1}$, $\mathbf{U} \in \mathbb{R}^{r\times d}$, and $\mathbf{V} \in \mathbb{R}^{r\times d}$. "T" represents the transposition operation. Attention scores can reflect the contribution of each instance to the bag aggregation, providing interpretability for the WSI classification. Many works develop the formulation in different ways for attention scores. For example, DSMIL considers the distance between an instance and the most salient instance as the attention score, defined as follows:

$$a_i = \frac{\exp(\langle \mathbf{q}_i, \mathbf{q}_m \rangle)}{\sum_{k=1}^n \exp(\langle \mathbf{q}_k, \mathbf{q}_m \rangle)}. \tag{5}$$

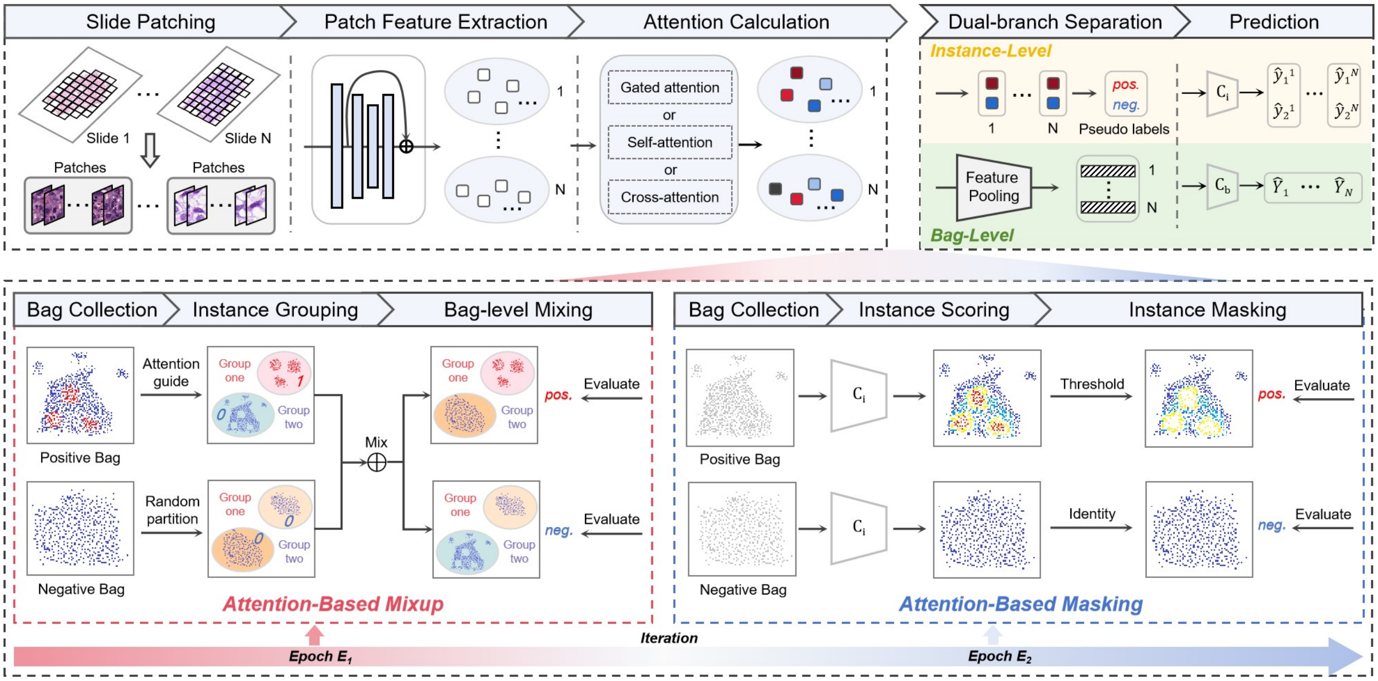


Fig. 3. Overview of our proposed FAMIL. FAMIL contains an instance branch and a bag branch. The red dotted box represents ABMix, and the blue dotted box represents ABMask, with a positive WSI and a negative WSI taken as examples. The solid rectangular box provides a t-SNE [55] visualization of the instances, where the scatter points represent the instance and each color represents a different score for the instance. *pos.* stands for the abbreviation of positive, and *neg.* stands for the abbreviation of negative.

Here, "$\langle \cdot, \cdot \rangle$" denotes the inner product of two vectors. q_i represents the query vector of z_i. Thus, the attention mechanism of DSMIL can be regarded as a variant of cross-attention. Another notable approach is TransMIL, which introduces a self-attention mechanism to effectively model the correlation between different instances, thereby enhancing the overall interpretability and classification performance.

B. Attention-Based Mixup

Although the attention mechanism facilitates MIL, the high degree of attention cannot always locate positive areas as expected. As discussed in Section I, we argue the reason behind the phenomenon is insufficient learning for positive instances. Ideally, for the positive WSI, a well-trained MIL model should capture positive areas in any scenario when the task is related to the tumor. However, the diversity of environments for positive instances in a single WSI is limited, leading to challenges in attention localization. Thus, this paper designs an attention-based Mixup that constructs diverse environments for positive samples to enhance the perception of the network for positive instances. Meanwhile, various negative bags are synthesized to avoid the unbalanced bag distribution for better training of the model. The red dotted box in Fig. 3 illustrates the proposed ABMix, which consists of two steps, namely group partitioning and group mixing. The detailed descriptions are listed in the following subsections.

1) Group Partitioning: Group partitioning divides the input WSI into two groups. We partition the positive WSI according to attention scores and randomly group the negative WSI. To be specific, for a positive bag $\mathbf{Z}^{pos} = \{\mathbf{z}_i^{pos}\}_{i=1}^{n_1}$ with n_1 instances, we feed it into the attention module $\mathcal{A}(\cdot)$ to get the attention scores A^{pos} of all the instances in the positive bag:

$$A^{pos} = \left[a_1^{pos}, a_2^{pos}, \ldots, a_{n_1}^{pos}\right] = \mathcal{A}(\mathbf{Z}^{pos}). \tag{6}$$

Then, we sort the instances in the positive bag from highest to lowest according to the attention scores:

$$I^{pos} = \left[i_1^{pos}, i_2^{pos}, \ldots, i_{n_1}^{pos}\right] = \mathrm{Sort}\left(A^{pos}\right), \tag{7}$$

where i_1^{pos} is the index of the instance with the highest attention score, while $i_{n_1}^{pos}$ is the index of the one with the lowest score in the positive bag.

For a negative bag $\mathbf{Z}^{neg} = \{\mathbf{z}_i^{neg}\}_{i=1}^{n_2}$ with n_2 instances, we sort the instances by randomly shuffling:

$$I^{neg} = \left[i_1^{neg}, i_2^{neg}, \ldots, i_{n_2}^{neg}\right] = \mathrm{Shuffle}\left(\mathbf{Z}^{neg}\right). \tag{8}$$

Using these index collections $I \in \{I^{pos}, I^{neg}\}$, we can conduct group partitioning for any input bag. Specifically, we divide all the instances into group one $\mathbf{G}^1$ and group two $\mathbf{G}^2$ using:

$$\left\{\mathbf{G}^1, \mathbf{G}^2\right\} = \mathcal{T}\left(I, k\right), \tag{9}$$

where $\mathcal{T}(I, k)$ represents that we select top k percent instances based on the instance index.

For a positive WSI, group one contains patches with high attention scores, and group two contains patches with low attention scores. According to the meaning of attention and the definition of bag labels in MIL, we assign a positive pseudo-label to group one and a negative pseudo-label to group two. For a negative WSI, both group one and group two are assigned negative pseudo-label.

2) Group Mixing: Group mixing randomly selects two WSIs $\mathbf{Z}_j = \left\{\mathbf{G}_j^1, \mathbf{G}_j^2\right\}$ and $\mathbf{Z}_k = \left\{\mathbf{G}_k^1, \mathbf{G}_k^2\right\}$, then mixes the group one of $\mathbf{Z}_j$ with the group two of $\mathbf{Z}_k$ to generate two synthesized WSIs $\mathbf{Z}_j^{'} = \left\{\mathbf{G}_j^1, \mathbf{G}_k^2\right\}$ and $\mathbf{Z}_k^{'} = \left\{\mathbf{G}_k^1, \mathbf{G}_j^2\right\}$, whose pseudo-labels are the same as $\mathbf{G}_j^1$ and $\mathbf{G}_k^1$, respectively. Different from previous Mixup methods, our labeling strategy follows the definition of MIL rather than Eq. 1, making it unnecessary to align the instance counts of the two input bags during our blending process. This operation promotes the model to focus on the positive instances, reducing the excessive attention to negative instances.

Considering the quality of synthesized bags, we introduce a quality evaluation to determine whether to use them. Specifically, for the synthesized positive WSI, we adopt the bag classifier $\mathcal{C}_b(\cdot)$ to predict the category of the bag. If the predicted positive possibility is greater than 50%, we add it to the dataset and use it to train the model in the corresponding training epoch; otherwise, the generated sample will be abandoned. For the synthesized negative bag, we only retain the samples mixed by two negative WSIs. Our quality assessment process significantly ensures the accuracy of synthetic bag labels.

C. Attention-Based Masking

To further modify the more challenging inaccurate attention on tumor areas, we design an ABMask technique to mine hard positive instances which is illustrated in the blue dotted box of Fig. 3. For an input bag $\mathbf{Z} = \{\mathbf{z}_i\}_{i=1}^n$ with n instances, ABMask uses an instance classifier $\mathcal{C}_i(\cdot)$ to generate the prediction for each instance of the bag, and adopts a Softmax function to map the classification results:

$$\hat{y}_i = Softmax\left(\mathcal{C}_i\left(\mathbf{z}_i\right)\right) \in \mathbb{R}^{1\times 2}, \tag{10}$$

where $\hat{y}_i$ is the prediction result of the i-th instance. Next, the predicted positive probability of the i-th instance is used as a salience score s_i. The higher the score, the more salient the instance. ABMask employs a salience threshold t to filter these salient instances and generates a masked bag $\mathbf{Z}^{mask}$ to replace the original input one. In the process, if $s_i \geq t$, the i-th instance is considered salient and is discarded; otherwise, it is retained.

Considering the imbalance of the tumor regions in the WSIs, we designed three different masking strategies combined with the quality evaluation as follows:

- **Fixed Masking:** For each positive WSI, fixed masking sets a fixed salience threshold and masks patches with salience scores higher than the threshold. Afterward, ABMask conducts a quality evaluation on the masked bag, which is the same as that in ABMix.
- **Random Masking:** Randomness is beneficial to reduce the risk of over-fitting. ABMask introduces a random masking strategy to obtain a masked bag. Specifically, for each positive WSI, random masking sets a random salience threshold and masks patches with salience scores higher than this threshold, which is a random number within a specified range. Subsequently, ABMask evaluates each masked bag.

Algorithm 1 Optimization Scheme in FAMIL

```
Input: Dataset $(\mathcal{X}, \mathcal{Y})$, Epoch $e$, $E_1$, $E_2$.
Load two WSIs $(X_j, Y_j)$, $(X_k, Y_k)$ from $(\mathcal{X}, \mathcal{Y})$;
if $e \geq E_1$ then
    Obtain representative instances and corresponding pseudo-labels;
    Calculate the instance loss $\mathcal{L}_{instance}$ by Eq. 12;
    if $e < E_2$ then
        for each $i \in (j, k)$ do
            if $Y_i = 1$ then
                Divide instances of a WSI into two groups $(G_i^1, G_i^2)$ by Eq. 6 and Eq. 7;
            else
                Divide instances of a WSI into two groups $(G_i^1, G_i^2)$ randomly by Eq. 8;
            end if
        end for
        Generate pseudo bags and corresponding labels $(X_j^{'}, Y_j^{'})$, $(X_k^{'}, Y_k^{'})$ by attention-based Mixup;
        Conduct quality evaluation and append the pseudo bags to the dataset;
    else
        for each $i \in (j, k)$ do
            if $Y_i = 1$ then
                Calculate the corresponding score $S_i$ by instance classifier $\mathcal{C}_i\left(.\right)$;
                Conduct Masking instances;
                Conduct quality evaluation and obtain the discarded bag;
            end if
        end for
    end if
end if
Calculate bag loss $\mathcal{L}_{bag}$ by Eq. 11;
Update parameters.
```

- **Step Masking:** Considering the specific nature of each positive bag, we design a Step Masking, which introduces multiple salience thresholds and adaptively selects the optimal threshold for each positive bag. Specifically, step masking first sets up a lower threshold (with a higher risk) to filter salient instances. Then the processed bag is evaluated to determine whether to use the threshold. If the masked bag fails to meet the requirement of evaluation. We adopt the higher threshold to mask fewer salient instances. If all thresholds cannot generate a satisfactory bag, we use the original bag for training. In this paper, we select three salience thresholds in step masking.

Unlike the hard instance mining approach [56] that abandons a certain proportion of high-attention instances, ABMask can mask bags flexibly. This is because ABMask adopts an instance classifier to quantify the salience of each instance, which enables the model to adaptively perceive salient positive instances and fully explore them. Such flexibility is essential given that, in numerous WSI datasets, the proportion of positive regions within each WSI can vary considerably.

D. Optimization

FAMIL is a two-branch framework as illustrated in Fig. 3. The bag-level branch is trained directly by the bag-level labels, the instance branch is trained by the instance-level labels distilled from the bag-level branch. The bag-level branch and the instance-level branch share the same fully connected layer of the encoder as described in Section IV-C, employing weight sharing.

1) Loss: In our FAMIL, bag classifier $\mathcal{C}_b(\cdot)$ outputs the bag-level prediction $\hat{Y}$. We adopt the cross entropy function to calculate the loss $\mathcal{L}_{bag}$ between the prediction $\hat{Y}$ and ground truth Y as follows:

$$\mathcal{L}_{bag} = Y\log\hat{Y} + (1-Y)\log\left(1-\hat{Y}\right). \tag{11}$$

In addition, for each positive WSI, we assign a positive pseudo-label to the instance with the highest attention score and a negative pseudo-label to the instance with the lowest attention score, and then train the instance classifiers $\mathcal{C}_i(\cdot)$ by them. The loss function of the instance $\mathcal{L}_{instance}$ is the cross-entropy between the network prediction $\hat{y}$ and the pseudo-label y:

$$\mathcal{L}_{instance} = y\log\hat{y} + (1-y)\log(1-\hat{y}). \tag{12}$$

2) Schedule: To achieve better convergence, during the early stages of FAMIL training, we primarily focus on optimizing the attention module and bag classifier. Once the attention module can effectively capture positive instances (by the E_1-th epoch), we introduce ABMix into the FAMIL training process and commence training the instance classifier. Upon convergence of the instance classifier (by the E_2-th epoch), we discontinue ABMix and integrate ABMask into the FAMIL training process to enhance the mining of challenging positive samples. Detailed steps are provided in Algorithm 1.

3) Multi-subtype classification: FAMIL can be easily developed to perform a multi-subtype classification task, where our instance-level classifier is trained by the instance with the largest attention score in a positive WSI of each subtype. The instance's label inherits the label of its bag. In ABMix, when two positive WSIs from different subtypes are extracted and mixed, the label of the synthetic WSI is the same as the group one of the original WSI.

IV. Experiments

A. Datasets

We evaluate our FAMIL on four public benchmarks, namely CAMELYON16, CAMELYON17, TCGA-NSCLC, and TCGA-RCC, which cover cases with balanced/unbalanced and single/multiple types of MIL problems. The CAMELYON16 [57] and CAMELYON17 [58] datasets are publicly available at the CAMELYON17 Grand Challenge website (https://camelyon17.grand-challenge.org/Data). The TCGA data (NSCLC, RCC) and corresponding labels are available from the National Institutes of Health genomic data commons (https://portal.gdc.cancer.gov).

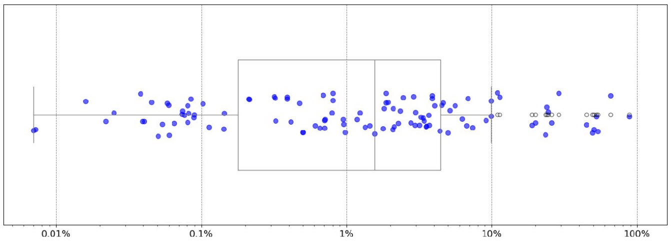


Fig. 4. A box plot showing the percentage of tumorous tiles (log-scaled) in tumorous slides in the training set of CAMELYON16. The grey, hollow points represent outliers, while the blue, circular points correspond to the data points themselves.

1) CAMELYON16: It is a public dataset for breast cancer metastasis detection containing 399 hematoxylins and eosin (H&E) stained WSIs from breast cancer patients. It consists of 270 training WSIs and 129 test WSIs. Among the training set, 110 of them are positive (metastasis) and the remaining are negative cases (normal). The test set consists of 50 positive WSIs and 79 negative WSIs. The average number of patches extracted per WSI is 11,559 at 20× magnification. This dataset presents a significant challenge among histological datasets due to the substantial variation in metastasis size from one slide to another. From a MIL perspective, this results in a significant disparity in the number of positive instances per bag. In some cases, there are only a few positive instances among tens of thousands of negative ones within a single bag, while in others, there may be nearly no negative instances. This variability is illustrated in Fig. 4 using a box plot, with the horizontal axis presented on a logarithmic scale.

2) CAMELYON17 (unseen): Similar to CAMELYON16, this dataset is a multi-center collection specifically designed for pathological N-staging in breast cancer. It comprises two categories of WSIs: those with lymph node metastasis and those without. The dataset originates from five distinct medical centers. To avoid overlap with CAMELYON16, WSIs from the same centers were excluded, resulting in 324 remaining WSIs (102 positive, 222 negative) for evaluating the domain generalization performance of the models. For our experiments, the CAMELYON17 dataset is utilized to assess the generalization capability of models trained on CAMELYON16 in diagnosing lymph node metastasis at the slide level. Due to the lack of publicly available annotations in the official CAMELYON17 test set, only the training set is used for evaluation, with slides labeled as isolated tumor cells excluded. After preprocessing, the dataset comprises a total of 3.77 million patches at 20× magnification, averaging 11,648 patches per WSI.

3) TCGA-NSCLC: It is a public lung cancer dataset that includes two subtypes, lung squamous cell carcinoma (LUSC) and lung adenocarcinoma (LUAD), with a total of 1,013 diagnostic WSIs including 512 LUADs and 501 LUSCs. In this dataset, the positive slides contain a relatively large area of tumor area (average 80% of the total cancer area per slide). After preprocessing, the average number of patches extracted per WSI was 2,894 at 10× magnification.

4) TCGA-RCC: It is a public renal cell carcinoma dataset that includes three subtypes, namely Kidney Chromophobe

TABLE I
RESULTS OF DIFFERENT MIL METHODS ON CAMELYON16, TCGA-NSCLC, AND TCGA-RCC DATASETS. EACH BOX INDICATES MEAN ± STANDARD DEVIATION. THE BEST PERFORMANCE IS MARKED IN **BOLD**.

Method	CAMELYON16			TCGA-NSCLC			TCGA-RCC		
	ACC (%)	F1 (%)	AUC (%)	ACC (%)	F1 (%)	AUC (%)	ACC (%)	F1 (%)	AUC (%)
Max-pooling	87.23±3.81	81.27±6.13	88.86±3.95	80.12±3.11	79.25±3.33	86.03±3.43	90.62±3.22	83.89±5.34	96.76±0.96
Mean-pooling	69.21±2.23	48.44±2.56	61.52±1.96	83.54±5.62	83.90±4.71	89.06±6.20	90.32±1.18	82.72±1.44	96.87±0.57
CLAM(SB)	84.50±2.61	76.57±4.16	86.98±2.65	85.34±3.39	85.04±3.25	91.04±4.36	92.01±2.53	86.13±4.49	97.18±1.22
CLAM(MB)	84.30±4.45	76.08±6.75	85.78±3.82	85.54±4.92	85.61±4.17	91.63±4.30	92.39±2.22	86.49±3.87	97.26±1.34
DTFD-MIL(AFS)	85.22±1.48	81.88±2.99	88.23±3.86	85.22±3.66	84.99±3.09	90.13±4.07	91.91±3.04	85.70±4.42	97.16±1.69
DTFD-MIL(MMS)	88.18±2.32	82.40±4.67	89.88±1.64	85.40±3.48	85.29±3.06	91.34±2.80	91.90±3.24	85.91±5.29	97.19±1.13
MS-CLAM*	85.12±2.46	78.96±3.86	87.23±2.37	-	-	-	-	-	-
MHIM-MIL	87.53±2.76	82.26±3.32	90.14±2.14	85.47±4.16	84.70±3.89	91.89±3.38	92.37±2.83	86.64±4.81	97.12±1.45
WiKG	88.27±1.84	82.32±2.68	90.31±2.09	85.47±3.68	85.74±3.72	91.89±3.04	92.53±2.17	87.22±4.53	97.22±1.39
MambaMIL	86.07±2.16	82.03±2.97	88.15±2.33	87.24±3.53	85.81±3.66	92.07±3.11	92.55±2.13	87.26±3.47	97.57±1.19
DSMIL	87.20±2.41	82.31±2.86	89.43±2.46	86.05±5.05	85.83±5.08	91.98±5.28	92.50±2.48	87.02±4.73	97.63±1.16
FAMIL(DSMIL)	90.36±1.27	86.49±2.22	91.96±1.68	87.84±3.83	**87.84±4.24**	92.60±4.21	93.24±1.83	87.90±3.45	**98.41±1.04**
ABMIL	87.38±2.81	82.41±3.43	89.26±2.56	86.13±4.22	85.55±4.39	91.95±3.57	92.35±2.93	86.73±4.71	97.53±1.36
FAMIL(ABMIL)	**92.38±1.32**	**89.19±2.44**	**92.61±1.96**	**87.93±3.42**	87.63±3.77	**93.81±2.97**	93.92±1.99	89.31±3.14	98.23±1.11
TransMIL	86.81±2.91	81.96±3.56	88.83±3.07	85.91±3.92	85.67±3.86	91.81±3.26	92.66±2.30	87.13±3.93	97.50±1.27
FAMIL (TransMIL)	89.93±2.29	86.14±2.87	91.55±2.43	87.04±3.82	87.45±3.53	92.49±2.98	**94.14±1.93**	**89.54±2.13**	98.37±1.02

* Note that, MS-CLAM is specifically designed for the positive and negative classification task.

Renal Cell Carcinoma (KICH), Kidney Renal Clear Cell Carcinoma (KIRC), and Kidney Renal Papillary Cell Carcinoma (KIRP), with a total of 880 diagnostic WSIs, including 110 WSIs of KICH, 488 WSIs of KIRC, and 282 WSIs of KIRP. TCGA-RCC is an unbalanced dataset among cancer subtypes, with large tumor areas in positive slides (average total tumor area per slide is 80%). After preprocessing, the number of patches extracted per WSI at 10× magnification is 3,482.

B. Experiment Setup and Evaluation Metrics

Following CLAM [40], each WSI is cropped into a series of 256 × 256 non-overlapping patches, where the background region is discarded. In CAMELYON16 [57], we divide the 270 training WSIs into training and validation sets at a ratio of 4:1 and tested on the official test set. For the TCGA dataset (https://camelyon16.grand-challenge.org/Data), the data is randomly split in the ratio of training:validation:test = 60:15:25. We use three evaluation metrics to report the performance of the model, including accuracy (ACC), F1 score (F1) and area under the curve (AUC). For CAMELYON16, we run experiments four times and report the averaged metrics. For TCGA, all experimental results are obtained by 4-fold cross-validation.

C. Implementation Details

We employ an Adam [59] optimizer with a learning rate of 0.001 and a weight decay of 0.0001 for optimizing the trainable weights of FAMIL. Following [40], instance features undergo embedding into a 1024-dimensional vector using an ImageNet pre-trained ResNet-50. During training, each feature embedding is compressed to 512 dimensions using a fully connected layer. In the inference stage, a softmax function is utilized to normalize the predictions for each class. All experiments are implemented on PyTorch 1.10.1 framework with an Nvidia RTX 3090 GPU. The hyperparameter k is consistently set to 0.5 across four datasets. Regarding the salience threshold t, a fixed masking strategy is employed, with t set to 0.98 for the CAMELYON16 dataset. For TCGA datasets, a step masking strategy is adopted, with each step t configured at 0.99, 0.98, and 0.97, respectively. The hyperparameters E_1 and E_2 are 50 and 300 respectively within the optimization scheme.

D. Comparisons with State-of-the-Art Methods

We present a comprehensive comparison of FAMIL with state-of-the-art methods on three datasets. The compared methods involve Max-pooling [38], Mean-pooling [39], ABMIL [6], DSMIL [7], DTFD-MIL (AFS) [35], DTFD-MIL (MMS) [35], CLAM (MB) [40], CLAM (SB) [40], MHIM-MIL [56], WiKG [60], TransMIL [9], MS-CLAM [17] and MambaMIL [61]. We obtain experimental results using their published code, where the hyperparameters of each method are set according to the implementation details described in their paper, all comparative experiments are conducted under identical experimental conditions, and only bag-level labels are available during the training process.

1) Quantitative Comparison: Table I lists the comparison results of different methods. The baselines of FAMIL are three representative attention-based MIL frameworks, namely ABMIL, DSMIL and TransMIL. Compared to other methods, Max-pooling and Mean-pooling perform poorly on three datasets. We attribute this to their insufficient modeling of the key instance information. This problem is especially severe on the CAMELYON16, where the proportion of positive instances is very small. The attention-based MIL methods achieve better results on all three datasets by identifying key instances. However, since the attention is not accurate, they are misled in some instances, resulting in limited performance. In particular, improved on attention-based MIL methods, MHIM-MIL benefits from the percentage mining of hard instances and

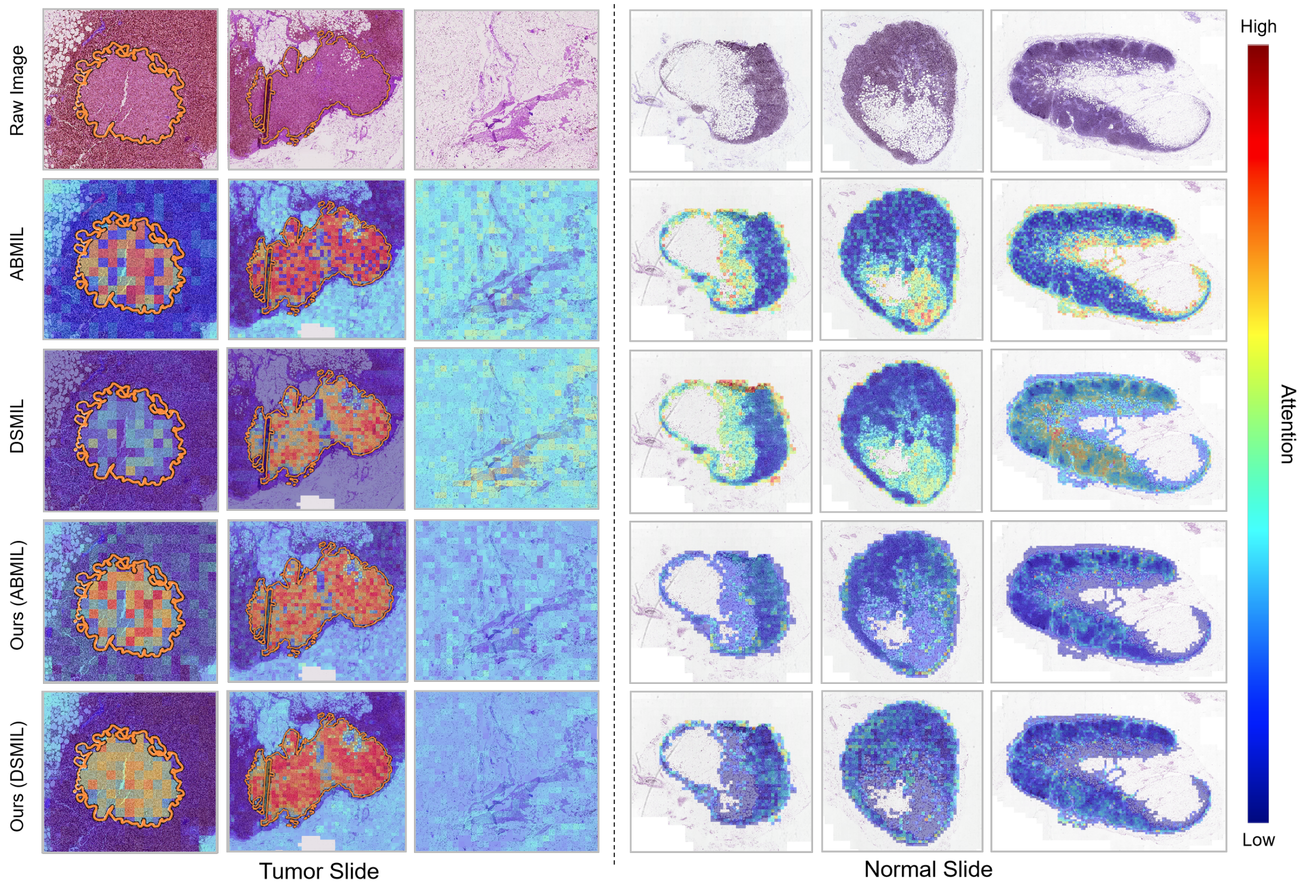


Fig. 5. Visualization of patches produced by ABMIL, DSMIL, and FAMIL. The orange curve outlines the area of the tumor. Redder patches indicate higher attention scores, whereas bluer patches indicate lower attention scores. Ideally, red patches should only cover the area inside the orange curve and blue patches only cover non-tumor areas. We show that our framework can significantly improve attention localization.

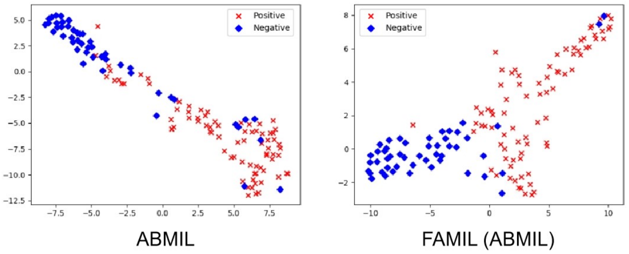


Fig. 6. Visualization of bag-level feature distribution with t-SNE on the CAMELYON16 dataset. Blue dots represent negative bags, while red crosses represent positive bags.

achieves the third-best performance on the three datasets, with 90.14% AUC on CAMELYON16, 91.89% AUC on TCGA-NSCLC and 97.12% AUC on TCGA-RCC. However, the percentage of tumors in different pathological images varies considerably, and the fixed ratio setting limits its generalization. The proposed FAMIL solves the two phenomena of inaccurate attention through ABMix and ABMask, breaking the performance bottleneck of the attention-based MIL. On CAMELYON16, our strategies significantly improve ABMIL, DSMIL and TransMIL by 3.35%, 2.53% and 2.72% in terms of AUC, respectively. Furthermore, we validate our framework on three different datasets, both of which can outperform the existing MIL methods.

2) Visualization Analysis: Fig. 5 visually illustrates the tumor localization abilities of various attention-based methods. From the figure, we have the following observations: (1) For positive WSI, ABMIL and DSMIL cannot accurately capture the tumor area (the second and third rows on the left), resulting in lower attention distribution in positive areas (first and second columns) and higher attention distribution in negative areas (third column). In contrast, FAMIL effectively improves this phenomenon (the fourth and fifth rows on the left), indicating that FAMIL has powerful tumor localization capability. (2) Fig. 5 also shows the attention distribution in negative samples, where the attention distribution of ABMIL and DSMIL is imbalanced (the second and third rows on the right), which may lead to poor generalization. Ideally, the model should treat each negative instance equally, meaning that attention should be balanced. In contrast, FAMIL mitigates the imbalanced attention distribution by increasing the diversity of the scene (the fourth and fifth rows on the right), achieving better performance.

To further analyze the impact of attention score on feature aggregation, we compared the aggregated bag-level features of ABMIL before and after FAMIL optimization on the CAMELYON16 test set. The t-SNE visualization results, as depicted in Fig. 6, clearly demonstrate that the FAMIL-enhanced ABMIL achieves more separable aggregated features, attributed to its enhanced attention localization capability.

TABLE II
Generalization capability evaluation results (mean ± standard deviation) of FAMIL. The best performance is marked in **bold**.

Method	Camelyon16 → Camelyon17 (unseen)		
	ACC (%)	F1 (%)	AUC (%)
CLAM (SB)	80.78±1.65	70.09±3.63	79.86±2.12
MS-CLAM	81.55±1.48	70.47±3.76	80.11±1.89
MHIM-MIL	81.32±1.94	70.19±3.58	79.86±2.61
ABMIL	81.49±1.76	69.97±3.91	79.33±2.34
FAMIL (ABMIL)	**84.34±0.91**	**73.16±2.80**	**82.03±0.72**
DSMIL	81.16±1.77	68.13±3.84	79.16±2.11
FAMIL (DSMIL)	83.23±0.96	70.34±3.14	81.47±0.83
TransMIL	77.87±2.58	65.09±4.23	71.31±3.04
FAMIL (TransMIL)	80.41±1.87	67.45±3.49	73.32±2.37

3) Generalization Capability: Generalization performance is a critical metric for evaluating MIL models, particularly given that test WSIs often display distinct visual characteristics from the training data due to variations in data acquisition processes. In this subsection, we assess the generalization capabilities of mainstream MIL models, with quantitative results summarized in Table II. The findings demonstrate that FAMIL excels not only in classifying familiar data but also in effectively handling previously unseen WSIs. On the unseen Camelyon17 dataset, FAMIL significantly enhances various classification metrics for attention-based methods. In particular, when integrated with ABMIL, FAMIL achieves impressed performance improvements, including a 2.85% increase in ACC, a 3.19% improvement in F1 score, and a 2.70% rise in AUC. These results underscore the robust generalization capacity of FAMIL in the domain of pathological image classification.

E. Ablation Study

1) Effects of Each Key Component in FAMIL: To validate the impact of each key component in FAMIL, we conduct a series of ablation studies on the CAMELYON16 dataset. The experiment results are reported in Table III. The ablation settings of each component are described below:

- Baseline: ABMIL is employed as a baseline, where two branches are applied for bag classification and instance classification.
- Baseline + ABMix: Compared with baseline, we introduce the ABMix to correct the inaccurate localization of attention outside the tumor.
- Baseline + ABMask: Compared with baseline, we incorporate the ABMask to enhance the network's ability to mine positive regions and reduce the neglect of tumor regions by the model's attention.
- Baseline + ABMix + ABMask: It is the framework FAMIL proposed in this work.

After the incorporation of ABMix, the model improves its ability to learn positive instance features, increasing the bag-level AUC of the baseline from 90.96% to 91.72%. And the instance-level AUC increased from 90.71% to 92.83%. Concurrently, ABMask facilitates the model mine difficult positive features, with the bag-level AUC increasing by 1.00% and the instance-level AUC increasing by 2.58%. When ABMix and ABMask are combined, the performance of the model is further improved, with the bag-level AUC reaching 92.61%, and the instance-level AUC reaching 95.53%. This substantiates the synergistic nature of these two methodologies, wherein their integration manifests as a mutually reinforcing mechanism, enhancing the model's power to obtain positive features.

TABLE III
Ablation experiments on the CAMELYON16 dataset based on ABMIL.

Setting			Bag-level		Instance-level	
Baseline	ABMix	ABMask	ACC (%)	AUC (%)	ACC (%)	AUC (%)
✓			88.41	90.96	89.05	90.71
✓	✓		90.03	91.72	91.27	92.83
✓		✓	90.48	91.96	91.86	93.29
✓	✓	✓	**92.38**	**92.61**	**93.32**	**95.53**

Moreover, a visualization result is shown in Fig. 7 which intuitively represents the improvement of our method on the baseline. In the figure, the red box represents the attention visualization of the tumor area, and the blue one denotes the attention visualization of the non-tumor area. We can observe that the baseline has two shortcomings: high attention distribution in non-tumor areas (the blue box in the second column) and low attention distribution in tumor areas (the red box in the second column). Notably, the incorporation of ABMix proves the advantages in mitigating high attention to non-tumor regions and simultaneously ameliorating the low attention distribution within tumor areas. Furthermore, ABMask serves to further diminish high attention to tumor areas, building upon the improvements facilitated by ABMix. These improvements further demonstrate the effectiveness of our FAMIL.

2) Effects of ABMix: As a flexible method, ABMix can be applied to any attention-based MIL framework. To prove the effectiveness of ABMix, we select popular mix-up strategies in pathological image analysis for comparison, including:

- **ReMix** [54]: The earliest mix-up-based approaches for MIL which mixes the prototypes of two bags within the same class;
- **Mixup** [53]: The original interpolation-based Mixup, in which two bags are aligned in the instance number before interpolation by random dropping instances from the bag with a larger instance number;
- **RankMix** [23]: An improved interpolation-based one, in which the instances of each bag are ranked sequentially according to attention scores, and delete instances with low attention for alignment;
- **PseMix** [22]: An instance-level mixup, in which Pseudo-bags are formed by sampling from prototype clusters,

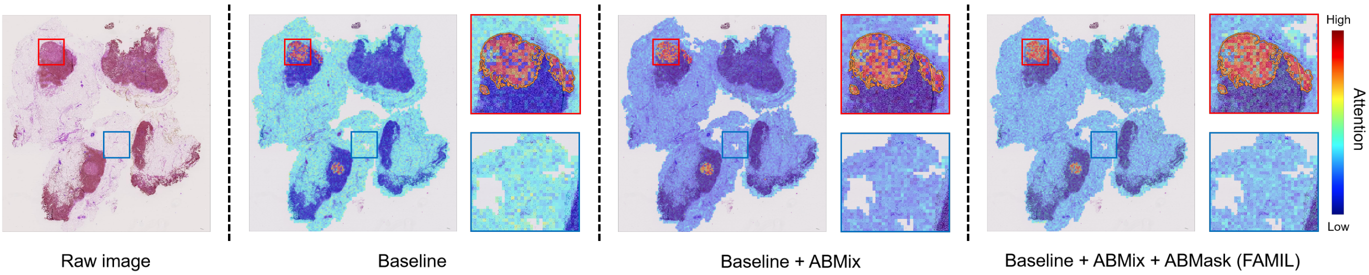


Fig. 7. Visualization of ablation experiments. The red box indicates the tumor area circled in orange. The blue box represents the non-tumor area. Redder patches indicate higher attention scores, whereas bluer patches indicate lower attention scores.

TABLE IV
RESULT OF DIFFERENT MIXUP METHODS ON CAMELYON16, TCGA-NSCLC, AND TCGA-RCC DATASETS. EACH BOX PRESENTS MEAN ± STANDARD DEVIATION. THE BEST PERFORMANCE ARE HIGHLIGHT IN **BOLD**.

Dataset	Method	ABMIL			DSMIL			Average		
		ACC (%)	F1 (%)	AUC (%)	ACC (%)	F1 (%)	AUC (%)	ACC (%)	F1 (%)	AUC (%)
CAMELYON16	Baseline	87.38±2.81	82.41±3.43	89.26±2.56	87.20±2.41	82.31±2.86	89.43±2.46	87.29	82.36	89.34
	w/ Remix	88.14±1.13	83.48±1.56	90.94±1.37	87.68±2.14	82.84±2.56	89.78±1.69	87.91	83.16	90.36
	w/ Mixup	88.01±2.23	83.56±2.80	90.89±1.82	87.84±1.67	82.63±2.15	89.63±1.39	87.92	83.10	90.26
	w/ RankMix	88.38±0.53	83.53±0.86	90.48±0.16	87.98±2.08	83.39±2.75	90.17±0.74	88.18	83.46	90.32
	w/ InstanceMix	86.82±0.97	81.13±1.26	89.91±0.89	87.22±2.62	80.88±4.49	87.95±2.81	87.02	81.01	88.93
	w/ PseMix	87.03±3.12	81.34±2.78	88.89±1.97	86.90±2.06	81.02±2.38	88.77±2.51	86.97	81.18	88.83
	w/ ABMix (ours)	**89.23±1.07**	**84.87±1.14**	**91.42±0.66**	**88.84±1.36**	**84.68±3.33**	**91.31±1.09**	**89.04**	**84.78**	**91.37**
	△ Over baseline	+1.85	+2.46	+2.16	+1.64	+2.37	+1.88	+1.75	+2.42	+2.02
TCGA-NSCLC	Baseline	86.13±4.22	85.55±4.39	91.95±3.57	86.05±5.05	85.83±5.08	91.98±5.28	86.09	85.69	91.97
	w/ Remix	86.54±3.43	85.88±3.57	92.43±2.89	85.86±3.13	85.84±3.02	91.77±4.59	86.20	85.86	92.10
	w/ Mixup	86.33±3.81	85.96±3.60	92.35±3.08	86.07±3.85	85.94±3.76	91.86±4.51	86.20	85.95	92.11
	w/ RankMix	84.54±3.51	84.55±3.16	91.52±2.81	85.97±2.98	85.82±2.46	91.53±3.85	85.26	85.19	91.53
	w/ InstanceMix	87.01±3.29	86.86±3.21	92.44±3.19	85.38±3.34	85.33±2.64	91.07±4.65	86.19	86.09	91.76
	w/ PseMix	87.06±3.31	86.97±3.10	92.53±2.75	86.15±3.54	86.11±2.51	92.07±3.66	86.61	86.54	92.30
	w/ ABMix	**87.29±3.41**	**87.34±3.02**	**92.93±2.66**	**86.73±3.14**	**86.91±2.53**	**92.22±3.56**	**87.01**	**87.13**	**92.58**
	△ Over baseline	+1.16	+1.79	+0.98	+0.68	+1.08	+0.24	+0.92	+1.44	+0.61
TCGA-RCC	Baseline	92.35±2.93	86.73±4.71	97.53±1.36	92.50±2.48	87.02±4.73	97.63±1.16	92.42	86.87	97.58
	w/ Remix	92.44±2.71	87.45±4.17	97.43±1.29	92.73±1.92	87.32±3.27	97.72±0.89	92.59	87.39	97.58
	w/ Mixup	92.61±2.50	87.32±4.05	97.68±1.24	92.76±1.81	87.57±3.23	97.81±0.97	92.68	87.44	97.74
	w/ RankMix	92.74±2.46	87.56±4.48	97.38±1.36	92.78±1.98	**87.81±3.51**	97.87±0.82	92.76	87.68	97.62
	w/ InstanceMix	91.97±2.77	87.48±1.45	97.12±1.43	92.69±1.28	87.59±1.96	97.53±0.95	92.33	87.53	97.32
	w/ PseMix	92.84±2.59	**88.31±3.18**	97.79±1.47	92.78±1.33	87.74±2.49	97.85±0.91	92.81	88.03	97.82
	w/ ABMix	**92.92±2.60**	88.29±2.77	**98.02±1.36**	**92.87±1.34**	87.79±2.29	**98.04±0.88**	**92.90**	**88.04**	**98.03**
	△ Over baseline	+0.57	+1.56	+0.49	+0.37	+0.77	+0.41	+0.47	+1.17	+0.45

pseudo-bag labels inherit parent bag labels, and Mixup is achieved through mixing pseudo-bags;

- **InstanceMix** [22]: An instance-level Mixup baseline, which randomly selects a certain proportion of instances from two bags and combines them into a mixed bag. The label setting of the mixed bag is consistent with Mixup.

Table IV shows the comparison results of the Mixup methods on three datasets, from which we can obtain the following observations: (1) The existing Mixup methods are powerless to generalize to various scenarios, for example, PseMix reduces ABMIL's AUC by 0.37% on the CAMELYON16 dataset; RankMix reduces ABMIL's AUC by 0.15% on the TCGA-RCC dataset; Mixup reduces the AUC of DSMIL by 0.12% on TCGA-NSCLC. This is because they do not emphasize the importance of positive examples, which is critical in MIL. In contrast, ABMix enhances the model's perception ability of positive areas in different scenarios by simulating different negative areas for positive areas, achieving improvements on each dataset. (2) Under different baselines, the performance improvement brought by ABMix is higher than that brought by other methods, for example, on TCGA-NSCLC, based on ABMIL, the AUC of ABMix is 0.40% higher than that of the suboptimal Mixup method (PseMix); on TCGA-RCC, based on DSMIL, the AUC of ABMix is 0.17% higher than that of the suboptimal Mixup method (RankMix). These results demonstrate the effectiveness of ABMix, which can further enhance the focus of attention-based methods on positive instances, increasing the classification accuracy of the model.

In addition, we visualized the attention of different Mixup methods on the CAMELYON16 dataset, as illustrated in Fig. 8. ABMix improves the label fusion definition to better align with the multi-instance learning framework and employs distinct mixing strategies for positive and negative bags. As a result, ABMix enables the model to concentrate more effectively on learning positive features, thereby achieving superior attention localization capabilities.

3) Effects of ABMask: ABMask is another core design in our framework. The main idea of this method is to combine the

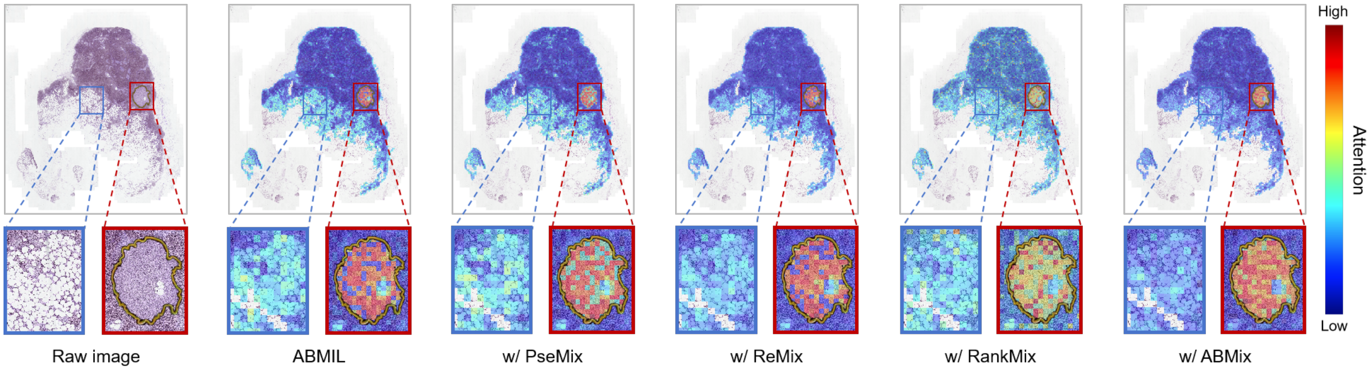


Fig. 8. Visualization of different Mixup method on CAMELYON16 dataset. The red box indicates the tumor area circled in orange. The blue box represents the non-tumor area. Redder patches indicate higher attention scores, whereas bluer patches indicate lower attention scores.

TABLE V
COMPARISON RESULTS (%) ON CAMELYON16 DATASET BASED ON ABMIL AMONG DIFFERENT MASKING STRATEGIES.

Masking	CAMELYON16			TCGA-NSCLC			TCGA-RCC		
	ACC	F1	AUC	ACC	F1	AUC	ACC	F1	AUC
Baseline	90.03	85.48	91.72	87.33	87.29	92.98	92.95	88.59	98.04
Fixed	**92.38**	**89.19**	92.61	87.64	87.33	93.59	93.67	89.16	98.15
Random	91.33	87.84	92.29	87.48	87.42	93.22	93.83	**89.71**	97.99
Step	91.92	88.51	**92.72**	**87.93**	**87.63**	**93.81**	**93.92**	89.31	**98.23**

instance branches to discard the most salient instances, thereby indirectly forcing the model to mine more difficult-to-learn positive instances to facilitate model training. On this basis, we design three ABMask strategies (Fixed Masking, Random Masking, and Step Masking) combined with the evaluation strategy and present their impact in Table V. After introducing the three strategies, the model boosts performance on all three datasets. Specifically, Fixed Masking shows more significant performance improvement on the CAMELYON16, while Random Masking obtains the highest F1 on TCGA-RCC. Furthermore, the more complex ABMask strategy (Step Masking) achieves the best performance on the TCGA-NSCLC dataset which has a larger proportion of positive instances. Overall, this experiment verifies the effectiveness of the ABMask strategy, and the diversity of the proposed strategy improves its applicability to different datasets.

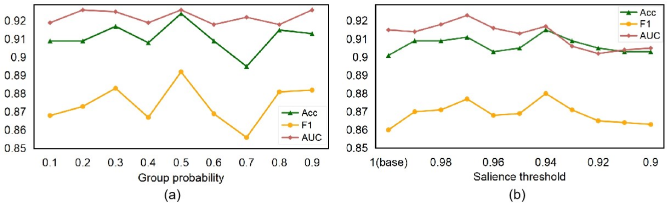


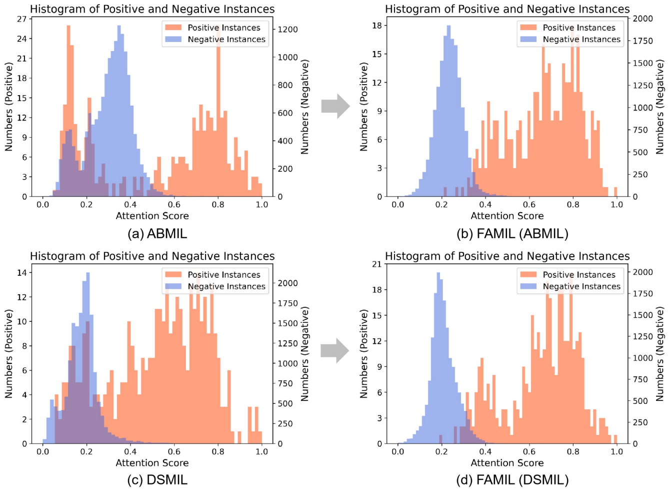
Fig. 9. Study on two key hyperparameters of FAMIL: (a) Group probability k and (b) Salience threshold t.

4) Hyperparameter Analysis: We conducted experiments to assess the impact of two crucial hyperparameters in the FAMIL framework: k, representing the proportion of group one in each bag, and t, denoting the salience threshold for masking. The outcomes of these tests are presented in Fig. 9. Regarding k, proximity to our default setting ($k = 0.5$) tends to yield optimal performance. In the case of t, we employ a Fixed Masking approach without an evaluation strategy to better elucidate its influence on the model during testing. As depicted in the right panel of Fig. 9, a systematic reduction in t from its maximum value of 1 initially enhances performance, followed by a subsequent decline. This phenomenon is attributed to the fact that a larger t results in discarding too few salient instances, limiting the model's efficacy in mining challenging positive cases. Conversely, a smaller t leads to the model discarding an excessive number of positive instances, impeding the learning of positive features.

V. DISCUSSION

To further discuss the relationship between instance discrimination and attention. we conducted experiments on the CAMELYON16 [57] dataset with pixel-level annotations, and Fig. 10 shows the results of different methods on test-001. As can be seen from the figure, ABMIL's [6] attention performs poorly in discriminating instances, evident in the misallocation of attention, wherein negative instances attract high attention and positive instances receive disproportionately low attention. Although DSMIL [7] demonstrates improved performance, the underlying issue persists. In contrast, FAMIL effectively enhances the coherence between instance distribution and attention. This improvement is reflected in the alignment of attention patterns with instance decision boundaries, thereby facilitating accurate tumor localization.

Previous feature-level enhancement methods for WSIs generally focused on the Mixup [24] technique, but these methods typically performed size alignment and semantic alignment according to the Mixup definition intended for natural images. It fails to take into account the particularity of the MIL framework, resulting in the neglect of positive instances, which are overwhelmed by the massive negative instances. Some studies have explored other data augmentation methods for mining hard instances. From the perspective of hard instance definition, existing methods are limited to setting a fixed ratio for hard instance mining, and ignore the huge difference in the tumor proportion of each WSI. To this end, the ABMix we



(a) ABMIL

(b) FAMIL (ABMIL)

(c) DSMIL

(d) FAMIL (DSMIL)

Fig. 10. The relationship between instance discrimination and attention, where (a) and (c) depict the attention distribution of ABMIL and DSMIL, respectively. Concurrently, (b) and (d) illustrate the attention distribution of our proposed FAMIL, with these two attention-based MIL methods serving as the respective baselines.

proposed re-optimizes the feature mixing and label generation of Mixup specifically for the MIL framework, and places emphasis on the feature enhancement of positive instances, effectively alleviating the problem of negative and positive patch imbalance. ABMask introduces an instance-level branch to quantify the salience of each instance, adapting to WSIs with different tumor proportions. In addition, FAMIL can be transferred to any attention-based MIL model. Therefore, the proposed FAMIL is a simple and effective MIL framework.

VI. Conclusion

This paper primarily addresses the issue of inaccurate attention localization in attention-based MIL methods, which is caused by insufficient learning of positive instances. This shortcoming impacts the classification performance and generalization of the model. Here, we propose a FAMIL framework and design ABMix and ABMask to enhance its ability to learn positive features. ABMix underscores the importance of positive instances in the MIL paradigm and effectively generalizes Mixup to the attention-based MIL without the need for alignment, thus improving instance feature representation. ABMask selectively masks salient positive instances, encouraging the network to mine challenging instances. Each approach includes an evaluation mechanism to ensure operational quality. Extensive experiments demonstrate that our approach mitigates the problem of inaccurate attention in attention-based MIL methods, and the proposed FAMIL framework exhibits superior performance compared to state-of-the-art methods. As a data-efficient approach, FAMIL offers valuable insights into the identification of rare diseases.

References

[1] G. Campanella, M. G. Hanna, L. Geneslaw, A. Miraflor, V. Werneck Krauss Silva, K. J. Busam, E. Brogi, V. E. Reuter, D. S. Klimstra, and T. J. Fuchs, "Clinical-grade computational pathology using weakly supervised deep learning on whole slide images," *Nature Medicine*, vol. 25, no. 8, pp. 1301–1309, 2019.

[2] J. N. Kather, A. T. Pearson, N. Halama, D. Jäger, J. Krause, S. H. Loosen, A. Marx, P. Boor, F. Tacke, U. P. Neumann *et al.*, "Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer," *Nature Medicine*, vol. 25, no. 7, pp. 1054–1056, 2019.

[3] H. Wieslander, P. J. Harrison, G. Skogberg, S. Jackson, M. Fridén, J. Karlsson, O. Spjuth, and C. Wählby, "Deep learning with conformal prediction for hierarchical analysis of large-scale whole-slide tissue images," *IEEE journal of biomedical and health informatics*, vol. 25, no. 2, pp. 371–380, 2020.

[4] S. Jiang, Z. Gan, L. Cai, Y. Wang, and Y. Zhang, "Multimodal cross-task interaction for survival analysis in whole slide pathological images," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2024, pp. 329–339.

[5] W. Hou, L. Yu, C. Lin, H. Huang, R. Yu, J. Qin, and L. Wang, "Hˆ 2-mil: Exploring hierarchical representation with heterogeneous multiple instance learning for whole slide image analysis," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 1, 2022, pp. 933–941.

[6] M. Ilse, J. Tomczak, and M. Welling, "Attention-based deep multiple instance learning," in *International Conference on Machine Learning*. PMLR, 2018, pp. 2127–2136.

[7] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 14 318–14 328.

[8] W. Shao, Z. Han, J. Cheng, L. Cheng, T. Wang, L. Sun, Z. Lu, J. Zhang, D. Zhang, and K. Huang, "Integrative analysis of pathological images and multi-dimensional genomic data for early-stage cancer prognosis," *IEEE Transactions on Medical Imaging*, vol. 39, no. 1, pp. 99–110, 2019.

[9] Z. Shao, H. Bian, Y. Chen, Y. Wang, J. Zhang, X. Ji *et al.*, "Transmil: Transformer based correlated multiple instance learning for whole slide image classification," *Advances in Neural Information Processing Systems*, vol. 34, pp. 2136–2147, 2021.

[10] G. Quellec, G. Cazuguel, B. Cochener, and M. Lamard, "Multiple-instance learning for medical image and video analysis," *IEEE reviews in biomedical engineering*, vol. 10, pp. 213–234, 2017.

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

[12] E. Mercan, S. Aksoy, L. G. Shapiro, D. L. Weaver, T. T. Brunyé, and J. G. Elmore, "Localization of diagnostically relevant regions of interest in whole slide images: a comparative study," *Journal of Digital Imaging*, vol. 29, pp. 496–506, 2016.

[13] P. Gupta, Y. Huang, P. K. Sahoo, J.-F. You, S.-F. Chiang, D. D. Onthoni, Y.-J. Chern, K.-Y. Chao, J.-M. Chiang, C.-Y. Yeh *et al.*, "Colon tissues classification and localization in whole slide images using deep learning," *Diagnostics*, vol. 11, no. 8, p. 1398, 2021.

[14] K. Das, S. Conjeti, A. G. Roy, J. Chatterjee, and D. Sheet, "Multiple instance learning of deep convolutional neural networks for breast histopathology whole slide classification," in *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*. IEEE, 2018, pp. 578–581.

[15] J. Yao, X. Zhu, J. Jonnagaddala, N. Hawkins, and J. Huang, "Whole slide images based cancer survival prediction using attention guided deep multiple instance learning networks," *Medical Image Analysis*, vol. 65, p. 101789, 2020.

[16] T. Lin, Z. Yu, H. Hu, Y. Xu, and C.-W. Chen, "Interventional bag multi-instance learning on whole-slide pathological images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 19 830–19 839.

[17] P. Tourniaire, M. Ilie, P. Hofman, N. Ayache, and H. Delingette, "Ms-clam: Mixed supervision for the classification and localization of tumors in whole slide images," *Medical Image Analysis*, 2023.

[18] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of Big Data*, vol. 6, no. 1, pp. 1–48, 2019.

[19] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, and Y. Yoo, "Cutmix: Regularization strategy to train strong classifiers with localizable features," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 6023–6032.

[20] J.-H. Kim, W. Choo, and H. O. Song, "Puzzle mix: Exploiting saliency and local statistics for optimal mixup," in *International Conference on Machine Learning*. PMLR, 2020, pp. 5275–5285.

[21] R. Takahashi, T. Matsubara, and K. Uehara, "Data augmentation using random image cropping and patching for deep cnns," *IEEE Transactions on Circuits and Systems for Video Technology*, 2019.

[22] P. Liu, L. Ji, X. Zhang, and F. Ye, "Pseudo-bag mixup augmentation for multiple instance learning-based whole slide image classification," *IEEE Transactions on Medical Imaging*, 2024.

[23] Y.-C. Chen and C.-S. Lu, "Rankmix: Data augmentation for weakly supervised learning of classifying whole slide images with diverse sizes and imbalanced categories," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 23 936–23 945.

[24] M. Gadermayr, L. Koller, M. Tschuchnig, L. M. Stangassinger, C. Kreutzer, S. Couillard-Despres, G. J. Oostingh, and A. Hittmair, "Mixup-mil: Novel data augmentation for multiple instance learning and a study on thyroid cancer diagnosis," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2023, pp. 477–486.

[25] T. Stegmüller, B. Bozorgtabar, A. Spahr, and J.-P. Thiran, "Scorenet: Learning non-uniform attention and augmentation for transformer-based histopathological image classification," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 6170–6179.

[26] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Pérez, "Solving the multiple instance problem with axis-parallel rectangles," *Artificial Intelligence*, vol. 89, no. 1-2, pp. 31–71, 1997.

[27] G. Xu, Z. Song, Z. Sun, C. Ku, Z. Yang, C. Liu, S. Wang, J. Ma, and W. Xu, "Camel: A weakly supervised learning framework for histopathology image segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 10 682–10 691.

[28] L. Hou, D. Samaras, T. M. Kurc, Y. Gao, J. E. Davis, and J. H. Saltz, "Patch-based convolutional neural network for whole slide tissue image classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2424–2433.

[29] M. Lerousseau, M. Vakalopoulou, M. Classe, J. Adam, E. Battistella, A. Carré, T. Estienne, T. Henry, E. Deutsch, and N. Paragios, "Weakly supervised multiple instance learning histopathological tumor segmentation," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part V 23*. Springer, 2020, pp. 470–479.

[30] P. Chikontwe, M. Kim, S. J. Nam, H. Go, and S. H. Park, "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*. Springer, 2020, pp. 519–528.

[31] S. Ding, J. Li, J. Wang, S. Ying, and J. Shi, "Multi-scale efficient graph-transformer for whole slide image classification," *IEEE Journal of Biomedical and Health Informatics*, 2023.

[32] M.-A. Carbonneau, V. Cheplygina, E. Granger, and G. Gagnon, "Multiple instance learning: A survey of problem characteristics and applications," *Pattern Recognition*, vol. 77, pp. 329–353, 2018.

[33] P. Chikontwe, M. Luna, M. Kang, K. S. Hong, J. H. Ahn, and S. H. Park, "Dual attention multiple instance learning with unsupervised complementary loss for covid-19 screening," *Medical Image Analysis*, vol. 72, p. 102105, 2021.

[34] Y. Sharma, A. Shrivastava, L. Ehsan, C. A. Moskaluk, S. Syed, and D. Brown, "Cluster-to-conquer: A framework for end-to-end multi-instance learning for whole slide image classification," in *Medical Imaging with Deep Learning*. PMLR, 2021, pp. 682–698.

[35] H. Zhang, Y. Meng, Y. Zhao, Y. Qiao, X. Yang, S. E. Coupland, and Y. Zheng, "Dtfd-mil: Double-tier feature distillation multiple instance learning for histopathology whole slide image classification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18 802–18 812.

[36] X. Shi, F. Xing, Y. Xie, Z. Zhang, L. Cui, and L. Yang, "Loss-based attention for deep multiple instance learning," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 04, 2020, pp. 5742–5749.

[37] L. Cai, S. Huang, Y. Zhang, J. Lu, and Y. Zhang, "Rethinking attention-based multiple instance learning for whole-slide pathological image classification: An instance attribute viewpoint," *arXiv preprint arXiv:2404.00351*, 2024.

[38] W. Zhu, Q. Lou, Y. S. Vang, and X. Xie, "Deep multi-instance networks with sparse label assignment for whole mammogram classification," in *Medical Image Computing and Computer Assisted Intervention- MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part III 20*. Springer, 2017, pp. 603–611.

[39] J. Feng and Z.-H. Zhou, "Deep miml network," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 31, no. 1, 2017.

[40] M. Y. Lu, D. F. Williamson, T. Y. Chen, R. J. Chen, M. Barbieri, and F. Mahmood, "Data-efficient and weakly supervised computational pathology on whole-slide images," *Nature Biomedical Engineering*, vol. 5, no. 6, pp. 555–570, 2021.

[41] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[42] C. Xie, H. Muhammad, C. M. Vanderbilt, R. Caso, D. V. K. Yarlagadda, G. Campanella, and T. J. Fuchs, "Beyond classification: Whole slide tissue histopathology analysis by end-to-end part learning," in *Medical Imaging with Deep Learning*. PMLR, 2020, pp. 843–856.

[43] X. Wang, H. Chen, C. Gan, H. Lin, Q. Dou, E. Tsougenis, Q. Huang, M. Cai, and P.-A. Heng, "Weakly supervised deep learning for whole slide lung cancer image analysis," *IEEE Transactions on Cybernetics*, vol. 50, no. 9, pp. 3950–3962, 2019.

[44] L. Qu, Y. Ma, X. Luo, Q. Guo, M. Wang, and Z. Song, "Rethinking multiple instance learning for whole slide image classification: A good instance classifier is all you need," *IEEE Transactions on Circuits and Systems for Video Technology*, 2024.

[45] H. Yang, B. Sun, B. Li, C. Yang, Z. Wang, J. Chen, L. Wang, and H. Li, "Iterative class prototype calibration for transductive zero-shot learning," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 3, pp. 1236–1246, 2022.

[46] X. Ding, B. Li, Y. Li, W. Guo, Y. Liu, W. Xiong, and W. Hu, "Web objectionable video recognition based on deep multi-instance learning with representative prototypes selection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 3, pp. 1222–1233, 2020.

[47] Y. Zhou, Y. Sun, L. Li, K. Gu, and Y. Fang, "Omnidirectional image quality assessment by distortion discrimination assisted multi-stream network," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 4, pp. 1767–1777, 2021.

[48] B. Song, J. Zhou, and H. Wu, "Multistage curvature-guided network for progressive single image reflection removal," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 10, pp. 6515–6529, 2022.

[49] R. Theagarajan and B. Bhanu, "An automated system for generating tactical performance statistics for individual soccer players from videos," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 2, pp. 632–646, 2020.

[50] S. Cheng, S. Liu, J. Yu, G. Rao, Y. Xiao, W. Han, W. Zhu, X. Lv, N. Li, J. Cai *et al.*, "Robust whole slide image analysis for cervical cancer screening using deep learning," *Nature Communications*, vol. 12, no. 1, p. 5639, 2021.

[51] D. Tellez, G. Litjens, P. Bándi, W. Bulten, J.-M. Bokhorst, F. Ciompi, and J. Van Der Laak, "Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology," *Medical Image Analysis*, vol. 58, p. 101544, 2019.

[52] H. Wang, Q. Wang, H. Zhang, J. Yang, and W. Zuo, "Constrained online cut-paste for object detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 10, pp. 4071–4083, 2020.

[53] W. Tang, S. Huang, X. Zhang, F. Zhou, Y. Zhang, and B. Liu, "mixup: Beyond empirical risk minimization," in *6th International Conference on Learning Representations, ICLR 2018*, May 2018.

[54] J. Yang, H. Chen, Y. Zhao, F. Yang, Y. Zhang, L. He, and J. Yao, "Remix: A general and efficient framework for multiple instance learning based whole slide image classification," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2022, pp. 35–45.

[55] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." *Journal of machine learning research*, vol. 9, no. 11, 2008.

[56] W. Tang, S. Huang, X. Zhang, F. Zhou, Y. Zhang, and B. Liu, "Multiple instance learning framework with masked hard instance mining for whole slide image classification," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2023, pp. 4078–4087.

[57] B. E. Bejnordi, M. Veta, P. J. Van Diest, B. Van Ginneken, N. Karssemeijer, G. Litjens, J. A. Van Der Laak, M. Hermsen, Q. F. Manson, M. Balkenhol *et al.*, "Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer," *Jama*, vol. 318, no. 22, pp. 2199–2210, 2017.

[58] P. Bandi, O. Geessink, Q. Manson, M. Van Dijk, M. Balkenhol, M. Hermsen, B. E. Bejnordi, B. Lee, K. Paeng, A. Zhong *et al.*, "From detection of individual metastases to classification of lymph node status

at the patient level: the camelyon17 challenge," *IEEE transactions on medical imaging*, vol. 38, no. 2, pp. 550–560, 2018.

[59] P. K. Diederik, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[60] J. Li, Y. Chen, H. Chu, Q. Sun, T. Guan, A. Han, and Y. He, "Dynamic graph representation with knowledge-aware attention for histopathology whole slide image analysis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 11 323–11 332.

[61] S. Yang, Y. Wang, and H. Chen, "Mambamil: Enhancing long sequence modeling with sequence reordering in computational pathology," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2024, pp. 296–306.

Hailun Cheng received the BS degree in Mechanical Engineering from Soochow University in 2022. He is currently working towards the MS degree at Tsinghua Shenzhen International Graduate School, Tsinghua University. His research interests include computational pathology and mutiple instance learning.

Shenjin Huang received a bachelor's and a master's degrees from the College of Mechanical and Electronic Engineering at Northwest Agriculture and Forestry University in 2019 and 2020, respectively. He is currently pursuing a Ph.D. at Harbin Institute of Technology. His research interests are remote sensing and medical image analysis.

Linghan Cai received the B.S. degreee from the Department of Information and Electrical Engineering, China Agricultural University, in 2020, the M.S. degree in the Department of Electronic Information Engineering, Beihang University, in 2023. He is currently pursuing the Ph.D. degree with Harbin Institute of Technology. His research interests include scene parsing and multi-modal learning.

Yangfan Xu received the BS degree in Biomedical Engineering from Beihang University in 2022. He is currently working towards the MS degree at Tsinghua Shenzhen International Graduate School, Tsinghua University. His research interests include computational pathology and deep learning.

Runming Wang received the Ph.D. degree in Bioinorganic Chemistry from The University in Hong Kong in 2018. He is currently an Assistant Professor in Institute of Biopharmaceutical and Health Engineering, Shenzhen International Graduate School, Tsinghua University. He is engaged in cross-disciplinary research in the fields of Metals in Chemical Biology, Synthetic Biology, Synthetic Chemistry and Computational Biology. Through the development of innovative drugs and related chemical/biological tools, the structure and function of key metallo-proteins (enzymes) in pathogenic bacteria, viruses and eukaryotic cells are finely regulated, and a certain research foundation has been made in overcoming major scientific problems such as emerging infectious diseases, malignant tumors and aging.

Yongbing Zhang received the BA degree in English and the MS and PhD degrees in computer science from Harbin Institute of Technology, Harbin, China, in 2004, 2006, and 2010, respectively. He is currently a Professor of Computer Science and Technology, Harbin Institute of Technology (Shenzhen), Shenzhen, China. He joined Tsinghua Shenzhen In ternational Graduate School from 2010 to 2020, and was a visiting scholar of University of California, Berkeley from 2016 to 2017. He was the receipt of the Best Student Paper Award at IEEE International Conference on Visual Communication and Image Processing in 2015 and the Best Paper Award at Pacific-Rim Conference on Multimedia in 2018. His current research interests include signal processing, machine learning, and computational imaging.