RPMIL: Rethinking Uncertainty-Aware Probabilistic Multiple Instance Learning for Whole Slide Pathology Diagnosis

Zhikang Zhao[1] , **Kaitao Chen**[1] and **Jing Zhao**[1,2,*]
[1]School of Computer Science and Technology, East China Normal University, Shanghai, China
[2]Key Laboratory of System Control and Information Processing, Ministry of Education, Shanghai, China
51265901012@stu.ecnu.edu.cn, 51215901053@stu.ecnu.edu.cn, jzhao@cs.ecnu.edu.cn

Abstract

Whole slide images (WSIs) are gigapixel digital scans of traditional pathology slides, offering substantial support for cancer diagnosis. Current multiple instance learning (MIL) methods for WSIs typically extract instance features and aggregate these into a single bag feature for prediction. We observe that these MIL methods rely on point estimation, where each bag is mapped to a deterministic embedding. Such MIL methods based on point estimation fail to capture the full spectrum of data variability due to the reliance on fixed embedding, especially when the number of trainable bags is limited. In this paper, we rethink probabilistic modeling in MIL and propose RPMIL, an uncertainty-aware probabilistic MIL method for whole slide pathology diagnosis. RPMIL learns a probabilistic aggregator to consolidate instance features into dynamic bag feature distributions instead of a deterministic bag feature. Specifically, we employ a variational autoencoder approach to compress multiple instance features into a low-dimension space with probabilistic representation and obtain the bag feature distribution formulated by the mean and variance. Furthermore, we drive the prediction by jointly leveraging the instance feature distribution and bag feature distribution. We evaluate the WSI classification performance on two public datasets: Camelyon16 and TCGA-NSCLC. Extensive experiments demonstrate that our method surpasses point estimation methods in MIL, achieving state-of-the-art levels.

1 Introduction

Deep neural networks have made significant strides in recent years, benefiting various fields worldwide. In the field of computational pathology, the application of various deep learning methods has led to new insights into disease diagnosis. Whole slide image (WSI) pathology diagnosis is regarded as the gold standard for identifying or excluding tumors and remains a hot topic of research. Due to the gigapixel resolution of WSIs and the absence of fine-grained patch labels, current methods predominantly follow the multiple instance learning (MIL) paradigm [Tellez *et al.*, 2019; Pinckaers *et al.*, 2020; Zhang *et al.*, 2022]. This approach divides a WSI (bag) into many small patches (instances), where the presence of even a single positive instance can lead to the bag as positive.

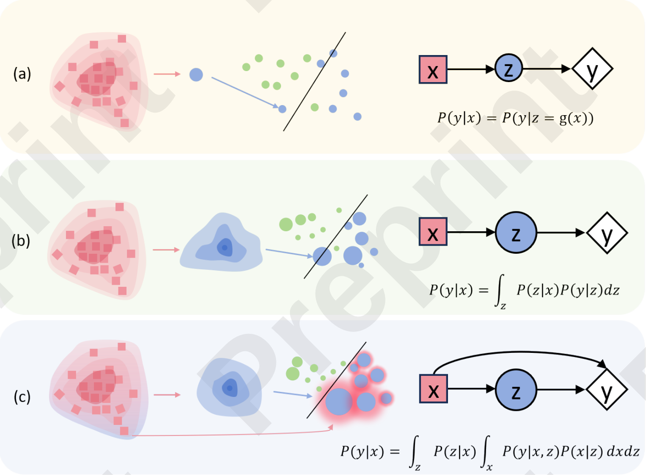
(a)
x
z
y
$P(y|x) = P(y|z = g(x))$
(b)
x
z
y
$P(y|x) = \int_z P(z|x)P(y|z)dz$
(c)
x
z
y
$P(y|x) = \int_z P(z|x)\int_x P(y|x,z)P(x|z)\,dxdz$

Figure 1: (a) Point estimation in multiple instance learning. **x**: instance features in a same bag; **z**: a deterministic bag feature aggregated from instances; y: the prediction. (b) The proposed probabilistic MIL utilizing bag uncertainty. **z**: a bag feature distribution. (c) The proposed probabilistic MIL utilizes both bag and instance uncertainties.

Existing research primarily focuses on enhancing instance feature representation, such as through self-supervised learning [Huang *et al.*, 2023; Lu *et al.*, 2024; Chen *et al.*, 2024b] or instance feature re-embedding [Chikontwe *et al.*, 2022; Tang *et al.*, 2024]. More work is dedicated to the design of bag aggregators, using methods like modifying instance weight allocation [Li *et al.*, 2021] or designing Transformer-based aggregators [Shao *et al.*, 2021]. Recently, some novel studies have begun to address false relationships in MIL [Lin *et al.*, 2023; Chen *et al.*, 2024a], recalculating intervention expectations to replace traditional likelihood estimation. In practice, training data is often limited, typically consisting of only a few hundred or a few thousand bags. In a deterministic embedding

[*] Corresponding author: Jing Zhao.

space, a small number of bag features can easily fit the decision boundary, but this fitting is often fragile and prone to overfitting. The bag feature obtained in this manner are suboptimal. Although existing research has explored data augmentation through splitting or mixing existing bags to generate pseudo-bags [Zhang *et al.*, 2022; Chen and Lu, 2023; Liu *et al.*, 2024], the number of training samples is typically only expanded by a limited factor.

From the perspective of uncertainty estimation [Kendall and Gal, 2017], we rethink the process of MIL modeling. We observe that these MIL methods fall under **point estimation**, where each bag is mapped to a deterministic embedding. Conversely, **uncertainty estimation** maps input samples to an embedding distribution, better capturing the bag feature distribution and aiding in understanding the confidence of the estimation. Figure 1(a) illustrates the process of point estimation in MIL, while Figure 1(b) depicts the probabilistic MIL utilizing bag uncertainty. The essential difference between the two lies in whether the bag feature is represented as a single value or a probability distribution. Furthermore, we consider the instance features within a bag as a distribution and assume that the instance feature distribution can also aid in prediction. This assumption aligns with the logical process in actual clinical diagnosis, where pathologists identify tumor regions (instance) to assess whether WSI is a tumor [Xiong *et al.*, 2023]. Consequently, we jointly utilize the bag distribution and the instance distribution to drive predictions, as shown in Figure 1(c). Two main issues need to be verified when applying uncertainty estimation in MIL: **i**) Can uncertainty estimation outperform the dominant point estimation MIL methods? Specifically, this involves comparing $P(y|\mathbf{z} = g(\mathbf{x}))$ and $\int_{\mathbf{z}} P(\mathbf{z}|\mathbf{x})P(y|\mathbf{z})\,\mathrm{d}\mathbf{z}$. **ii**) Can instance distribution and bag distribution jointly enhance classification results? This involves comparing likelihood $P(y|\mathbf{z})$ and $P(y|\mathbf{z}, \mathbf{x})$.

In this paper, we rethink probabilistic modeling in MIL and propose RPMIL, an uncertainty-aware probabilistic MIL method for whole slide pathology diagnosis. RPMIL learns a probabilistic aggregator that consolidates instance features into a bag feature distribution rather than the fixed bag feature. Specifically, the probabilistic aggregator uses a variational autoencoder approach [Cemgil *et al.*, 2020; Michelucci, 2022] to compress multiple instance features into two low-dimensional latent space features, representing the mean and variance of the bag feature distribution, respectively. By reparameterization sampling [Kingma *et al.*, 2015], an arbitrary number of bag feature representations can be obtained. We use Kullback-Leibler (KL) divergence and mean squared error (MSE) loss to constrain the bag feature distribution, preventing distribution drift. Importantly, we propose combining instance and bag feature distribution for the final prediction. Through extensive experiments and analysis, we reach two significant conclusions: **i**) Uncertainty estimation based on bags surpasses point estimation methods in MIL. **ii**) Combining instance feature distribution and bag feature distribution for prediction significantly outperforms using a single distribution. Through ablation studies, we further find that as the number of sampling in the bag feature distribution increases, the results become more pronounced.

2 Related Work

2.1 Multiple Instance Learning for WSIs

There are two main methods for predicting bags: the instance-based method and the bag-based method. A typical instance-based approach trains a instance classifier to assign pseudo-labels to each instance based on bag labels [Qu *et al.*, 2024b; Lin *et al.*, 2024]. Due to the large number of instances, the top-k policy is often used to select instances [Xu *et al.*, 2019; Chikontwe *et al.*, 2020], necessitating a very large number of WSIs [Chen and Lu, 2023]. Given the fact that WSIs lack pixel-level labels and contain both cancerous and normal regions, bag-based approaches have become dominant. The rapid development of deep learning, particularly the exploration of attention mechanisms [Vaswani, 2017; Ilse *et al.*, 2018], has led to the rise of bag-based methods that can directly predict the label of a bag by aggregating instance features into a bag feature [Li *et al.*, 2021; Shao *et al.*, 2021; Zhang *et al.*, 2022]. The advantage of this approach is that it does not require the labels of numerous instances, so only the bag labels are needed for classification. Some studies also explore improving bag feature representation, employing multi-scale learning [Xiong *et al.*, 2023; Qu *et al.*, 2023] or fusing multimodal information [Qu *et al.*, 2024a; Li *et al.*, 2024; Shi *et al.*, 2024] to achieve a better representation of the bag. The above MIL methods fall under point estimation and fail to capture the full spectrum of data variability. Therefore, we propose RPMIL, an uncertainty-aware probabilistic MIL method to learns a probabilistic aggregator.

2.2 Probabilistic Methods in MIL

Introducing a probabilistic model in multiple instance learning can adequately take into account uncertainty [Haußmann *et al.*, 2017; Kendall and Gal, 2017]. DGMIL [Qu *et al.*, 2022] is an instance-based MIL method that considers instance feature distribution while assigning pseudo-labels to instances and focuses on constructing instance-level feature distribution. Bayes-MIL [Cui *et al.*, 2023] uses the uncertainty of the attention weight of each instance as a measure of the accuracy of guessing whether the instance is positive or negative. However, our RPMIL is a bag-based method to construct distributions for bag features and we drive the prediction by jointly utilizing instance feature distribution and bag feature distribution. AGP [Schmidt *et al.*, 2023] is also a bag-based method that utilizes Gaussian processes to dynamically obtain attention scores for each instance. In contrast, our RPMIL learns a probabilistic aggregator based on a generative model, constructing the bag features as a distribution, which allows for sampling to capture uncertainty. We note that Causal inference methods such as IBMIL [Lin *et al.*, 2023] and CaMIL [Chen *et al.*, 2024a] have also made adjustments to the probability calculation formula by using interventional likelihood expectations. Yet, these causal methods still fall under point estimation, and they rely on sampling based on clustering. Conversely, our RPMIL constructs a normal distribution about bag features. We perform sampling via reparameterization and utilize MCMC [Rubinstein and Kroese, 2016] to approximate the desired quantity.

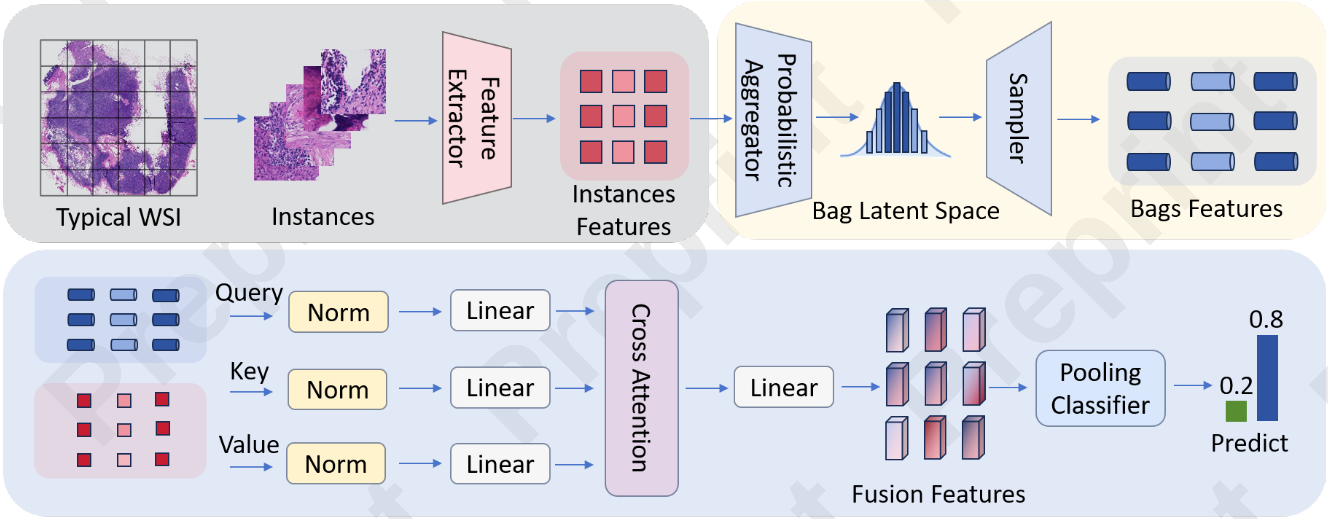

Figure 2: The overall architecture of our proposed uncertainty-aware probabilistic MIL method. RPMIL learns a probabilistic aggregator and maps the WSI to the bag feature distribution rather than the deterministic bag feature. Furthermore, we drive the prediction by jointly leveraging the instance feature distribution and bag feature distribution.

3 Methodology

3.1 Preliminary

The MIL methodology involves the following steps: first, instance features are obtained using a feature extractor with fixed weights. Then, these instance features are aggregated into a bag feature by assigning attention scores or through global integration. Finally, these bag features are used for classification. For a given dataset of WSIs, each WSI has a corresponding label $y_i \in \{0, 1\}$ for the binary classification task, which determines whether a certain cancer is present. Moreover, each WSI referred to as a bag is divided into many patches p_i, i.e., instances, $\{p_{i,1}, p_{i,2}, \ldots, p_{i,n_i}\}$ and n_i is the number of instances from a bag. The bag label y_i is defined as follows:

$$y_i = \begin{cases} 0, \text{ if } \sum_{j=1}^{n_i} y_{i,j} = 0, \\ 1, \text{ else}, \end{cases} \tag{1}$$

where $y_{i,j}$ corresponds to the label of $p_{i,j}$.

A challenging problem arises because the number of instances usually reaches thousands or more. We only have access to bag labels y_i. Our goal is to predict the bag label when the instance labels are unknown. This whole process of MIL can be summarized as follows:

$$\mathbf{x}_{i,j} = f(p_{i,j}), \mathbf{z}_i = g\left(\{\mathbf{x}_{i,j}\}_{j=1}^{n_i}\right), \hat{y}_i = h(\mathbf{z}_i), \tag{2}$$

where $f(\cdot)$ is a pre-trained feature extractor used to obtain the instance features $\mathbf{x}$, $g(\cdot)$ is an aggregator that converges all instance features $\mathbf{x}$ into a bag feature $\mathbf{z}$, and $h(\cdot)$ is a classifier to get the predicative result $\hat{y}$. The optimization objective is to minimize the cross-entropy (CE) loss between the real label y and the prediction $\hat{y}$.

There is a dependency among all the variables due to their sequential relationship, as illustrated in Figure 1(a). The traditional MIL methods fall under point estimation, and the prediction probability can be defined as follows:

$$P(y|\mathbf{x}) = P(y|\mathbf{z} = g(\mathbf{x})). \tag{3}$$

In this approach, a bag is mapped to a deterministic embedding, which is then used to predict its output. $P(y|\mathbf{z})$ is the final prediction probability.

3.2 Uncertainty-Aware Probabilistic MIL

MIL methods based on point estimation fail to capture the full spectrum of data variability due to their reliance on a fixed embedding, especially when the number of trainable bags is limited. Conversely, uncertainty estimation is better equipped to capture data diversity, as it maps input samples to an embedded distribution. Thus, we propose an uncertainty-aware probabilistic MIL method for whole slide pathology diagnosis, as shown in Figure 2. Our method learns a probabilistic aggregator that aggregates instance features into a bag feature distribution $P(\mathbf{z}|\mathbf{x})$ rather than a fixed bag feature $\mathbf{z}$. Then, we obtain $\mathbf{z}$ by sampling from the dynamic distribution space, allowing us to derive a range of different values that provide openings for the bag classifier. Here, the specific prediction probability equation is different from Equation (3) and is defined as follows:

$$P(y|\mathbf{x}) = \int_{\mathbf{z}} P(\mathbf{z}|\mathbf{x}) P(y|\mathbf{z}) \, \mathrm{d}\mathbf{z}. \tag{4}$$

Now, we obtain the probability by sampling $\mathbf{z}$ from the distribution $P(\mathbf{z}|\mathbf{x})$ and then performing marginal integration. In an implementation, we use Markov Chain Monte Carlo (MCMC) [Rubinstein and Kroese, 2016] to approximate the desired quantity.

Furthermore, we conceptualize the instance features within a bag as representing a distribution. We consider that jointly leveraging these distributions can enhance the final prediction, as shown in Figure 1(c). This approach mirrors the clinical diagnostic process, wherein pathologists use tumor areas

Method	Camelyon16			TCGA-NSCLC		
	ACC	F1-score	AUC	ACC	F1-score	AUC
Mean-pooling [Wang *et al.*, 2018]	65.89±1.35	31.75±2.72	56.22±2.76	83.21±2.26	82.67±2.45	90.93±1.55
Max-pooling [Wang *et al.*, 2018]	81.40±1.55	74.16±1.13	82.45±0.97	85.50±1.90	84.03±1.01	94.73±0.80
ABMIL [Ilse *et al.*, 2018]	84.50±1.55	76.74±1.83	87.17±1.25	85.09±1.76	84.35±1.53	92.74±1.10
CLAM-SB [Lu *et al.*, 2021]	80.62±3.09	72.09±3.74	83.42±2.42	87.83±1.52	85.22±2.81	94.67±0.54
CLAM-MB [Lu *et al.*, 2021]	81.40±2.32	74.23±2.69	86.81±1.23	87.07±1.90	85.71±1.84	94.07±0.83
TransMIL [Shao *et al.*, 2021]	86.05±3.10	80.95±3.36	91.02±2.35	87.45±1.52	86.89±1.46	94.51±1.57
DGMIL [Qu *et al.*, 2022]	82.17±2.28	79.13±1.70	85.71±2.03	89.93±1.70	88.91±1.43	95.43±1.44
DTFD-MaxS [Zhang *et al.*, 2022]	88.37±0.77	82.35±1.37	89.13±1.58	85.55±1.14	83.26±1.84	91.64±1.46
DTFD-MaxMinS [Zhang *et al.*, 2022]	87.59±1.63	82.75±2.64	92.27±1.48	88.59±1.14	88.06±0.74	95.23±1.27
DTFD-AFS [Zhang *et al.*, 2022]	90.69±0.78	86.95±0.96	93.46±0.71	88.97±1.90	87.71±1.85	95.01±0.88
MMIL [Zhang *et al.*, 2023]	91.18±2.33	88.91±2.64	94.83±1.36	90.11±1.67	88.57±1.64	96.03±0.60
DGRMIL [Zhu *et al.*, 2024]	91.47±1.55	89.07±1.67	93.02±1.57	90.38±0.65	89.02±1.00	95.69±0.74
RPMIL(Ours+ResNet-50)	91.72±1.07	89.36±1.25	95.13±1.58	90.49±1.53	89.34±1.84	96.31±1.01
RPMIL(Ours+PLIP)	**92.42±0.95**	**89.89±0.63**	**95.71±1.32**	**90.76±1.45**	**90.99±1.57**	**97.08±0.83**

Table 1: Comparison performance of WSIs classification on the Camelyon16 and the TCGA-NSCLC dataset.

(instance) to assess WSI. We modify the probability $P(y|\mathbf{x})$ in Equation (4) to obtain the $P(y|\mathbf{x},\mathbf{z})$, and the final prediction is as follows:

$$P(y|\mathbf{x}) = \int_{\mathbf{z}} P(\mathbf{z}|\mathbf{x}) \int_{\mathbf{x}} P(y|\mathbf{x},\mathbf{z}) P(\mathbf{x}|\mathbf{z}) \mathrm{d}\mathbf{x}\mathrm{d}\mathbf{z}. \tag{5}$$

For detailed formula derivations, please refer to the supplementary materials. To calculate $P(\mathbf{x}|\mathbf{z})$, we use a cross-attention [Vaswani, 2017] module, where $\mathbf{z}$ is act as query and $\mathbf{x}$ is act as key. We fuse the joint distributions of instance and bag:

$$\widetilde{\mathbf{z}} = \text{Softmax}\left(\frac{(\mathbb{N}(\mathbf{z})\mathbf{W}_q(\mathbb{N}(\mathbf{x})\mathbf{W}_k)^\top}{\sqrt{d}}\right)\mathbb{N}(\mathbf{x})\mathbf{W}_v, \tag{6}$$

where $\mathbb{N}(\cdot)$ is the layer normalization, $\mathbf{z} \in \mathbb{R}^{k\times d}$, $\mathbf{x} \in \mathbb{R}^{n\times d}$, $\mathbf{W}_q \in \mathbb{R}^{d\times d}$, $\mathbf{W}_k \in \mathbb{R}^{d\times d}$ and $\mathbf{W}_v \in \mathbb{R}^{d\times d}$ are learnable parameters of query, key and value, respectively.

Eventually, we use the pooling operator and feed it to the classifier $h(\cdot)$ to get the final predicted labels:

$$\hat{y} = h(\text{Pooling}(\widetilde{\mathbf{z}})). \tag{7}$$

3.3 Building Bag Distribution

In our probabilistic framework, it is necessary to construct a distribution of bag features. Specifically, we compress multiple instance features into a new distribution for classification purposes. Here, we use a variational autoencoder to compress the instance features into two low-dimensional latent space features, representing the mean and variance of the bag feature distribution, respectively. We use KL loss to regularise the distribution $P(\mathbf{z}|\mathbf{x})$. We assume that $P(\mathbf{z}|\mathbf{x})$ obeys a normal distribution, with the mean μ and variance σ^2 obtained by the encoder. Given the assumption that $Q(\mathbf{z})$ is a prior that follows a standard normal distribution. The KL loss is defined as follows:

$$\text{KL}(P(\mathbf{z}|\mathbf{x})||Q(\mathbf{z})) = -\frac{1}{2}(\log(\sigma^2) - \sigma^2 - \mu^2 + 1). \tag{8}$$

The benefit of keeping $P(\mathbf{z}|\mathbf{x})$ close to a standard normal distribution is the benefit of ensuring that the feature space is more compact and distinct, which aids the classifier in making better distinctions. Now, we have reformulated Equation (2), and the bag feature distribution is represented as follows:

$$\mathbf{z}_i \sim \mathcal{N}\left(\mu_i, \sigma_i^2\right), \text{ where } \mu_i, \sigma_i = \widetilde{g}\left(\{\mathbf{x}_{i,j}\}_{j=1}^{n_i}\right). \tag{9}$$

By reparameterization sampling, an arbitrary number of bag features $\mathbf{z}$ can be obtained. To prevent excessive differences between features sampled from uncertainty estimation and those obtained from point estimation, we use MSE to constrain the differences between the mean of the sampled features $\bar{\mathbf{z}}$ and the point-estimated feature $\mathbf{z}'$. The overall loss is summarised as follows:

$$\begin{aligned} \mathcal{L} = {} & \lambda_1 \text{CE}(y, \hat{y}) + \lambda_2 \text{KL}(P(\mathbf{z}|\mathbf{x})||Q(\mathbf{z})) \\ & + \lambda_3 \text{MSE}(\bar{\mathbf{z}}, \mathbf{z}'), \end{aligned} \tag{10}$$

where λ_1, λ_2 and λ_3 are hyperparameters.

4 Experiments

4.1 Datasets

The effectiveness of the proposed method is examined on both the **Camelyon16** [Bejnordi *et al.*, 2017] and the **TCGA-NSCLC** datasets. The Camelyon16 dataset, from a competition organized by ISBI to classify breast cancer, is officially divided into a training set and a test set, with 270 WSIs in the training set and 129 in the test set.

The TCGA is a joint project of the National Cancer Institute and the National Human Genome Research Institute, including clinical data on a wide range of human cancers. We select two types of non-small cell lung cancer (NSCLC) WSIs for classification: lung adenocarcinoma (LUAD) and lung squamous carcinoma (LUSC). TCGA-NSCLC dataset contains a total of 1053 WSIs, with 541 LUAD from 478 patients and 512 LUSC from another 478 patients. We divide the dataset into training, validation, and test sets in a 6:1.5:2.5 ratio.

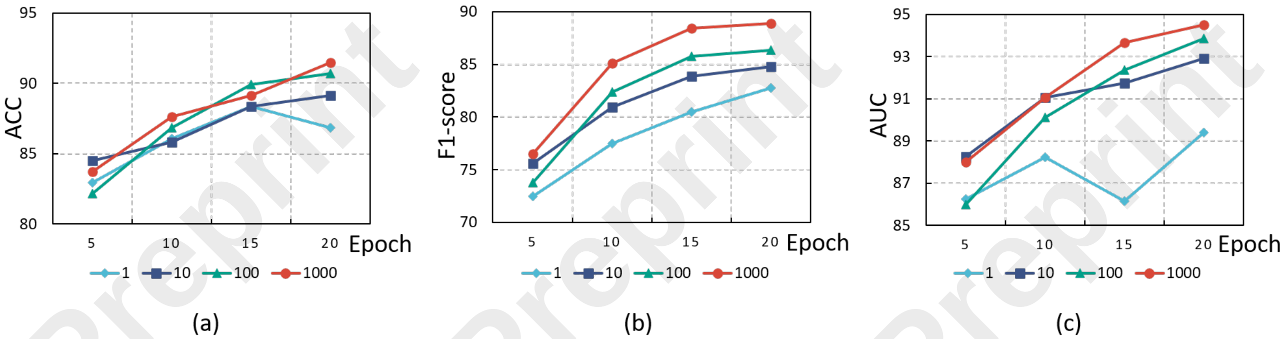

Figure 3: The effect of the number of sampling (1, 10, 100, and 1,000) on the Camelyon16 dataset. (a) Results on ACC. (b) Results on F1-score. (c) Results on AUC. A larger sampling size effectively improves model performance and stability.

4.2 Preprocessing and Implement Details

Following CLAM [Lu *et al.*, 2021], all WSIs are divided into non-overlapping patches at 20× magnification with a window size of 256×256 pixels. Our statistics show that the Camelyon16 has a total of approximately 3.617 million patches, with an average of about 9,066 patches per WSI, while the TCGA-NSCLC has a total of about 12.731 million patches, with an average of about 12,102 patches per WSI.

All patches are extracted using ResNet-50 pre-trained on ImageNet-1k [He *et al.*, 2016], resulting in features with a dimension of 1024. Additionally, PLIP pre-trained on large-scale pathology images is used to obtain 512-dimensional features [Huang *et al.*, 2023]. We use the Adam optimizer with an initial learning rate of 1e-4 and a weight decay of 1e-5, adjusting the learning rate using a cosine annealing scheme. The minibatch is set to 1, the number of epochs is set to 100, and the hyperparameter λ_1 is 1, λ_2 and λ_3 are both set to 0.5. We record accuracy (ACC), F1 score, and area under curve (AUC) as criteria for evaluating our models. All experiments are performed on an NVIDIA GeForce RTX 3090 and are repeated five times, with all metrics reported as averages.

4.3 Baseline Methods

We compare many classical multiple instance learning methods for WSIs classification, including ABMIL, CLAM, TransMIL, DGMIL, DTFD, MMIL and DGR-MIL. We use the vanilla attention-based method as part of our default bag probabilistic aggregator. ABMIL enables the aggregation of instance features by assigning different attention scores to each instance. CLAM employs a multi-branch attention network that trains the instance classifier to obtain higher-scoring instances. TransMIL and MMIL are multiple instance learning networks based on the Transformer architecture that take into account information between instances. DTFD is a dual-layer framework which introduces pseudo-bags to increase the number of bags, and the pseudo-bags are still at the same resolution. DGMIL performs WSI classification from the perspective of data distribution, while it focuses on instances rather than bags. DGR-MIL takes into account the relationship between global vectors and instance embeddings, but diverse global vectors make training unstable.

Dataset	Distribution	ACC	F1-score	AUC
Camelyon16	$P(y\|\mathbf{z}=g(\mathbf{x}))$	84.50	76.74	87.17
	$P(y\|\mathbf{z})$	86.05	80.43	89.16
	$P(y\|\mathbf{x},\mathbf{z})$	**91.72**	**89.36**	**95.13**
TCGA-NSCLC	$P(y\|\mathbf{z}=g(\mathbf{x}))$	85.09	84.35	92.74
	$P(y\|\mathbf{z})$	86.61	86.27	93.23
	$P(y\|\mathbf{x},\mathbf{z})$	**90.49**	**89.34**	**96.31**

Table 2: Impact of joint distribution of instance feature and bag feature on model performance.

4.4 Experimental Results

As shown in Table 1, our method achieves the best results on both the Camelyon16 and the TCGA-NSCLC datasets, in terms of all three evaluation metrics: ACC, F1-score, and AUC. Especially, our method has a significant improvement on the Camelyon16 dataset compared with DGMIL, a method for constructing instance feature distribution, 9.55% increase on ACC, 10.23% increase on F1-score, and 9.42% increase on AUC. This improvement can be attributed to the limitations of DGMIL, which relies on clustering techniques to construct instance distributions and requires positive instances to train the instance classifier. On the Camelyon16 dataset, however, the small cancer regions in the WSI positive samples result in a limited number of available positive instances, hindering DGMIL's performance. In contrast, our method handles such imbalances more effectively, demonstrating better performance under these challenging conditions. Notably, these methods usually achieve better results on the TCGA-NSCLC compared with the Camelyon16. In particular, the two constructed baselines, mean-pooling and max-pooling [Wang *et al.*, 2018], also performe admirably on the latter. Specifically, the TCGA-NSCLC dataset contains a larger number of WSIs, and the distribution of positive and negative samples is more balanced compared to the Camelyon16 dataset. The average tumor region in positive samples of the TCGA-NSCLC dataset is significantly larger, often around 80%, whereas, in the Camelyon16 dataset, the tumor region in positive samples is much smaller, averaging only about 10% [Shao *et al.*, 2021].

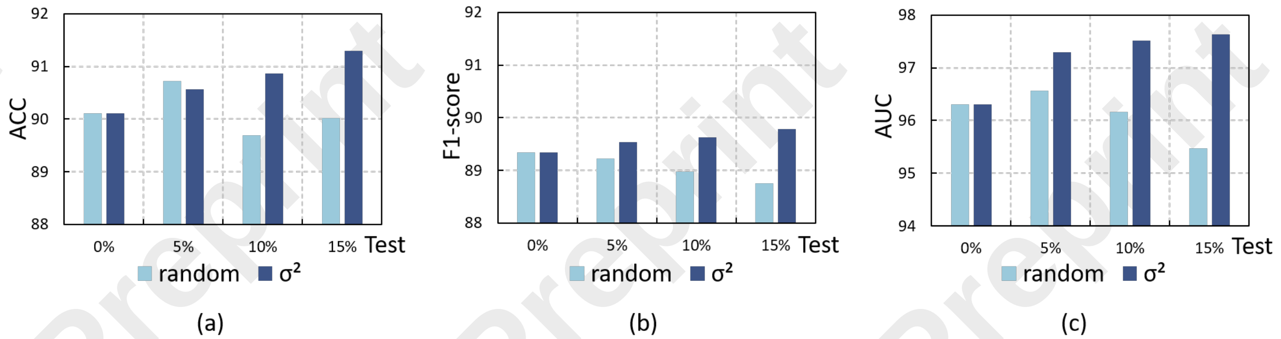

Figure 4: Importance of the variance on the TCGA-NSCLC dataset. Random stands for random deletion, and σ^2 stands for deletion of those with higher variance. (a) Results on ACC. (b) Results on F1-score. (c) Results on AUC.

Dataset	Loss	ACC	F1-score	AUC
Camelyon16	RPMIL	**91.72**	**89.36**	**95.13**
	w/o MSE	85.27	80.85	90.03
	w/o KL	87.60	83.67	91.45
	w/o both	84.50	79.21	87.40
TCGA-NSCLC	RPMIL	**90.49**	**89.34**	**96.31**
	w/o MSE	86.31	85.48	93.88
	w/o KL	87.83	85.96	94.60
	w/o both	85.55	84.92	93.42

Table 3: Impact of loss function limitations of the encoder and decoder on model performance.

4.5 Main Analysis

In this subsection, we primarily address the two main questions posed in the introduction: **i**) Can uncertainty estimation outperform the dominant point estimation MIL methods? **ii**) Can instance distribution and bag distribution jointly enhance classification results?

Comparison between Point Estimation and Uncertainty Estimation ($P(y|\mathbf{z} = g(\mathbf{x}))$ in Equation (3) vs $\int_{\mathbf{z}} P(\mathbf{z}|\mathbf{x})P(y|\mathbf{z})\,\mathrm{d}\mathbf{z}$ in Equation (4). In point estimation methods, multiple instance features are aggregated into a single feature, whereas uncertainty methods aggregate them into a probability distribution. To compare these two types of MIL methods, we conduct experiments with both parameterized aggregators (using vanilla attention) and non-parameterized aggregators (using mean-pooling and max-pooling). The results, as shown in Table 2, reveal that under parameterized aggregators, uncertainty methods significantly outperform point estimation methods. We acknowledge that a high-performance aggregator can enhance the performance of MIL models. However, our experiments reveal that even with a basic pooling method, MIL models can still achieve significant benefits from uncertainty estimation and the joint distribution of instance and bag features. This finding suggests that, despite the focus on aggregator design in MIL, uncertainty estimation methods should be given more attention in practical WSI applications.

Aggregator	Pooling	ACC	F1-score	AUC
Mean	Mean	89.92	86.32	94.62
Mean	Max	88.37	86.02	92.78
Max	Mean	89.15	84.78	93.93
Max	Max	91.47	87.64	92.88
Attention	Mean	**91.72**	**89.36**	**95.13**
Attention	Max	90.70	86.36	93.95

Table 4: Impact of different aggregators and pooling methods on Camelyon16.

Comparison between Two Predictive Distributions ($P(y|\mathbf{z})$ in Equation (4) vs $P(y|\mathbf{z}, \mathbf{x})$ in Equation (5)). The limitation of point estimation methods lies in their sole focus on the aggregated, fixed bag feature representation, overlooking the inherent distribution of instance features within the bag. To better utilize the information from instance feature distribution, we propose a method that jointly utilizes instance distribution and bag feature distribution. In this process, we remove the cross-attention branch and simplify the modeling $P(y|\mathbf{z})$. As shown in Table 2, the experimental results indicate that using only $P(y|\mathbf{z})$ leads to a significant shift in the bag feature distribution. In contrast, our method, which jointly leverages instance distributions and bag feature distributions, significantly improves the model's predictive performance, highlighting the importance of considering instance distributions. The effectiveness of this approach aligns with clinical observations, as tumor patches can confirm the tumor WSI.

4.6 Ablation Results

In this subsection, we investigate the proposed method in more detail, showing the contribution of each component through a series of ablation experiments.

Effect of Sampling Number. In our study, the probabilistic model makes predictions by learning the distribution of bag features, and the sampling number significantly impacts the experimental results. In Figure 3, we test different sampling numbers (1, 10, 100, and 1,000) and find that a larger sampling size effectively improves model performance and

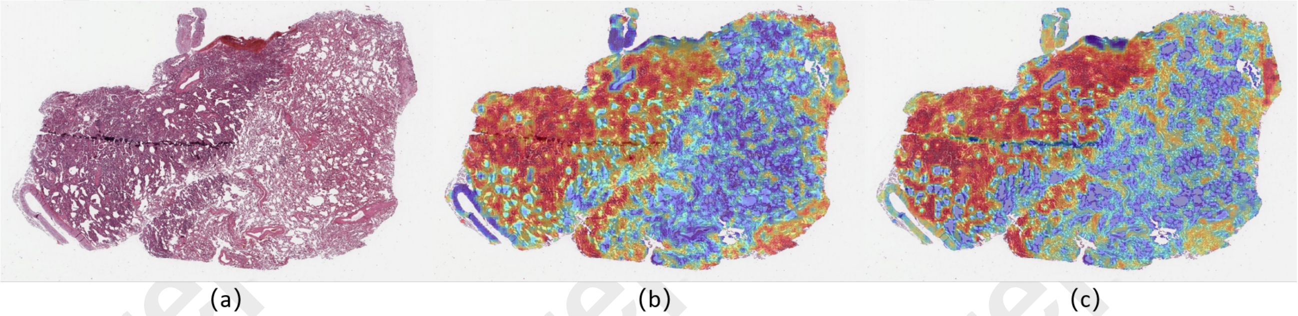

Figure 5: Comparison of attention score visualisations in our Approach and the CLAM Approach. The redder areas represent higher attention scores, which are the pathological tissues that the model attends to. (a) Original WSI. (b) CLAM heatmap. (c) Ours RPMIL heatmap.

stabilizes training. With a sampling size of 1,000, the model obtains more enriched bag representations, achieving the best performance. When sampling sizes are small, the sampled bag features often fail to represent the overall distribution, leading to a significant deviation of the sample mean from the actual center of the bag feature distribution.

Importance of the Variance. Due to the introduction of variance, the results of each iteration will fluctuate, and the degree of fluctuation is related to the magnitude of the variance. We roughly think that larger variance indicates more uncertainty in the bag features, making the model unable to adequately estimate the results of classifying such bag features. We verify the above thought by removing some test samples with a large variance in bag features. We compare this with randomly removing a portion of the sample. We observe that the results become more favorable in Figure 4. The ACC improves by 0.46%, 0.75% and 1.19%, the F1-score improves by 0.20%, 0.29% and 0.44% and the AUC improves by 0.99%, 1.21%, and 1.32% when 5%, 10%, and 15% of the test set with high variance are removed on the TCGA-NSCLC. As the number of removed samples increases, there is an improvement, indicating that variance can be an effective measure of uncertainty to some extent.

Impact of KL and MSE Loss. Compared to the single feature in point estimation, we learn an enriched feature distribution. To prevent significant deviations in the features, we introduce additional KL and MSE loss functions to constrain the bag feature distribution, in addition to using cross-entropy loss. From Table 3, it can be concluded that the absence of either two loss functions degrades the model performance, especially on the Camelyon16. The effect of the missing MSE loss is greater than the KL loss, which keeps the pre-encoding and post-decoding bag features from deviating significantly, while the KL loss contributes to reducing the variance of the classification results.

Combination of Probabilistic Aggregator and Bag Pooling. We compare the effectiveness of learning bag features using three different aggregators: mean-pooling, max-pooling, and vanilla attention-based aggregators. Additionally, there is another bag pooling operation before classification to pool the fusion bag features, and we compare two pooling: mean-pooling and max-pooling. In Table 4, the best combination is the attention-based probabilistic aggregator and the bag mean-pooling. The former is able to learn a fine bag distribution, and the latter enables bag features sampled from the distribution to be near the center of the distribution.

4.7 Visualization

In this subsection, we follow the visualization technique in CLAM to assess whether the attention mechanism in the distribution aggregator is appropriately focused on the positive tissue regions. This evaluation is crucial as it helps to provide further confidence in the reliability of our approach and holds significant promise for real-world clinical applications, where precision and interpretability are paramount for diagnosis. In Figure 5, both our model and the classical CLAM successfully identify the general regions of interest that are deemed positive. While CLAM assigns high attention to areas that are contaminated or irrelevant in the original WSI, our method, enhanced by the incorporation of uncertainty, refrains from considering these areas as positive. This distinction is vital because such contaminated regions, often arising due to issues like uneven staining or sample folding during WSI acquisition, can lead to misclassifications. By preventing these areas from being mistakenly labeled as positive, our approach maintains a higher level of accuracy in classification, demonstrating robustness in the presence of common challenges faced during sample collection and preparation.

5 Conclusion

In this paper, we rethink the process of MIL modeling from the perspective of uncertainty estimation. We propose RPMIL, an uncertainty-aware probabilistic MIL method tailored for WSI pathology diagnosis. Unlike traditional methods that rely on point estimates, RPMIL constructs a bag feature distribution, enabling more excellent results. Our experimental results demonstrate that uncertainty estimation surpasses conventional point estimation approaches in MIL, achieving SOTA performance. Moreover, combining instance feature distribution with bag feature distribution for prediction yields significantly better results than relying on a single distribution alone. We also observe that increasing the number of samples in the bag distribution leads to more pronounced improvements in classification performance. Additionally, the variance in uncertainty serves as a supplementary support for enhancing classification accuracy.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Project 62476089 and the Foundation of Key Laboratory of System Control and Information Processing, Ministry of Education.

References

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

[Cemgil *et al.*, 2020] Taylan Cemgil, Sumedh Ghaisas, Krishnamurthy Dvijotham, Sven Gowal, and Pushmeet Kohli. The autoencoding variational autoencoder. *Advances in Neural Information Processing Systems*, 33:15077–15087, 2020.

[Chen and Lu, 2023] Yuan-Chih Chen and Chun-Shien Lu. Rankmix: Data augmentation for weakly supervised learning of classifying whole slide images with diverse sizes and imbalanced categories. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23936–23945, 2023.

[Chen *et al.*, 2024a] Kaitao Chen, Shiliang Sun, and Jing Zhao. Camil: Causal multiple instance learning for whole slide image classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 1120–1128, 2024.

[Chen *et al.*, 2024b] Richard J Chen, Tong Ding, Ming Y Lu, Drew FK Williamson, Guillaume Jaume, Andrew H Song, Bowen Chen, Andrew Zhang, Daniel Shao, Muhammad Shaban, et al. Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 30(3):850–862, 2024.

[Chikontwe *et al.*, 2020] Philip Chikontwe, Meejeong Kim, Soo Jeong Nam, Heounjeong Go, and Sang Hyun Park. Multiple instance learning with center embeddings for histopathology classification. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part V 23*, pages 519–528. Springer, 2020.

[Chikontwe *et al.*, 2022] Philip Chikontwe, Soo Jeong Nam, Heounjeong Go, Meejeong Kim, Hyun Jung Sung, and Sang Hyun Park. Feature re-calibration based multiple instance learning for whole slide image classification. In *International Conference on Medical Image Computing and Computer-assisted Intervention*, pages 420–430. Springer, 2022.

[Cui *et al.*, 2023] Yufei Cui, Ziquan Liu, Xiangyu Liu, Xue Liu, Cong Wang, Tei-Wei Kuo, Chun Jason Xue, and Antoni B. Chan. Bayes-MIL: A new probabilistic perspective on attention-based multiple instance learning for whole slide images. In *The Eleventh International Conference on Learning Representations*, 2023.

[Haußmann *et al.*, 2017] Manuel Haußmann, Fred A Hamprecht, and Melih Kandemir. Variational bayesian multiple instance learning with gaussian processes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6570–6579, 2017.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[Huang *et al.*, 2023] Zhi Huang, Federico Bianchi, Mert Yuksekgonul, Thomas J Montine, and James Zou. A visual–language foundation model for pathology image analysis using medical twitter. *Nature Medicine*, 29(9):2307–2316, 2023.

[Ilse *et al.*, 2018] Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *International Conference on Machine Learning*, pages 2127–2136. PMLR, 2018.

[Kendall and Gal, 2017] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in Neural Information Processing Systems*, 30, 2017.

[Kingma *et al.*, 2015] Durk P Kingma, Tim Salimans, and Max Welling. Variational dropout and the local reparameterization trick. *Advances in Neural Information Processing Systems*, 28, 2015.

[Li *et al.*, 2021] Bin Li, Yin Li, and Kevin W Eliceiri. Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14318–14328, 2021.

[Li *et al.*, 2024] Hao Li, Ying Chen, Yifei Chen, Rongshan Yu, Wenxian Yang, Liansheng Wang, Bowen Ding, and Yuchen Han. Generalizable whole slide image classification with fine-grained visual-semantic interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11398–11407, 2024.

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

[Lin *et al.*, 2024] Weiping Lin, Zhenfeng Zhuang, Lequan Yu, and Liansheng Wang. Boosting multiple instance learning models for whole slide image classification: a model-agnostic framework based on counterfactual inference. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 3477–3485, 2024.

[Liu *et al.*, 2024] Pei Liu, Luping Ji, Xinyu Zhang, and Feng Ye. Pseudo-bag mixup augmentation for multiple instance learning-based whole slide image classification. *IEEE Transactions on Medical Imaging*, 43(5):1841–1852, 2024.

[Lu *et al.*, 2021] Ming Y Lu, Drew FK Williamson, Tiffany Y Chen, Richard J Chen, Matteo Barbieri, and Faisal Mahmood. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature Biomedical Engineering*, 5:555–570, 2021.

[Lu *et al.*, 2024] Ming Y Lu, Bowen Chen, Drew FK Williamson, Richard J Chen, Ivy Liang, Tong Ding, Guillaume Jaume, Igor Odintsov, Long Phi Le, Georg Gerber, et al. A visual-language foundation model for computational pathology. *Nature Medicine*, 30(3):863–874, 2024.

[Michelucci, 2022] Umberto Michelucci. An introduction to autoencoders. *arXiv preprint arXiv:2201.03898*, 2022.

[Pinckaers *et al.*, 2020] Hans Pinckaers, Bram Van Ginneken, and Geert Litjens. 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.

[Qu *et al.*, 2022] Linhao Qu, Xiaoyuan Luo, Shaolei Liu, Manning Wang, and Zhijian Song. Dgmil: Distribution guided multiple instance learning for whole slide image classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 24–34. Springer, 2022.

[Qu *et al.*, 2023] Linhao Qu, Zhiwei Yang, Minghong Duan, Yingfan Ma, Shuo Wang, Manning Wang, and Zhijian Song. Boosting whole slide image classification from the perspectives of distribution, correlation and magnification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21463–21473, 2023.

[Qu *et al.*, 2024a] Linhao Qu, Kexue Fu, Manning Wang, Zhijian Song, et al. The rise of ai language pathologists: Exploring two-level prompt learning for few-shot weakly-supervised whole slide image classification. *Advances in Neural Information Processing Systems*, 36, 2024.

[Qu *et al.*, 2024b] Linhao Qu, Yingfan Ma, Xiaoyuan Luo, Qinhao Guo, Manning Wang, and Zhijian 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*, page 9732–9744, 2024.

[Rubinstein and Kroese, 2016] Reuven Y Rubinstein and Dirk P Kroese. *Simulation and the Monte Carlo method*. John Wiley & Sons, 2016.

[Schmidt *et al.*, 2023] Arne Schmidt, Pablo Morales-Alvarez, and Rafael Molina. Probabilistic attention based on gaussian processes for deep multiple instance learning. *IEEE Transactions on Neural Networks and Learning Systems*, 35(8):10909–10922, 2023.

[Shao *et al.*, 2021] Zhuchen Shao, Hao Bian, Yang Chen, Yifeng Wang, Jian Zhang, Xiangyang Ji, et al. Transmil: Transformer based correlated multiple instance learning for whole slide image classification. *Advances in Neural Information Processing Systems*, 34:2136–2147, 2021.

[Shi *et al.*, 2024] Jiangbo Shi, Chen Li, Tieliang Gong, Yefeng Zheng, and Huazhu Fu. ViLa-MIL: Dual-scale vision-language multiple instance learning for whole slide image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11248–11258, 2024.

[Tang *et al.*, 2024] Wenhao Tang, Fengtao Zhou, Sheng Huang, Xiang Zhu, Yi Zhang, and Bo Liu. Feature re-embedding: Towards foundation model-level performance in computational pathology. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11343–11352, 2024.

[Tellez *et al.*, 2019] David Tellez, Geert Litjens, Jeroen Van der Laak, and Francesco Ciompi. Neural image compression for gigapixel histopathology image analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43:567–578, 2019.

[Vaswani, 2017] Ashish Vaswani. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017.

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

[Xiong *et al.*, 2023] Conghao Xiong, Hao Chen, Joseph JY Sung, and Irwin King. Diagnose like a pathologist: transformer-enabled hierarchical attention-guided multiple instance learning for whole slide image classification. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pages 1587–1595, 2023.

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

[Zhang *et al.*, 2022] Hongrun Zhang, Yanda Meng, Yitian Zhao, Yihong Qiao, Xiaoyun Yang, Sarah E Coupland, and Yalin Zheng. Dtfd-mil: Double-tier feature distillation multiple instance learning for histopathology whole slide image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18802–18812, 2022.

[Zhang *et al.*, 2023] Ruijie Zhang, Qiaozhe Zhang, Yingzhuang Liu, Hao Xin, Yan Liu, and Xinggang Wang. Multi-level multiple instance learning with transformer for whole slide image classification. *arXiv preprint arXiv:2306.05029*, 2023.

[Zhu *et al.*, 2024] Wenhui Zhu, Xiwen Chen, Peijie Qiu, Aristeidis Sotiras, Abolfazl Razi, and Yalin Wang. Dgrmil: Exploring diverse global representation in multiple instance learning for whole slide image classification. In *European Conference on Computer Vision*, pages 333–351. Springer, 2024.