Medical Image Analysis 97 (2024) 103294

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

# E$^2$-MIL: An explainable and evidential multiple instance learning framework for whole slide image classification

Jiangbo Shi [a], Chen Li [a,*], Tieliang Gong [a], Huazhu Fu [b]

[a] School of Computer Science and Technology, Xi'an Jiaotong University, Xi'an, Shaanxi, 710049, China
[b] Institute of High Performance Computing (IHPC), Agency for Science, Technology and Research (A*STAR), 138632, Singapore

ARTICLE INFO

*Keywords:*
Histopathology
Whole slide image analysis
Multiple instance learning
Uncertainty estimation

ABSTRACT

Multiple instance learning (MIL)-based methods have been widely adopted to process the whole slide image (WSI) in the field of computational pathology. Due to the sparse slide-level supervision, these methods usually lack good localization on the tumor regions, leading to poor interpretability. Moreover, they lack robust uncertainty estimation of prediction results, leading to poor reliability. To solve the above two limitations, we propose an explainable and evidential multiple instance learning (E$^2$-MIL) framework for whole slide image classification. E$^2$-MIL is mainly composed of three modules: a detail-aware attention distillation module (DAM), a structure-aware attention refined module (SRM), and an uncertainty-aware instance classifier (UIC). Specifically, DAM helps the global network locate more detail-aware positive instances by utilizing the complementary sub-bags to learn detailed attention knowledge from the local network. In addition, a masked self-guidance loss is also introduced to help bridge the gap between the slide-level labels and instance-level classification tasks. SRM generates a structure-aware attention map that locates the entire tumor region structure by effectively modeling the spatial relations between clustering instances. Moreover, UIC provides accurate instance-level classification results and robust predictive uncertainty estimation to improve the model reliability based on subjective logic theory. Extensive experiments on three large multi-center subtyping datasets demonstrate both slide-level and instance-level performance superiority of E$^2$-MIL.

## 1. Introduction

Pathological examination plays a crucial role in the early diagnosis and treatment of cancer (Cui and Zhang, 2021). With the rapid advancement of digital scanning devices, traditional pathological slides can be quickly digitized into whole slide images (WSI). As shown in the first column of Fig. 1, WSI is typically enormous in size, consisting of $10^4 \times 10^4$ pixels at the highest magnification (0.25 μm/pixel). This makes obtaining a large number of pixel-level and even region-level annotations extremely time-consuming and error-prone. Consequently, weakly supervised learning methods (Laleh et al., 2022) have been widely adopted in the field of computational pathology, relying solely on WSI-level labels. Multiple instance learning (MIL) (Ilse et al., 2018; Lu et al., 2021; Shao et al., 2021; Schmidt et al., 2023) is the most popular weakly supervised learning paradigm for WSI analysis and it has demonstrated satisfactory performance in WSI-level prediction tasks. In particular, the MIL approaches enable the approximate identification of regions of interest within the WSI, eliminating the need for pixel-level annotation.

The common MIL for digital pathology follows a three-step pipeline: (1) splitting the WSI (*i.e.*, bag) into a series of non-overlapping small patches (*i.e.*, instances) of equal size; (2) using a pre-trained encoder (*e.g.*, ResNet50 He et al., 2016) to extract features of each patch; (3) pooling all the patch-level features into a slide-level representation for the final task. Following this pipeline, recent attention-based MIL methods (Ilse et al., 2018; Lu et al., 2021; Li et al., 2021; Shao et al., 2021; Zhao et al., 2024; Godson et al., 2024; Liu et al., 2024) learn an attention weight for each instance in the pooling step to aggregate all the instance-level representations linearly with different weights. The attention mechanism can improve the slide-level performance and exhibit a rudimentary level of interpretability (*i.e.*, higher attention values represent a higher likelihood to be the cancerous area). Despite showing promising results in various pathological diagnostic tasks, the application of these methods in real-world scenarios is still limited by two following issues:

Firstly, the attention maps generated by the current MIL-based methods lack precise localization ability for entire cancerous regions, leading to poor interpretability. Accurate instance-level classification facilitates the determination of tumor area proportions, enabling precise quantitative diagnosis and providing supporting evidence for

* Corresponding author.
*E-mail addresses:* cli@xjtu.edu.cn (C. Li), hzfu@ieee.org (H. Fu).

https://doi.org/10.1016/j.media.2024.103294
Received 28 February 2024; Received in revised form 25 July 2024; Accepted 30 July 2024
Available online 6 August 2024

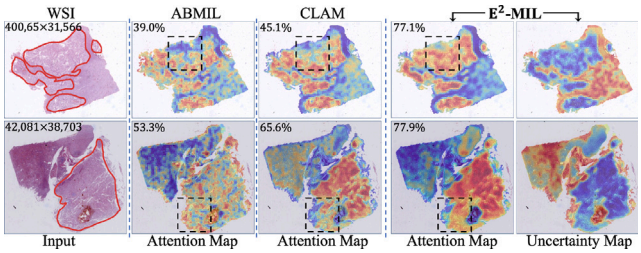


**Fig. 1.** Motivation. The first column shows two whole slide images with tumor regions outlined in red. The second and third columns display the attention maps produced by two existing MIL-based methods (ABMIL Ilse et al., 2018 and CLAM Lu et al., 2021), with the instance-level classification accuracy indicated in the top left corner of each image, respectively. The last two columns present the attention map and uncertainty map generated by our $E^2$-MIL framework.

pathologists during secondary reviews, ultimately enhancing diagnostic efficiency. As depicted in Fig. 1, it is evident that ABMIL (Ilse et al., 2018) and CLAM (Lu et al., 2021) exhibit limitations in their ability to accurately identify positive instances. These approaches tend to identify only partial positive instances and may also produce false-positive results. Specifically, the instance-level classification accuracy is only around 50%, far below the level required by clinical pathologists for model interpretability. A likely explanation is that these models tend to focus on the most discriminative patches related to the slide category, overlooking many other cancerous patches due to the sparse supervision of slide-level labels.

Secondly, existing methods lack robust uncertainty estimation of prediction results, leading to poor reliability. In practice, deep learning models remain vulnerable to perturbations. Even minor artifacts in the data, such as dust, tissue folds, and air bubbles on the slides, can substantially degrade model performance (Brixtel et al., 2022). Moreover, variations in data acquisition procedures across clinics (*e.g.*, staining procedures and scanners) often induce distribution shifts between training and deployment settings. This frequently results in unpredictable model behavior on new data and generates unreliable prediction results (Stacke et al., 2020). Therefore, for the high-risk medical scenario, in addition to the high accuracy and interpretability, the diagnostic model should also be able to notify the pathologists when further examination is required and reject problematic inputs that could lead to erroneous predictions (Pocevičiūtė et al., 2020).

In this work, to overcome the above two limitations, we propose an explainable and evidential multiple instance learning framework ($E^2$-MIL) for whole slide image classification. The primary objective of $E^2$-MIL is to improve the model's interpretability and reliability, thereby enhancing its clinical utility.

To solve the first limitation, we propose a detail-aware attention distillation module (DAM) and a structure-aware attention refined module (SRM). For DAM, we empirically observe that the complementary sub-bags can help the attention-based MIL models discover more positive instances (*i.e.*, tumor regions) compared with taking the whole big bag as the input. Specifically, DAM comprises a global network receiving the whole bag as the input and a local network generating attention maps at a finer grain by taking the local sub-bags as the input. These two networks are constrained by an online distillation loss, which transfers discriminative attention knowledge from the local network to the global network during training. To bridge the gap between the slide-level label and the instance-level classification, we also propose a masked self-guidance loss. Based on the attention map from the global network, instances with higher attention scores are masked in the original bag to build a new masked bag. This masked bag is then fed into the global network again with supervision by the self-guidance loss.

For SRM, we utilize the clustering-based pooling method to aggregate the instances first, significantly reducing instances number and the computational complexity. Then, the spatial relevance between the clustering instances is sufficiently modeled by a vision Transformer (ViT) network (Dosovitskiy et al., 2021). We generate a structure-aware attention map based on the multi-head self-attention map in ViT, helping the framework locate the entire tumor region and further improve the instance-level classification performance. The two complementary attention maps generated by DAM and SRM are combined as the final attention map.

To address the second limitation, we propose an uncertainty-aware instance classifier (UIC) based on the subjective logic theory (Sensoy et al., 2018). UIC generates accurate instance-level classification results and reliable uncertainty estimates to highlight suspicious areas. Specifically, UIC is trained on the pseudo-labels generated from the above final attention map. However, these pseudo-labels usually contain some noise, potentially affecting classifier training. The uncertainty map is efficiently utilized to filter instances with noisy labels during training. In this way, $E^2$-MIL demonstrates significant potential in enhancing model interpretability and reliability by accurately identifying positive instances and providing reliable instance-level uncertainty estimates.

Extensive comparison experiments on three subtyping datasets demonstrate that $E^2$-MIL averagely outperforms the current best results by 4.3% in AUC, 7.6% in ACC and 9.0% in F1 across three datasets, respectively, in the instance-level classification task. The reliability analysis under the settings of domain shift and label noise further indicates that $E^2$-MIL can provide better-calibrated uncertainty estimation and reliable prediction results compared with the current benchmarking uncertainty estimation methods. In summary, the main contributions of this work are as follows:

- We propose an explainable and evidential multiple instance learning ($E^2$-MIL) framework for WSI classification that improves model interpretability and reliability.
- We propose a detail-aware attention distillation module (DAM) and a structure-aware attention refined module (SRM) to improve the model's interpretability by significantly improving the localization ability of positive instances.
- We propose an uncertainty-aware instance classifier (UIC) to improve the model's reliability by providing robust instance-level predictive uncertainty estimation.
- Extensive comparisons and ablation studies on three multi-center subtyping datasets demonstrate that $E^2$-MIL outperforms other state-of-the-art methods for whole slide image classification.

## 2. Related work

### 2.1. Multiple instance learning in WSI

Recently, many MIL-based methods (Ilse et al., 2018; Lu et al., 2021; Li et al., 2021; Zhang et al., 2022; Shao et al., 2021; Chen et al., 2021; Zheng et al., 2022; Shi et al., 2023a; Xiang et al., 2023; Wang et al., 2022; Chikontwe et al., 2022; Shi et al., 2024) have been utilized to solve different kinds of pathological diagnostic tasks in computational pathology, such as cancer subtyping (Lin et al., 2023; Qu et al., 2022), cancer staging (Shi et al., 2023b), tissue segmentation (Xie et al., 2021; Zhang et al., 2023; Jiang et al., 2023), and survival analysis (Di et al., 2022; Shao et al., 2023). For example, Ilse et al. (2018) proposed an attention-based aggregation function by measuring the contribution of each instance embedding in a parameterized way through a neural network. Subsequently, Lu et al. (2021) utilized a ResNet50 model pretrained on the ImageNet to extract the patch features and modeled the WSI classification task as a weakly-supervised slide-level prediction task. Following this design paradigm, a series of embedding-based MIL methods has been proposed. For example, DSMIL (Li et al., 2021) presents a dual-stream architecture with trainable distance measurement and a pyramidal fusion mechanism for multi-scale patch features. DTMIL (Zhang et al., 2022) proposes a double-tier MIL framework

by introducing pseudo-bags to address the challenge of small sample cohorts. TransMIL (Shao et al., 2021) first applies the multi-head self-attention mechanism in Transformer to model the morphological and spatial information. Chen et al. (2021) proposed a context-aware, spatially-resolved patch-based graph convolutional network by aggregating the instance features in the high magnification to model the local topological structures. Graph-Transformer (GTMIL) (Zheng et al., 2022) integrates the graph-based representation and a vision Transformer to fully model the local and global tissue structures in an end-to-end way. Later, IBMIL (Lin et al., 2023) proposes a new scheme based on the backdoor adjustment to achieve the interventional training. Li et al. (2023) proposed an efficient fine-tuning framework for WSI, leveraging on the information bottleneck theory and supervised by slide-level labels. Shi et al. (2023b) introduced a structure-aware hierarchical graph-based multi-instance learning framework for pT staging. While these methods primarily aim to enhance bag-level classification performance, they often struggle with the more challenging instance-level classification task. In contrast, our $E^2$-MIL framework not only excels at delineating the entire tumor regions with high precision but also significantly improves the model's overall reliability.

### 2.2. Predictive uncertainty estimation in WSI

Besides the high accuracy and good interpretability, some recent work (Lakshminarayanan et al., 2017; Gal and Ghahramani, 2016; Chai, 2018; Van Amersfoort et al., 2020; Schmidt et al., 2023) has emphasized the importance of utilizing uncertainty estimation methods to enhance the model's reliability. Current state-of-the-art (SOTA) uncertainty estimation methods are mainly based on ensemble (Lakshminarayanan et al., 2017), dropout (Gal and Ghahramani, 2016), the Bayesian neural network (Chai, 2018), and the deterministic (Van Amersfoort et al., 2020) approaches. In the context of computational pathology, Linmans et al. (2023) conducted a rigorous empirical evaluation of the recent developments in uncertainty estimation on large-scale digital pathology datasets. Dolezal et al. (2022) proposed a clinically-oriented approach to uncertainty quantification for WSIs by estimating uncertainty using dropout and determining thresholds on training data. Mehrtens et al. (2023) first compared the most common methods for uncertainty estimation under the setting of domain shift and label noise. Pocevičiūtė et al. (2022) evaluated the clinical value of uncertainty estimate for deep learning predictions in digital pathology. Although lots of work has evaluated the effect of current benchmarking uncertainty estimation methods in the WSI analysis, most of them rely on pixel-level annotations for training. Several studies (Cui et al., 2023; Schmidt et al., 2023) have also focused on integrating uncertainty estimation with the MIL-based approach for WSI analysis. For example, Bayes-MIL (Cui et al., 2023) proposes a probabilistic MIL framework that induces patch-level uncertainty as a new measurement of interpretability. Schmidt et al. (2023) proposed a probabilistic attention mechanism based on Gaussian processes. However, both methods rely on variational inference to compute the posterior distribution, which is difficult to optimize during training.

Recently, evidential deep learning (Sensoy et al., 2018; Han et al., 2022) has gained attention for uncertainty estimation. Sensoy et al. (2018) utilized the subjective logic theory to treat neural network predictions as subjective opinions for estimating predictive uncertainty. Zou et al. (2022) proposed a trusted brain tumor segmentation network to provide reliable voxel-wise uncertainty estimation. Inspired by these works (Wang et al., 2023), we propose an uncertainty-aware instance classifier based on the subjective logic theory to improve MIL reliability. Our approach does not require expensive patch-level labels and excessive computational burden. Since instance labels are not available during training, we use the predictive uncertainty to filter noises in the generated instance pseudo-labels, stabilizing the training process.

## 3. Methodology

### 3.1. Overview

In this section, we describe the proposed explainable and evidential multiple instance learning framework ($E^2$-MIL) for whole slide image analysis in detail. The framework is presented in Fig. 2. First, we introduce the graph representation learning module as a backbone model to capture the local tissue spatial relations between instances. Then, we present the detail-aware attention distillation module and the structure-aware attention refined module to improve instance-level classification performance. Finally, we present the uncertainty-aware instance classifier, which aims to provide reliable instance-level prediction uncertainty estimations.

### 3.2. Graph representation learning module

In MIL, a WSI $W$ is taken as a bag containing multiple instances $I = \{I_1, I_2, \dots, I_N\} \in \mathbb{R}^{N\times N_0\times N_0\times 3}$, in which $N$ is the instance number and $N_0$ denotes the size of each instance. Following the current embedding-based MIL methods (Ilse et al., 2018; Lu et al., 2021), we utilize the non-overlapping sliding window method to crop a series of patches as instances. Then, a pretrained feature encoder $E(\cdot)$ is utilized to extract patch features as $H = \{H_1, H_2, \dots, H_N\} \in \mathbb{R}^{N\times d_0}$, where $d_0$ denotes the feature dimension of each patch.

To make each instance perceive the spatial relation of local tissues, the WSI $W$ is first modeled as a tissue graph $G$. Specifically, the patch feature matrix $H$ is fed into a linear layer to reduce the dimension of the patch features from $d_0$ to $d$. All patches in $H$ are chosen as nodes in the tissue graph. Then, the node matrix is denoted as $V \in \mathbb{R}^{N\times d}$. Nodes within the graph are interconnected based on the coordinates of the patch center, following an 8-connectivity scheme. This means that inner patches are linked to their eight nearest neighbors, while patches located on the image border have fewer connections due to the missing neighboring patches. The topology of the tissue graph is represented by a binary adjacency matrix $E \in \{0,1\}^{N\times N}$. Finally, the tissue graph is formulated as $G = \{V, E\}$. To capture the local tissue structure, the tissue graph $G$ is passed through a graph convolutional network (GCN) $E_g(\cdot)$ with $T$ GCN layers of dense connections. The message passing of the $t$th GCN layer is formulated as:

$$h_{ij}^t = \mathrm{ReLU}(v_i^t + v_j^t), j \in \mathcal{N}(i), \tag{1}$$

$$m_{ij}^t = \sum_{j\in\mathcal{N}(i)} \frac{\exp(h_{ij}^t)}{\sum_{j'\in\mathcal{N}(i)} \exp(h_{ij'}^t)} \cdot h_{ij}^t, \tag{2}$$

$$v_i^{t+1} = \mathrm{MLP}^t(m_{ij}^t + v_i^t), \tag{3}$$

where $v_*^t \in \mathbb{R}^d$ for $t \in [1, T]$ denotes the input node features of the $t$th GCN layer and $\mathcal{N}(i)$ denotes the neighboring nodes of the $i$th node. Similarly to the attention mechanism, $m_{ij}^t$ is fused by considering the importance of all neighboring nodes. $v_i^{t+1}$ is generated by combining the input with the fused feature $m_{ij}^t$ and then passing through a multi-layer perception $\mathrm{MLP}^t(\cdot)$. After passing through the GCN network, the tissue graph is updated as $G' = \{V' \in \mathbb{R}^{N\times d'}, E\}$, where $d' = T \times d$. Each node in tissue graph $G'$ aggregates information from the local tissue contextual structure.

### 3.3. Detail-aware attention distillation module

The node matrix of the updated tissue graph $G'$ is read out as $X \in \mathbb{R}^{N\times d'}$. The attention-based MIL (ABMIL) (Ilse et al., 2018) and its variants (Li et al., 2021; Lu et al., 2021) utilize the attention-based trainable aggregation function to obtain the bag-level representation, which is formulated as follows:

$$X_i' = \mathbf{W_a} X_i, \tag{4}$$

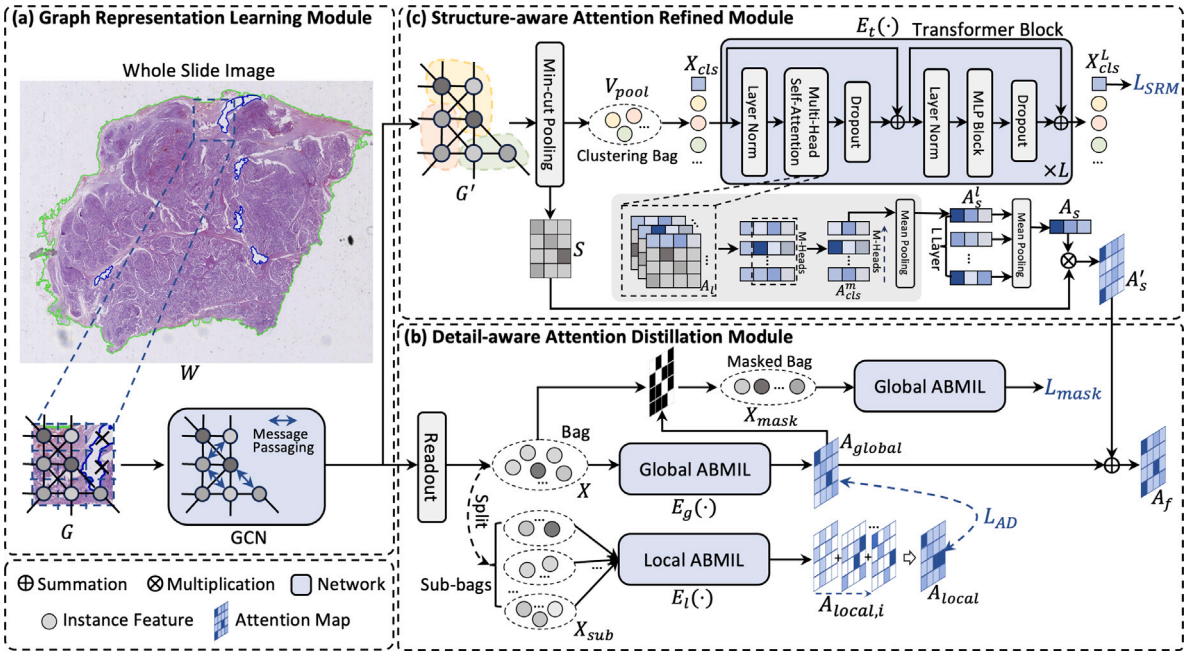


**Fig. 2.** Overview of the proposed E²-MIL Framework. (a) The graph representation learning module projects the whole slide image into a local contextual tissue graph first; (b) the detail-aware attention distillation module employs a simple and effective attention knowledge transfer method to help the attention map locate more detail-aware instances; (c) the structure-aware attention refined module utilizes the min-cut pooling method (Bianchi et al., 2020) to build the clustering bag and generates the structure-aware attention map by utilizing the long-distance modeling ability of multi-head self-attention. These two kinds of complementary attention maps are further fused to generate the final attention map.

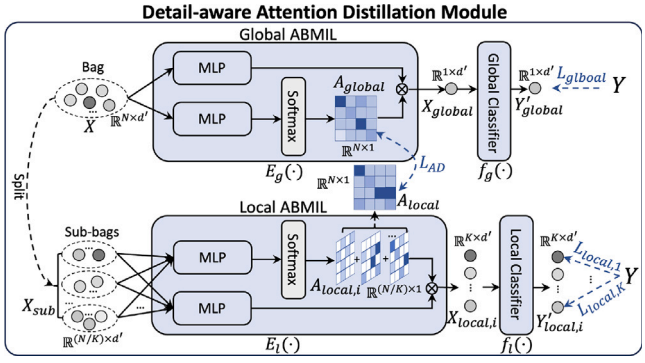


**Fig. 3.** Illustration of the detail-aware attention distillation module.

$$A_i = \frac{\exp\{\mathbf{W}_\mathbf{b}^\top(\tanh(\mathbf{V}X_i'^\top))\}}{\sum_{j=1}^{N}\exp\{\mathbf{W}_\mathbf{b}^\top(\tanh(\mathbf{V}X_j'^\top))\}}, \tag{5}$$

$$X_{bag} = \mathbf{W}_\mathbf{c}\sum_{i=1}^{N} A_i X_i', \tag{6}$$

where $\mathbf{W_a}$, $\mathbf{W_c}$, $\mathbf{V}$ are all trainable weight matrices with the same dimension of $d' \times d'$; $\mathbf{W_b}$ is also a trainable weight matrix with the dimension of $d' \times 1$; $\tanh(\cdot)$ is the activation functions; and $A \in \mathbb{R}^{N\times 1}$ is the attention map that scores the weight of each instance in the bag-level representation. Due to the large number of instances in the bag $X$ and the sparse weak slide-level labels, the attention map $A$ usually can only locate a few most critical positive instances, resulting in poor instance-level classification performance.

Our detail-aware attention distillation module (DAM) leverages complementary sub-bags randomly split from the whole bag to distill detailed attention knowledge from the local network to the global network, as shown in Fig. 3. Specifically, DAM consists of a global $E_g(\cdot)$ and a local $E_l(\cdot)$ ABMIL network, both of which are attention-based aggregation functions with the same structure but different initialization. For the whole bag $X$, we randomly split it into disjoint $K$ sub-bags $X_{sub} = \{X_{s,1}, X_{s,2}, \ldots, X_{s,K}\}$ with an equal number of instances in each sub-bag. The global ABMIL network takes the entire bag $X$ as input and produces a bag-level representation $X_{global}$ and an attention map $A_{global}$. For each sub-bag $X_{s,i}$, the local ABMIL network takes it as the input and produces a sub-bag-level representation $X_{local,i}$ and corresponding attention map $A_{local,i}$. The bag-level representation $X_{global}$ and each sub-bag-level representation $X_{local,i}$ are separately fed to two classifiers $f_g(\cdot)$ and $f_l(\cdot)$ to obtain predictions of each bag $Y'_{global}$ and $Y'_{local,i}$, respectively. Then, the bag-level classification losses for two networks are defined as:

$$L_{global} = \mathrm{CE}(Y'_{global}, Y), \tag{7}$$

$$L_{local} = \frac{1}{K}\sum_{i=1}^{K}\mathrm{CE}(Y'_{local,i}, Y), \tag{8}$$

where $\mathrm{CE}(\cdot,\cdot)$ denotes the cross-entropy loss and $Y$ is the slide label. To transfer the detailed attention knowledge from the local network to the global network, we define an attention distillation loss as follows:

$$L_{AD} = |\cup_{i=1}^{K} A_{local,i} - A_{global}|. \tag{9}$$

Specifically, the local attention maps $A_{local,i}$ for sub-bags are assembled to generate $A_{local}$. The attention distillation loss is formulated by measuring the difference between $A_{global}$ and $A_{local}$. In this way, the global attention map $A_{global}$ can locate more positive instances by accumulating the rich instance detail knowledge in an online way.

Moreover, to bridge the gap between the slide-level label and the instance-level classification task, we propose a masked self-guidance loss, which leverages the global attention map $A_{global}$ to provide more auxiliary supervisory information. Specifically, the masked bag $X_{mask}$ is generated by preserving the instances in bag $X$, whose attention values in $A_{global}$ are no greater than a threshold $m$,

$$X_{mask} = \{X_i : i \in \mathrm{INDEX}(A_{global} \leq m)\}, \tag{10}$$

where INDEX(·) is an operator returning indexes of attention values smaller than or equal to $m$ in $A_{global}$. Then, the representations (*cf.* Eq. (6)) of masked bag $X_{mask}$ are fed into the global network $E_g(\cdot)$ again to produce the bag-level representation $X_m$. The masked self-guidance loss is formulated as follows:

$$L_{mask} = \text{Sigmoid}(f_g(X_m))_{[Y]}, \tag{11}$$

where Sigmoid(·) is the activation function. We mask out the most prominent cancer patches (*i.e.*, positive instances with the highest attention scores) to force the global network to attend to more cancer patches, thereby increasing the coverage of cancer patches in the final attention map. The gap between the sparse slide-level labels and the instance-level classification tasks can be effectively mitigated. Finally, the total loss for DAM is formulated as follows:

$$L_{DAM} = L_{global} + L_{local} + L_{AD} + L_{mask}. \tag{12}$$

### 3.4. Structure-aware attention refined module

After training with the DAM, the attention map $A_{global}$ can locate more positive instances. However, the model lacks perception of the whole tumor region structure. We propose a structure-aware attention refined module (SRM) that utilizes the tissue contextual information to refine the attention map further. Specifically, for the tissue graph $G' = \{V' \in \mathbb{R}^{N\times d'}, E\}$, we first utilize the min-cut pooling method (Bianchi et al., 2020) to significantly reduce the node number,

$$\mathbf{S} = \text{ReLU}(V'W_{pool}), \tag{13}$$

$$V_{pool} = \mathbf{S}^\top V', \tag{14}$$

where $W_{pool} \in \mathbb{R}^{d'\times N_p}$ is a trainable weight matrix; $\mathbf{S} \in \mathbb{R}^{N\times N_p}$ is the assignment matrix for soft node clustering; $V_{pool} \in \mathbb{R}^{N_p\times d'}$ is the aggregated bag after clustering ($N_p \ll N$); and $N_p$ is the number of clusters after min-cut pooling (since all instances in a cluster are aggregated into one, each cluster is then treated as a new "instance" in the aggregated bag). To enhance the quality of instance clustering and ensure that the clustered instances accurately represent various tissue components, we incorporate the unsupervised min-cut pooling loss (Bianchi et al., 2020) for further regularization:

$$L_{mincut} = -\frac{\text{Tr}(\mathbf{S}^\top\widetilde{\mathbf{A}}\mathbf{S})}{\text{Tr}(\mathbf{S}^\top\widetilde{\mathbf{D}}\mathbf{S})} + \left\| \frac{\mathbf{S}^\top\mathbf{S}}{\|\mathbf{S}^\top\mathbf{S}\|_F} - \frac{\mathbf{I_p}}{\sqrt{p}} \right\|_F, \tag{15}$$

where $\|\cdot\|_F$ indicates the Frobenis norm. The first item encourages nodes that are strongly connected to be clustered together and the second item encourages the clusters to have a similar size.

Following the setting of ViT (Dosovitskiy et al., 2021), we concatenate a learnable class token $X_{cls} \in \mathbb{R}^{1\times d'}$ to $V_{pool}$. Then, the input of the Transformer encoder $E_t(\cdot)$ is denoted as $V_{cls} = [X_{cls}||V_{pool}] \in \mathbb{R}^{(N_p+1)\times d'}$. Because each instance in $V_{pool}$ is already local structure-aware, it is not necessary to add the position embedding. The Transformer encoder $E_t(\cdot)$ is composed of $L$ Transformer blocks. Each Transformer block includes a multi-head self-attention (MHSA) layer and a multi-layer perceptron (MLP) with two fully connected layers. The output of the $l$th Transformer block is calculated as:

$$V'_l = \text{MHSA}(\text{LN}(V_{(l-1)})) + V_{(l-1)}, \tag{16}$$

$$V_l = \text{MLP}(\text{LN}(V'_l)) + V'_l, \tag{17}$$

where LN(·) denotes the layer normalization. After passing through the $L$ Transformer blocks, the input feature matrix $V_{cls}$ is updated as $V^L_{cls} \in \mathbb{R}^{(N_p+1)\times d'}$. The updated $cls$ token $X^L_{cls}$ is extracted from $V^L_{cls}$, which represents the whole cluster. Then, the bag-level classification result is obtained by $Y_{bag} = \text{MLP}(X^L_{cls})$. The bag-level classification loss is formulated as follows:

$$L_{cls} = \text{CE}(Y_{bag}, Y). \tag{18}$$

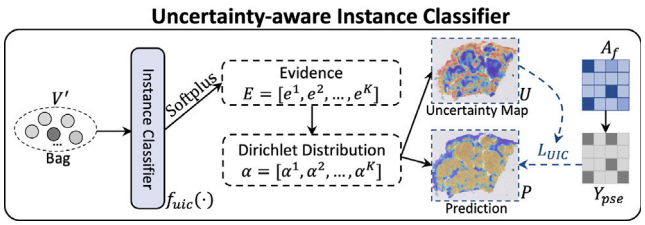


**Fig. 4.** Illustration of the uncertainty-aware instance classifier.

Note that the prediction $Y_{bag}$ is regarded as the bag-level prediction for the WSI $W$. Finally, the total loss for SRM is defined as follows:

$$L_{SRM} = L_{cls} + L_{mincut}. \tag{19}$$

Then, we generate the structure-aware attention map $A_s$ to locate more positive instances by capturing the entire tumor structure. Specifically, suppose the $l$th Transformer block has $M$ heads, $Q$ and $K$ are $d'/M$-dimensional query vectors and key vectors of all tokens. Then, the self-attention weight $A_l \in \mathbb{R}^{(N_p+1)\times(N_p+1)}$ for each head can be calculated by:

$$A_l = \text{softmax}(\frac{QK^T}{\sqrt{d'/M}}). \tag{20}$$

Based on $A_l$, for the $m$th head, we extract the attention matrix $A^m_{cls} \in \mathbb{R}^{1\times N_p}$ between the $cls$ token and all other patch tokens. Then, the $l$th structure-aware attention map $A^l_s$ can be generated by averaging the $A^m_{cls}$ of each head as $A^l_s = \sum_{m=1}^{M} A^m_{cls}/M$. Finally, the structure-aware attention map $A_s$ is calculated by $A_s = \prod_{i=1}^{T} A^l_s$. The assignment matrix $\mathbf{S}$ is further utilized to scale the dimension of $A_s$ to match $A_{global}$:

$$A'_s = \mathbf{S}^\top \cdot A_s. \tag{21}$$

The final attention map $A \in \mathbb{R}^{N\times 1}$ is generated by fusing the global attention map $A_g$ (*i.e.*, $A_{global}$) and the scaled structure-aware attention map $A'_s$, which is formulated as follows:

$$A_f = \frac{A_g - \min(A_g)}{\max(A_g) - \min(A_g)} + \frac{A'_s - \min(A'_s)}{\max(A'_s) - \min(A'_s)}, \tag{22}$$

where max(·) and min(·) denote the maximum and minimum value, respectively.

### 3.5. Uncertainty-aware instance classifier

To provide reliable instance-level prediction results, we further propose an uncertainty-aware instance classifier (UIC) based on the subjective logic theory (Sensoy et al., 2018), as shown in Fig. 4. Specifically, for the $C$-category classification task with $C$+1 mass maps, including $C$ belief mass maps and an uncertainty mass map, which are all non-negative. For each instance, the sum of all the belief mass maps and the uncertainty mass map is 1.

The input of UIC is the updated bag $V'$, which is first passed through an instance classifier $f_{uic}(\cdot)$ and an Softplus(·) activation function to obtain the evidence $E = \text{Softplus}(f_{uic}(V'))$. Then, the evidence $E \in \mathbb{R}^{N\times C}$ is parameterized to the Dirichlet distribution as $\alpha_i = E_i + 1$. Next, the belief masses and the corresponding uncertainty map can be calculated separately as $B_i = E_i/S = (\alpha_i - 1)/S$, and $U_i = C/S$, where $S = \sum_{c=1}^{C}(E^c_i + 1)$ is the Dirichlet intensities. For the $c$-th category, the prediction probability will be higher when more evidence is observed for this category. Otherwise, if the total obtained evidence is less, the uncertainty will be higher. Based on the Dirichlet distribution, the probability for the $i$th instance is computed as $P_i = \alpha_i/S$.

To optimize the UIC, based on attention map $A_f$, we first generate the pseudo-label $Y_{pse} \in \mathbb{R}^{N\times 1}$ for all the instances. If the attention map value in $A_f$ is larger than $t_f$, then the corresponding pseudo-label $Y_{pse}$

**Table 1**
Dataset statistics.

| Cancer type | | Kidney cancer | | Lung cancer |
|---|---|---|---|---|
| Data Source | | TFAH-RCC | TCGA-RCC | TCGA-Lung |
| Number of WSIs | | 480 | 332 | 658 |
| Number of Samples in Each Slide Category | | 186/176/118 | 147/128/57 | 350/ 308 |
| Dataset Split | Training | 288 | 200 | 394 |
| | Validation | 96 | 66 | 132 |
| | Test | 96 | 66 | 132 |
| Number of Instances | | 1,670,362 | 1,286,212 | 1,903,894 |
| Number of Positive/Negative Instances | | 993,984/676,378 | 657,482/ 628,730 | 826,156/1,077,738 |

**Table 2**
Bag-level classification results (presented in %) on TFAH-RCC, TCGA-RCC, And TCGA-Lung subtyping datasets. The best results are marked in bold, and the comparable results are denoted by superscript * based on the paired t-test ($p$-value >0.05).

| Dataset | TFAH-RCC | | | TCGA-RCC | | | TCGA-Lung | | |
|---|---|---|---|---|---|---|---|---|---|
| Metric | AUC | ACC | F1 | AUC | ACC | F1 | AUC | ACC | F1 |
| ABMIL | 87.2 ± 2.8 | 74.6 ± 4.6 | 72.6 ± 5.0 | 94.8 ± 1.6 | 81.8 ± 3.3 | 78.2 ± 5.8 | 87.0 ± 3.3 | 81.0 ± 3.3 | 81.3 ± 3.1 |
| CLAM | 93.5 ± 0.6 | 82.1 ± 2.9 | 82.5 ± 2.9 | 98.4 ± 1.1* | 89.3 ± 5.5 | 87.4 ± 8.8 | 91.3 ± 1.6 | 83.3 ± 1.9 | 83.1 ± 2.0 |
| DSMIL | 90.7 ± 2.0 | 78.3 ± 2.4 | 77.2 ± 2.9 | 95.8 ± 1.1 | 82.1 ± 4.4 | 75.6 ± 11.5 | 84.2 ± 2.9 | 77.8 ± 2.7 | 77.7 ± 2.7 |
| GraphMIL | 93.6 ± 1.4 | 79.6 ± 5.0 | 79.4 ± 4.9 | 98.2 ± 1.6 | 91.5 ± 4.1 | 90.4 ± 4.2 | 92.5 ± 2.0* | 84.8 ± 2.7 | 84.7 ± 2.7 |
| TransMIL | 91.6 ± 1.7 | 79.6 ± 3.5 | 78.9 ± 3.3 | 98.7 ± 1.2* | 90.3 ± 2.5 | 89.1 ± 2.6 | 91.7 ± 1.8 | 84.8 ± 2.0 | 84.7 ± 2.1 |
| DTMIL | 92.0 ± 2.5 | 81.3 ± 4.8 | 80.7 ± 4.7 | 98.1 ± 1.4 | 89.7 ± 5.2 | 89.2 ± 4.8 | 91.0 ± 1.2 | 84.2 ± 3.0 | 84.2 ± 2.5 |
| GTMIL | 92.5 ± 2.0 | 81.0 ± 5.3 | 80.6 ± 5.1 | 97.9 ± 1.2 | 89.1 ± 1.1 | 88.9 ± 1.5 | 92.6 ± 1.4* | 84.5 ± 2.8 | 84.3 ± 2.9 |
| Bayes-MIL | 90.7 ± 2.3 | 81.5 ± 2.8 | 81.8 ± 2.8 | 97.5 ± 1.8 | 88.1 ± 6.9 | 88.0 ± 5.2 | 91.7 ± 0.7 | 85.2 ± 1.2* | 85.1 ± 1.3* |
| MHIM-MIL | 93.8 ± 1.8 | 83.5 ± 4.3 | 82.9 ± 1.5 | 97.7 ± 3.4 | 89.5 ± 4.7 | 90.5 ± 5.5 | 92.5 ± 1.2* | 84.9 ± 2.8* | 85.0 ± 3.6* |
| **E²-MIL** | **95.2 ± 1.2** | **85.2 ± 3.7** | **84.8 ± 3.7** | **98.8 ± 1.3** | **92.7 ± 5.1** | **91.8 ± 5.3** | **92.9 ± 1.5** | **85.9 ± 3.2** | **85.8 ± 3.1** |

is 1; otherwise, $Y_{pse}$ is 0. Then, the instance-level classification loss is defined as:

$$L_{UIC} = \frac{1}{N}\sum_{i=1}^{N} \mathrm{CE}(P_i, Y_{pse,i}). \tag{23}$$

Based on Sensoy et al. (2018), Eq. (23) can be rewritten as follows:

$$L_{UIC} = \frac{1}{N}\sum_{c=1}^{C}\sum_{i=1}^{N} Y^{c}_{pse,i}(\psi(S_i) - \psi(\alpha^{c}_{i})), \tag{24}$$

where $\psi(\cdot)$ denotes the digamma function. Since the noise in the generated pseudo-labels may affect the training process, we further utilize the generated uncertainty map $U_i$ to filter out the noise in the pseudo-label $Y_{pse}$, which makes the training process more stable. Then, the instance-level classification loss is updated as:

$$\begin{aligned} L_{UIC} &= \frac{1}{N}\sum_{c=1}^{C}\sum_{i=1}^{N} M_i \cdot Y^{c}_{pse,i}(\psi(S_i) - \psi(\alpha^{c}_{i})), \\ where\ M_i &= \begin{cases} 1 & if\ U_i < t_u, \\ 0 & otherwise, \end{cases} \end{aligned} \tag{25}$$

where $t_u$ is the threshold to determine whether the pseudo-label is noisy based on the uncertainty map $U_i$. Finally, our framework E²-MIL is trained end-to-end by minimizing the following loss:

$$L = L_{DAM} + L_{SRM} + L_{UIC}. \tag{26}$$

## 4. Experiments

We first briefly describe the collected datasets in Section 4.1. The implementation details and evaluation metrics are described in Section 4.2 and Section 4.3, respectively. Then, Section 4.4 describes the comparison baselines in detail and reports the experiment results. Subsequently, interpretability analysis and reliability analysis are described in Section 4.5 and Section 4.6, respectively. Finally, a series of ablation studies is conducted in Section 4.7 to evaluate the effectiveness of E²-MIL.

### 4.1. Datasets

To validate our E²-MIL framework on the subtyping task, we created three multi-center and multi-cancer datasets, namely TFAH-RCC, TCGA-RCC, and TCGA-Lung. The dataset statistics are reported in Table 1. The specific descriptions of each dataset are as follows:

**TFAH-RCC:** This dataset consists of kidney cancer slides collected by our research team from The First Affiliated Hospital of Xi'an Jiaotong University (TFAH). The slides were stained with hematoxylin and eosin (H&E) and scanned by the KF-PRO-005 digital slice scanner at 20× magnification with 0.5 μm/pixel resolution. There are 480 slides with slide-level subtyping labels. This study has been approved by the institutional ethics review committee of TFAH.

**TCGA-RCC:** To verify E²-MIL on the multi-center data, we collected 332 kidney cancer slides from TCGA.[1] All the slides were also annotated with slide-level subtyping labels and pixel-level tumor region annotations by three professional pathologists. The data in TFAH-RCC and TCGA-RCC are divided into three categories: clear cell (CCRCC), papillary (PRCC), and chromophobe renal cell carcinoma (CRCC).

**TCGA-Lung:** To verify the generalizability of E²-MIL across multiple cancer types, we collected 658 lung cancer slides (TCGA-Lung) from TCGA. The slides were annotated with slide-level subtyping labels and pixel-level tumor region annotations. TCGA-Lung includes two subtypes: lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC).

Note that for each dataset, the pixel-level tumor region annotations were meticulously annotated by three professional pathologists.

### 4.2. Implementation details

The whole tissue in the original WSI is segmented out from the white background by Otsu's binarization algorithm. The z-score normalization is performed for the stain of all the patches extracted from the slides. All patches are cropped at 10× magnification, and the size of each patch is 256 × 256 pixels ($N_0 = 256$). In total, 4.9 million patches are extracted from 1470 WSIs for the experiments. The feature encoder $E(\cdot)$ is implemented as a truncated ResNet50 model (He et al., 2016) pre-trained on ImageNet with an output dimension of 1024 (*i.e.*, $d_0 = 1024$). The reduced feature dimension $d$ is 128. The GCN network $E_g(\cdot)$ has $T = 3$ layers. In DAM, the number of sub-bags $K$ is 5. The threshold

[1] A public cancer data consortium that contains diagnostic WSIs and pathological reports. The download link is http://www.cancer.gov/tcga.

**Table 3**
Instance-level classification results (presented in %) on TFAH-RCC, TCGA-RCC, and TCGA-Lung subtyping datasets. The best results are marked in bold, and the comparable results are denoted by superscript * based on the paired t-test ($p$-value >0.05).

| Dataset | TFAH-RCC | | | TCGA-RCC | | | TCGA-Lung | | |
|---|---|---|---|---|---|---|---|---|---|
| Metric | AUC | ACC | F1 | AUC | ACC | F1 | AUC | ACC | F1 |
| ABMIL | 59.5 ± 2.3 | 46.5 ± 2.5 | 38.3 ± 2.8 | 68.4 ± 1.6 | 54.9 ± 1.9 | 47.3 ± 2.5 | 53.3 ± 1.9 | 52.9 ± 2.3 | 38.2 ± 1.4 |
| CLAM | 58.1 ± 1.0 | 57.8 ± 8.3 | 51.2 ± 8.7 | 72.3 ± 2.1 | 59.5 ± 5.0 | 52.0 ± 6.3 | 57.0 ± 2.7 | 56.7 ± 3.2 | 39.6 ± 1.8 |
| DSMIL | 46.3 ± 1.5 | 40.8 ± 0.9 | 20.2 ± 0.9 | 46.2 ± 1.8 | 37.6 ± 3.1 | 18.6 ± 0.5 | 51.8 ± 2.6 | 48.3 ± 2.9 | 37.7 ± 1.6 |
| GraphMIL | 74.0 ± 3.5 | 59.4 ± 10.4 | 50.9 ± 12.1 | 67.8 ± 2.0 | 66.8 ± 5.3 | 61.4 ± 6.0 | 70.6 ± 3.2 | 65.6 ± 2.1 | 54.7 ± 3.7 |
| TransMIL | 49.0 ± 5.9 | 40.6 ± 2.0 | 29.6 ± 3.1 | 54.2 ± 3.1 | 45.0 ± 2.3 | 34.7 ± 0.7 | 60.2 ± 4.6 | 54.6 ± 3.4 | 35.1 ± 1.4 |
| DTMIL | 72.3 ± 3.9 | 57.9 ± 5.4 | 50.9 ± 8.7 | 66.7 ± 2.9 | 59.4 ± 3.4 | 60.9 ± 3.7 | 62.3 ± 4.9 | 55.6 ± 4.4 | 52.3 ± 2.4 |
| GTMIL | 64.5 ± 8.7 | 58.4 ± 9.7 | 50.8 ± 10.5 | 72.7 ± 2.2 | 64.0 ± 3.5 | 57.4 ± 4.0 | 59.2 ± 4.9 | 56.1 ± 5.6 | 41.4 ± 3.8 |
| Bayes-MIL | 71.5 ± 1.8 | 57.4 ± 1.1 | 49.3 ± 6.5 | 71.0 ± 1.2 | 57.8 ± 1.5 | 55.0 ± 1.0 | 68.5 ± 3.0 | 61.9 ± 2.8 | 48.2 ± 1.2 |
| MHIM-MIL | 72.3 ± 5.6 | 54.7 ± 5.9 | 51.0 ± 7.8 | 72.2 ± 2.9 | 65.8 ± 3.9 | 60.2 ± 2.4 | 69.3 ± 3.9 | 64.7 ± 3.5 | 53.9 ± 2.4 |
| **E$^2$-MIL** | **78.9 ± 2.6** | **76.4 ± 1.9** | **69.4 ± 2.0** | **79.6 ± 1.2** | **71.1 ± 1.8** | **66.3 ± 2.2** | **71.8 ± 3.5** | **67.2 ± 1.3** | **58.6 ± 3.2** |

$m$ for building the masked bag is 0.5. In SRM, the instance number $N_p$ after min-cut pooling is 64. The Transformer encoder $E_t(\cdot)$ includes $L = 3$ Transformer blocks. The head number $M$ in each Transformer block is 8. In UIC, the category number $C$ is 2, which represents the cancer and normal instances, respectively. The instance classifier $f_{uic}(\cdot)$ consists of two fully connected layers. The threshold of $t_f$ and $t_u$ are 0.7 and 0.8, respectively. Note that all hyper-parameters are heuristically tuned on the TFAH-RCC dataset and applied to the other datasets. We adopt the Adam optimizer (Kingma and Ba, 2015) with a learning rate of $1\times10^{-4}$ and weight decay of $1\times10^{-5}$. The minimum training epoch number is 50, and the early stopping strategy is adopted if the accuracy does not continuously increase for 20 epochs. The batch size is 1. Our framework E$^2$-MIL and all the baselines are implemented with PyTorch and the PyG (Fey and Lenssen, 2019) library on a workstation equipped with eight NVIDIA 2080-Ti GPUs.

### 4.3. Evaluation metrics

We report the area under the curve (AUC), accuracy (ACC), and F1 score (F1) to measure the classification performance at the bag and instance level. To evaluate the model's reliability, the expected calibration error (ECE) (Guo et al., 2017) is also adopted by measuring the uncertainty calibration. The prediction confidence (*i.e.*, one minus the uncertainty) for each instance should, on average, match the correctness of the prediction:

$$\mathrm{ECE} = \sum_{i=1}^{M} \frac{|B_m|}{n} \left|\mathrm{conf}(B_m) - \mathrm{acc}(B_m)\right|, \tag{27}$$

where $|B_m|$ denotes the number of instances in the $m$th bin; $n$ is the total number of instances; $\mathrm{conf}(\cdot)$ and $\mathrm{acc}(\cdot)$ denote the average confidence and accuracy, respectively, in the $m$th bin; and $M$ is the predetermined number of bins ($M = 20$ in the experiment). Five-fold cross-validation is adopted to evaluate the model's performance, and each metric's mean and standard deviation are calculated based on the five-fold results. The paired t-test is also adopted to determine the statistical comparability of two paired results. Whenever the $p$-value is $> 0.05$, the two results are comparable.

### 4.4. Comparisons with state-of-the-art

We compare our framework E$^2$-MIL with several baseline WSI classification methods, including:

**ABMIL** (Ilse et al., 2018): It proposes an attention-based aggregation function to generate the bag-level representation by measuring the importance of each instance;

**CLAM** (Lu et al., 2021): It proposes a global pooling operator trained for weakly-supervised slide-level prediction tasks;

**DSMIL** (Li et al., 2021): It utilizes the concatenated multi-scale patches as the input and proposes a non-local operation to aggregate all the instances;

**GraphMIL** (Chen et al., 2021): It proposes a context-aware patch-based convolutional network that aggregates instance features to model the topological structures (Note that this method is denoted as Patch-GCN in Chen et al. 2021);

**TransMIL** (Shao et al., 2021): It utilizes the Transformer model to explore the spatial relations between instances;

**DTMIL** (Zhang et al., 2022): It proposes a double-tier MIL framework by introducing the concept of pseudo-bags;

**GTMIL** (Zheng et al., 2022): It utilizes a graph-based representation of a WSI and a vision Transformer to aggregate all instances.

**Bayes-MIL** (Cui et al., 2023): It proposes a multiple instance framework from a probabilistic perspective and utilizes the induced patch-level uncertainty as the measure of MIL interpretability.

**MHIM-MIL** (Tang et al., 2023): This framework utilizes a Siamese structure with a consistency constraint to effectively mine hard instances.

For each method, we conduct experiments five times. During each run, we randomly split the datasets into training, validation, and test sets following a prescribed proportion of 4:3:3. We then calculate the mean and standard deviation values from the results of these five iterations. Besides, all the methods utilize the same patch feature extraction method and training hyperparameters. Tables 2 and 3 report the bag-level and instance-level classification results of all methods on three subtyping datasets, respectively.

#### 4.4.1. Bag-level classification task

As shown in Table 2, compared with the state-of-the-art WSI classification methods in computational pathology, E$^2$-MIL achieves superior performance on all three datasets in all metrics. Specifically, compared with current best methods, E$^2$-MIL achieves an improvement of 1.4% 1.7% and 1.9% on the metrics of AUC, ACC, and F1, respectively, across all three datasets. This indicates its effectiveness in handling various cancer subtyping tasks. The integration of the graph representation learning module and the structure-aware attention refined module enables the effective capture of both local contextual information among small patches and global tissue-level spatial relations among aggregated clusters. Moreover, the detail-aware attention distillation module enhances the model's discriminatory power at the bag level by leveraging knowledge transfer from randomly split sub-bags under the guidance of online distillation loss.

#### 4.4.2. Instance-level classification task

In addition to the high slide-level classification performance on the subtyping task, E$^2$-MIL also significantly improves the instance-level classification performance compared with all the current best methods. Specifically, as shown in Table 3, E$^2$-MIL averagely outperforms the second-best methods by 4.3% in AUC, 7.6% in ACC and 9.0% in F1 across three datasets. This indicates E$^2$-MIL's superior ability to identify comprehensive positive tumor regions accurately. The fused attention map, generated by the detail-aware attention distillation module and the structure-aware attention refined module, effectively captures detailed instances and the global tissue spatial

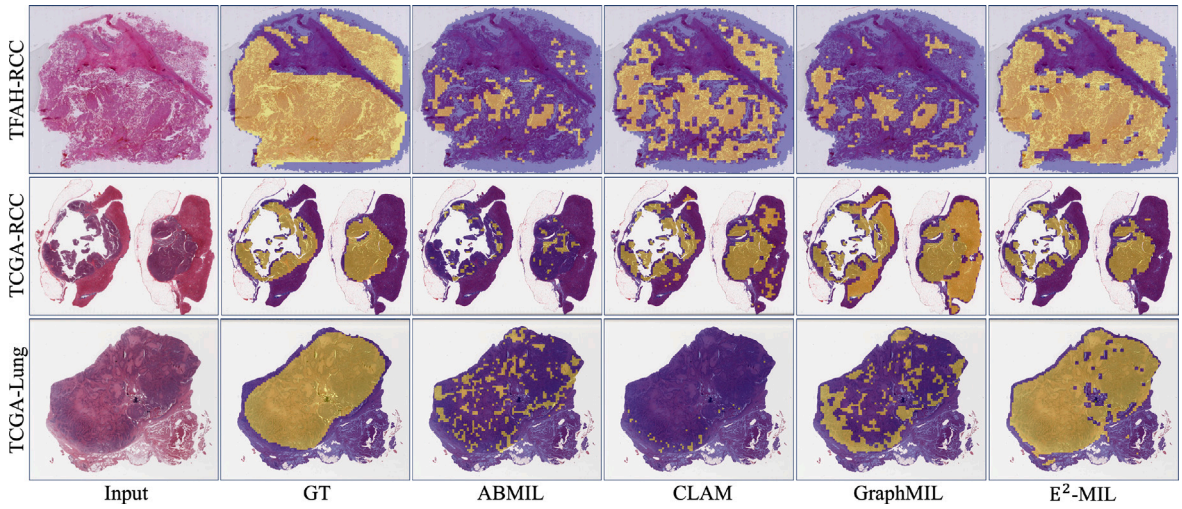


**Fig. 5.** Instance-level classification results (yellow for cancer and purple for normal region) of a sample from the TFAH-RCC and TCGA-RCC datasets.

**Table 4**
Results (presented in %) of multi-center cross-evaluation between TFAH-RCC and TCGA-RCC datasets.

| Setting | TFAH-RCC → TCGA-RCC | | | | TCGA-RCC → TFAH-RCC | | | |
|---|---|---|---|---|---|---|---|---|
| Task | Bag-level | | Instance-level | | Bag-level | | Instance-level | |
| Metric | ACC | F1 | ACC | F1 | ACC | F1 | ACC | F1 |
| CLAM | 79.4 ± 3.9 | 74.3 ± 3.8 | 55.4 ± 2.2 | 50.3 ± 3.1 | 75.6 ± 4.7 | 73.7 ± 4.1 | 47.2 ± 6.5 | 36.4 ± 5.4 |
| GraphMIL | 76.2 ± 1.6 | 71.9 ± 1.7 | 63.3 ± 6.8 | 56.2 ± 9.1 | 71.5 ± 3.3 | 68.9 ± 4.8 | 52.9 ± 8.7 | 46.3 ± 7.5 |
| TransMIL | 78.2 ± 3.7 | 75.9 ± 3.2 | 44.9 ± 1.2 | 32.5 ± 2.7 | 72.4 ± 4.5 | 71.9 ± 3.5 | 35.3 ± 3.7 | 25.9 ± 4.3 |
| GTMIL | 77.1 ± 7.0 | 74.8 ± 3.8 | 57.2 ± 4.2 | 50.7 ± 5.9 | 72.9 ± 5.2 | 69.2 ± 3.4 | 45.8 ± 7.9 | 38.7 ± 6.4 |
| Bayes-MIL | 80.6 ± 6.3 | 78.5 ± 5.9 | 55.6 ± 3.2 | 42.5 ± 3.6 | 77.9 ± 6.3 | 74.5 ± 6.2 | 44.3 ± 5.4 | 39.6 ± 7.4 |
| **E²-MIL** | **83.8 ± 5.6** | **80.5 ± 5.3** | **68.7 ± 3.0** | **63.4 ± 2.5** | **79.3 ± 4.9** | **77.8 ± 5.1** | **62.5 ± 4.2** | **55.3 ± 3.5** |

structures. Additionally, the uncertainty-aware classifier precisely filters out noisy instance pseudo-labels with high uncertainty values, significantly boosting instance-level performance.

#### 4.4.3. Multi-center cross-evaluation experiments

To evaluate the domain adaption ability of our E²-MIL framework, we conduct multi-center cross-evaluation experiments between the TFAH-RCC and TCGA-RCC datasets. The "TFAH-RCC → TCGA-RCC" denotes that the model is trained in the TFAH-RCC dataset and tested in the TCGA-RCC dataset. The "TCGA-RCC → TFAH-RCC" denotes that the model is trained in the TCGA-RCC dataset and tested in the TFAH-RCC dataset. The experiment results are reported in Table 4. Specifically, the performances of all methods have been affected by variations in data distribution. However, compared to the current best baselines, our E²-MIL achieves superior ACC and F1 performances under both domain adaption settings. This underscores the superior domain adaption ability of our E²-MIL, demonstrating its robustness to variations in data distribution across multiple centers.

#### 4.4.4. More experiment results on Camelyon16 dataset

To further evaluate the performance of our E²-MIL framework, we also conduct experiments on the more challenging Camelyon16 dataset. The split ratio for training, validation, and test datasets is 6:2:2. Besides, the data preprocessing steps and training settings are the same as the other datasets. The specific experiments are reported in Table 5. Compared to several state-of-the-art baselines, our E²-MIL framework outperforms them significantly in both bag-level and instance-level classification tasks. Specifically, for the bag-level classification task, the improvements are 1.7%, 1.5%, and 2.6% in the metrics of AUC, ACC, and F1, respectively. For the instance-level classification task, the improvements are 2.4% and 4.4% in the metrics of ACC and F1, respectively. Despite the slides of metastatic cancer posing challenges in locating the entire tumor regions, our E²-MIL framework still effectively

**Table 5**
Classification results (presented in %) on the Camelyon16 dataset. The best results are marked in bold, and the comparable results are denoted by superscript * based on the paired t-test (*p*-value >0.05).

| Task | Bag-level | | | Instance-level | |
|---|---|---|---|---|---|
| Metric | AUC | ACC | F1 | ACC | F1 |
| ABMIL | 76.0 ± 7.5 | 76.3 ± 4.6 | 73.0 ± 6.8 | 62.1 ± 5.1 | 40.2 ± 1.7 |
| CLAM | 79.8 ± 3.0 | 78.5 ± 3.2 | 77.8 ± 4.2 | 64.3 ± 5.0 | 42.4 ± 2.5 |
| GraphMIL | 76.9 ± 7.2 | 75.5 ± 6.1 | 72.5 ± 8.0 | 75.8 ± 1.3 | 67.7 ± 3.4 |
| TransMIL | 76.5 ± 6.3 | 71.3 ± 2.6 | 69.6 ± 2.5 | 68.8 ± 0.9 | 59.7 ± 1.2 |
| Bayes-MIL | 82.7 ± 4.5 | 80.8 ± 2.6 | 78.0 ± 3.4 | 74.9 ± 1.4 | 67.2 ± 1.5 |
| **E²-MIL** | **84.4 ± 3.7** | **82.3 ± 1.2** | **80.6 ± 1.7** | **78.2 ± 0.7** | **72.1 ± 1.3** |

mines the hard-to-classified instances during the training with the help of the proposed DAM and SRM. Moreover, as illustrated in Fig. 6, the generated attention map accurately locates entire tumor regions. Additionally, the uncertainty maps highlight the suspicious area for further examination by pathologists, thereby enhancing the clinical utility of our framework.

#### 4.4.5. Data efficiency analysis

In clinical scenarios, challenges like patient privacy concerns and the rarity of certain diseases make collecting large datasets difficult. To thoroughly evaluate model performance under such constraints, we adopt the same TFAH-RCC data split used in prior experiments. Then, we randomly select 50% of the TFAH-RCC training set to train all models. As reported in Table 6, the performance of all models has decreased due to the reduced number of bag-level training samples. However, E²-MIL consistently significantly outperforms all comparative methods in both bag-level and instance-level classification tasks by effectively leveraging the weak bag-level labels. This demonstrates E²-MIL's robustness in data-limited environments.

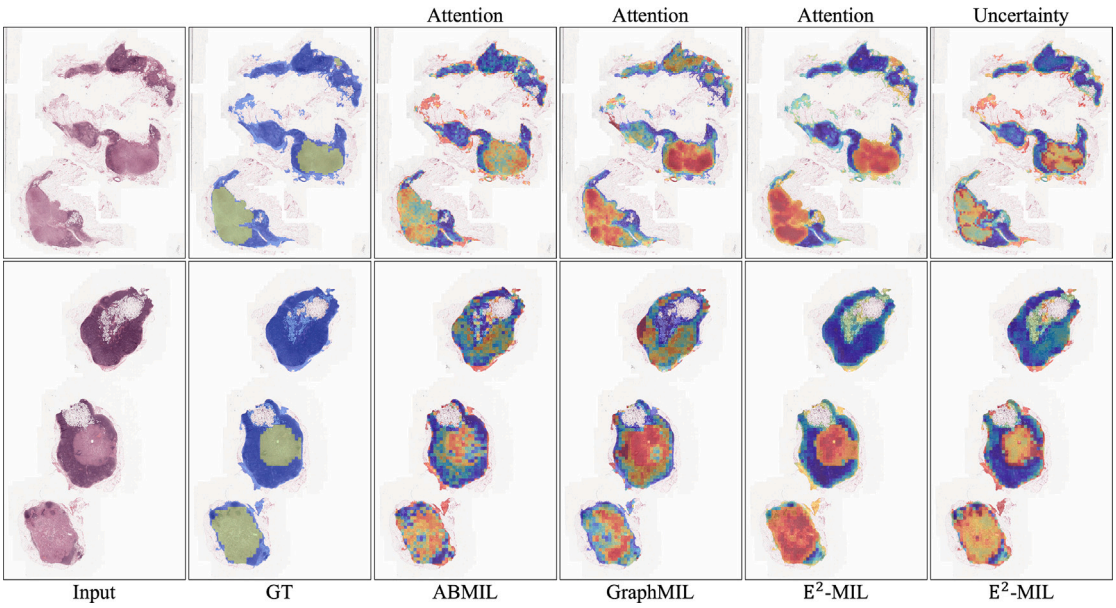


**Fig. 6.** Qualitative attention results of two samples from the Camelyon16 dataset. The first and second columns display the original WSI and corresponding pixel-level annotations. The third and fourth columns present attention maps of two baselines. The last two columns on the right show the attention map and uncertainty map of our E$^2$-MIL framework.

**Table 6**
Classification results (presented in %) on the TFAH-RCC subtyping dataset with 50% training data. The best results are marked in bold, and the comparable results are denoted by superscript * based on the paired t-test ($p$-value >0.05).

| Task | Bag-level | | | Instance-level | |
|---|---|---|---|---|---|
| Metric | AUC | ACC | F1 | ACC | F1 |
| ABMIL | 86.0 ± 2.0 | 69.8 ± 5.0 | 59.5 ± 11.1 | 46.6 ± 3.7 | 37.7 ± 4.4 |
| CLAM | 90.3 ± 1.7 | 78.8 ± 2.7 | 78.1 ± 2.7 | 52.9 ± 9.0 | 45.2 ± 10.5 |
| DSMIL | 87.8 ± 2.9 | 73.3 ± 3.4 | 71.0 ± 4.6 | 28.0 ± 1.5 | 15.7 ± 0.7 |
| GraphMIL | 91.3 ± 2.1* | 79.0 ± 6.5 | 78.5 ± 6.5 | 56.7 ± 8.9 | 48.0 ± 10.3 |
| TransMIL | 89.4 ± 2.5 | 76.9 ± 4.0 | 75.6 ± 4.6 | 40.5 ± 1.3 | 29.5 ± 3.0 |
| DTMIL | 90.2 ± 2.2 | 77.9 ± 3.0 | 76.6 ± 4.5 | 51.5 ± 6.3 | 41.7 ± 7.0 |
| GTMIL | 90.3 ± 1.7 | 77.1 ± 2.9 | 76.7 ± 2.7 | 53.4 ± 6.8 | 45.6 ± 8.7 |
| **E$^2$-MIL** | **91.8 ± 2.2** | **81.3 ± 4.2** | **80.9 ± 4.0** | **69.0 ± 5.1** | **61.1 ± 5.1** |

**Table 7**
Comparison of the computational costs of state-of-the-art MIL-based classification methods measured in the test set of the TFAH-RCC dataset, which includes average inference time, FLOPs, and trainable model parameters.

| Methods | Average inference time (s) | FLOPs (G) | Paramters (M) |
|---|---|---|---|
| CLAM | 0.035 | 2.746 | 0.789 |
| GraphMIL | 0.097 | 3.333 | 0.955 |
| TransMIL | 0.043 | 11.688 | 3.200 |
| GTMIL | 0.102 | 1.809 | 5.165 |
| Bayes-MIL | 0.036 | 1.830 | 0.525 |
| E$^2$-MIL | 0.121 | 2.728 | 5.429 |

#### 4.4.6. Computational cost analysis

To evaluate the computational cost of our E$^2$-MIL framework, we conduct a computational cost analysis on the test set of the TFAH-RCC dataset. Three metrics are measured, including average inference time per slide (measured in seconds, s), FLOPs (measured in billions, G), and the model's trainable parameters (measured in millions, M). The specific experiment results are reported in Table 7. Compared with several selected baselines, the computation cost of our E$^2$-MIL framework is comparable in all metrics. Despite the inclusion of DAM, SRM, and UIC in our E$^2$-MIL framework, which slightly increases the number of learnable parameters, the significant performance enhancements observed in both the bag-level and instance-level classification tasks justify the associated computational costs as acceptable.

### 4.5. Interpretability analysis

To show the interpretability of E$^2$-MIL, we visualize the instance-level prediction results. As shown in Fig. 5, we selected one case from all three datasets. The instance-level predictive results of three baseline models are also visualized. The results show that current MIL-based methods usually can only locate a small number of critical positive instances and may also produce some false-positive instances that affect the diagnostic process. Compared with ground truth (GT), E$^2$-MIL exhibits a high degree of interpretability, as it is capable of accurately identifying and locating nearly all positive instances. (Note that some white blocks in the prediction results of E$^2$-MIL are regions of blood vessels with white backgrounds.) This illustrates that E$^2$-MIL can achieve more fine-grained interpretability and precise boundaries of tumor regions. As shown in Fig. 7, we also visualize the attention maps generated by each module. The global attention map $A_{global}$ locates detailed positive instances, and the structure-aware attention map $A'_s$ learns the whole structure of the tumor region well. The fused attention map $A_f$, generated by fusing these two complementary attention maps, obtains a better localization ability for tumor regions. Moreover, the uncertainty map $U$ showcases all the potentially suspicious regions, with a majority of them being the boundaries of tumor regions, requiring further review by pathologists.

### 4.6. Reliability analysis

To evaluate the reliability of our proposed uncertainty-aware instance classifier (UIC), we compare UIC with the current benchmarking uncertainty estimation methods (Gal and Ghahramani, 2016; Lakshminarayanan et al., 2017; Guo et al., 2017; Linmans et al., 2023) under the scenarios of domain shift and label noise. To ensure the comparison fairness, all the methods utilize the same baseline classifier with two fully connected layers. The details of selected uncertainty estimation methods are described as follows:

**MC Dropout** (Gal and Ghahramani, 2016): The dropout operation can be regarded as a regularization method to stabilize model training. Multiple forward passes of the same input with activated dropout layers are utilized. Based on the baseline classifier, a dropout layer with a drop rate of 0.5 is added between two fully connected layers to implement the MC Dropout method.

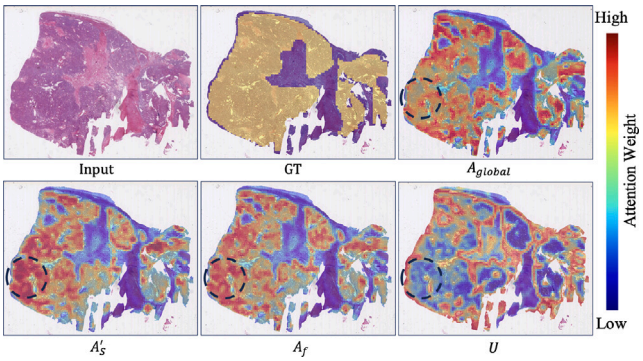


**Fig. 7.** Attention maps. GT denotes the ground truth mask of cancer (yellow) and non-cancer (purple) region; $A_{global}$ is the global attention map from DAM; $A'_s$ is the structure-aware attention map from SRM; $A_f$ is the combined attention map; and $U$ denotes the uncertainty map.

**Deep Ensemble** (Lakshminarayanan et al., 2017): This method consists of multiple models with the same structure but different random initializations. The mean of the predictions of all models is taken as the final output. Three baseline classifiers are ensembled to implement the Deep Ensemble method.

**Multi-head Ensemble** (Linmans et al., 2023): It is an improved version of the Deep Ensemble method. Based on the baseline classifier, the first connected layer is shared between different models, and only the last fully connected layer is ensembled. The head number is set to three in our experiments.

**Temperature Scaling** (Guo et al., 2017): It introduces a temperature scaling factor (set to 1.5 in our experiments) on the model's output that improves the model calibration.

#### 4.6.1. Domain shift

To evaluate the classifier's generalization ability across multiple centers, we use the TFAH-RCC dataset to train all classifiers and directly test on the TCGA-RCC dataset. (Note that the instance-level labels are utilized to train all classifiers to avoid the effect of label noise on the results.) As shown in Table 8, compared with the other uncertainty estimation methods, our proposed UIC achieves the best results under all three metrics. Specifically, the best improvement is 0.60%, 0.63%, and 0.81% in metrics of ACC, F1, and ECE, respectively, compared with the best baselines, which indicates that UIC has better generalization ability across domains.

#### 4.6.2. Label noise

To simulate label noise, we utilize the instance pseudo label $Y_{pse}$ generated from the attention map $A_f$ to train all the classifiers on the TFAH-RCC dataset. As shown in Table 9, UIC achieves significant improvements of 1.01% and 2.64% in the metrics of F1 and ECE, respectively. In the ACC metric, UIC is comparable with the best method (*i.e.*, Deep Ensemble). These experimental results indicate that UIC is more robust and reliable, which can effectively mitigate the impact of label noise and stabilize the training process. Given the large amount of noise in the generated pseudo-label, the training of all classifiers has been significantly affected. In UIC, we introduce an uncertainty mass map leveraging the subjective logic theory. By transforming the input bag into obtained evidence and parameterizing it into the Dirichlet distribution, the calculated uncertainty map effectively filters out the noise in the pseudo-label and improves the model's performance.

**Table 8**
Experiment results (presented in %) of different uncertainty estimation methods under domain shift. The best results are marked in bold, and the comparable results are denoted by superscript * based on the paired t-test ($p$-value >0.05).

| Setting | TFAH → TCGA | | |
|---|---|---|---|
| Metric | ACC ↑ | F1 ↑ | ECE ↓ |
| MC Dropout | 80.65 ± 1.29 | 76.88 ± 1.47 | 5.49 ± 1.41 |
| Deep Ensemble | 80.71 ± 1.22 | 76.94 ± 1.43 | 5.31 ± 1.28 |
| Multi-head Ensemble | 80.67 ± 1.20 | 76.89 ± 1.38 | 6.08 ± 1.48 |
| Temperature Scaling | 80.63 ± 1.26 | 76.85 ± 1.46 | 5.10 ± 1.18 |
| **UIC** | **81.31 ± 1.22** | **77.57 ± 1.40** | **4.29 ± 0.28** |

**Table 9**
Experiment results (presented in %) of different uncertainty estimation methods under label noise. The best results are marked in bold, and the comparable results are denoted by superscript * based on the paired t-test ($p$-value >0.05).

| Dataset | TFAH-RCC | | |
|---|---|---|---|
| Metric | ACC ↑ | F1 ↑ | ECE ↓ |
| MC Dropout | 74.27 ± 3.48 | 66.20 ± 5.99 | 21.19 ± 3.64 |
| Deep Ensemble | **76.96 ± 3.33** | 68.41 ± 5.22 | 19.33 ± 5.36 |
| Multi-head Ensemble | 76.57 ± 3.19* | 67.65 ± 6.09 | 20.05 ± 5.59 |
| Temperature Scaling | 73.97 ± 8.50 | 66.79 ± 9.52 | 19.69 ± 6.77 |
| **UIC** | 76.41 ± 1.92* | **69.42 ± 1.97** | **16.69 ± 3.41** |

### 4.7. Ablation studies

#### 4.7.1. Effect of each module in $E^2$-MIL

To evaluate the impact of each module in $E^2$-MIL, we perform an ablation study on the TFAH-RCC and TCGA-RCC datasets, as shown in Table 10. The specific module ablation settings are described as follows:

- w/o $L_{AD}$: The attention distillation loss is removed from the objective function $L_{DAM}$ for the DAM to evaluate the impact of attention distillation loss on locating more positive instances.
- w/o $L_{mask}$: The mask self-guidance loss is removed from the objective function $L_{DAM}$ for the DAM to evaluate the impact of auxiliary supervisory information.
- w/o $L_{SRM}$: The structure-aware attention refined module (SRM) is removed. The bag-level prediction is replaced by the output $Y'_{global}$ of the global ABMIL network in DAM. Additionally, the structure-aware attention map $A'_s$ is also removed in the final attention map $A_f$.
- w/o $L_{UIC}$: The UIC classifier is removed and the pseudo-label $Y_{pse}$ is taken as the instance-level prediction result.

As reported in Table 10, after removing the graph representation learning module (*i.e.*, w/o GCN), both bag-level and instance-level classification performance have slightly decreased. This underscores the effectiveness of the graph representation learning module as the backbone in capturing local patch contextual information. After removing the attention distillation loss $L_{AD}$ in DAM, the instance-level performances drop significantly in all metrics, which indicates that the complementary local sub-bags can help the global network learn more detailed attention knowledge. After removing the masked self-guidance loss $L_{mask}$, all the metrics decrease slightly. This illustrates that masking the instances with high attention can help bridge the gap between sparse slide-level labels and instance-level classification tasks, further improving the model performance.

After removing the structure-aware attention refined module $L_{SRM}$, the instance-level metrics are decreased by 12.8% and 17.9% in metrics of ACC and F1, respectively, in the TCGA-RCC dataset, which indicates that $A'_s$ can help locate more structure-aware positive instances and the attention maps $A_g$ and $A'_s$ are complementary to each other. After removing the uncertainty-aware instance classifier $L_{UIC}$, both the bag-level and instance-level performances decrease, indicating that UIC can enhance the accuracy and reliability of the model.

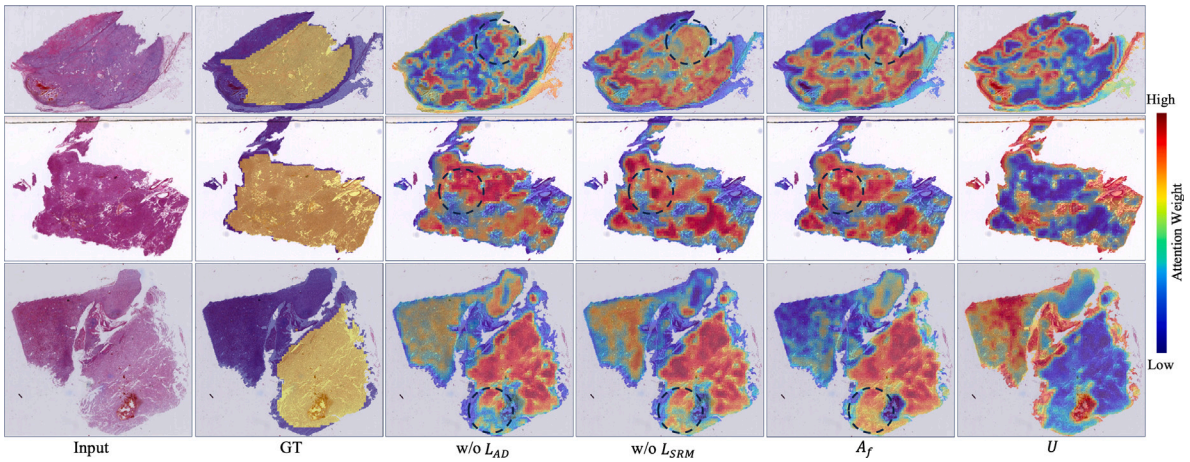


**Fig. 8.** Ablation results of attention map. Three cases are selected from the TFAH-RCC dataset to evaluate the effect of proposed loss functions on the attention maps. The first two columns are the input WSI and ground truth, respectively. The third column denotes the attention map $A_s'$ when removing the attention distillation loss $L_{AD}$. The fourth column denotes the global attention map $A_{global}$ when removing the $L_{SRM}$ loss. The fused attention map $A_f$ and uncertainty map $U$ are also visualized.

**Table 10**
Ablation experiment (presented in %) on the TFAH-RCC and TCGA-RCC datasets.

| Dataset | TFAH-RCC | | | | | TCGA-RCC | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Task | Bag-level | | | Instance-level | | Bag-level | | | Instance-level | |
| Metric | AUC | ACC | F1 | ACC | F1 | AUC | ACC | F1 | ACC | F1 |
| w/o GCN | 94.8 ± 1.4 | 83.5 ± 3.7 | 83.1 ± 3.2 | 73.5 ± 5.2 | 67.1 ± 6.2 | 98.3 ± 1.5 | 91.4 ± 3.4 | 90.7 ± 3.9 | 68.7 ± 4.6 | 63.5 ± 5.7 |
| w/o $L_{AD}$ | 94.6 ± 1.8 | 80.6 ± 2.8 | 80.2 ± 2.5 | 57.8 ± 13.3 | 47.4 ± 15.6 | 98.4 ± 1.8 | 91.5 ± 3.1 | 90.8 ± 3.7 | 59.8 ± 5.3 | 50.0 ± 9.9 |
| w/o $L_{mask}$ | 94.7 ± 1.6 | 83.8 ± 4.2 | 83.4 ± 4.1 | 74.6 ± 5.1 | 67.9 ± 5.1 | 98.3 ± 1.2 | 90.3 ± 3.3 | 90.0 ± 3.4 | 68.1 ± 4.4 | 62.6 ± 5.8 |
| w/o $L_{SRM}$ | 94.4 ± 1.4 | 81.6 ± 3.5 | 81.2 ± 3.5 | 58.4 ± 8.6 | 50.3 ± 10.0 | 98.3 ± 1.2 | 91.5 ± 4.8 | 90.8 ± 5.1 | 58.3 ± 3.0 | 48.4 ± 4.2 |
| w/o $L_{UIC}$ | 94.3 ± 1.6 | 83.1 ± 3.6 | 82.7 ± 3.6 | 73.4 ± 3.7 | 65.8 ± 3.8 | 98.3 ± 1.7 | 91.5 ± 3.9 | 90.6 ± 4.3 | 68.6 ± 4.5 | 63.0 ± 5.2 |
| **E²-MIL** | **95.2 ± 1.2** | **85.2 ± 3.7** | **84.8 ± 3.7** | **76.4 ± 1.9** | **69.4 ± 2.0** | **98.9 ± 1.3** | **92.7 ± 5.1** | **91.8 ± 5.3** | **71.1 ± 1.8** | **66.3 ± 2.2** |

**Table 11**
Experiment results (presented in %) of different magnifications on the TFAH-RCC subtyping dataset. The best results are marked in bold, and the comparable results are denoted by superscript * based on the paired t-test ($p$-value >0.05).

| Task | Bag-level | | Instance-level | |
|---|---|---|---|---|
| Metric | AUC (10×) | AUC (20×) | F1 (10×) | F1 (20×) |
| CLAM | 93.5 ± 0.6 | 93.7 ± 1.2 | 51.2 ± 8.7 | 50.9 ± 7.4 |
| GraphMIL | 93.6 ± 1.4 | 94.0 ± 1.9 | 50.9 ± 12.1 | 49.8 ± 9.8 |
| TransMIL | 91.6 ± 1.7 | 91.5 ± 2.0 | 29.6 ± 3.1 | 31.1 ± 4.6 |
| Bayes-MIL | 90.7 ± 2.3 | 91.0 ± 3.1 | 49.3 ± 6.5 | 49.0 ± 5.6 |
| **E²-MIL** | **95.2 ± 1.2** | **95.4 ± 2.2** | **69.4 ± 2.0** | **68.9 ± 1.9** |

**Table 12**
Result (presented in %) of the ABMIL and graphMIL methods on the TFAH-RCC dataset after incorporating the attention distillation loss.

| Task | Bag-level | | | Instance-level | |
|---|---|---|---|---|---|
| Metric | AUC | ACC | F1 | ACC | F1 |
| ABMIL | 87.2 ± 2.8 | 74.6 ± 4.6 | 72.6 ± 5.0 | 46.5 ± 2.5 | 38.3 ± 2.8 |
| ABMIL+$L_{AD}$ | 91.4 ± 1.6 | 77.7 ± 3.1 | 76.1 ± 3.6 | 65.3 ± 3.8 | 59.5 ± 3.6 |
| △ | **4.2** ↑ | **3.1** ↑ | **3.5** ↑ | **18.8** ↑ | **21.2** ↑ |
| GraphMIL | 91.6 ± 1.7 | 79.6 ± 3.5 | 79.4 ± 4.9 | 59.4 ± 10.4 | 50.9 ± 12.1 |
| GraphMIL+$L_{AD}$ | 95.5 ± 1.2 | 84.0 ± 2.5 | 83.5 ± 2.3 | 64.1 ± 3.6 | 56.6 ± 3.5 |
| △ | **3.9** ↑ | **4.4** ↑ | **4.1** ↑ | **4.7** ↑ | **5.7** ↑ |

We also evaluate the effect of the proposed module on the generated attention maps. As depicted in Fig. 8, after removing the attention distillation loss in ADM, many details are ignored in the attention map. This further illustrates that the proposed attention distillation loss $L_{AD}$ can help the global network capture more detail-aware attention knowledge. After removing the structure-aware attention refined loss $L_{SRM}$, the attention map can learn enough details by the $L_{AD}$, but the whole tumor structures are not modeled well. Overall, the fused attention map $A_f$ demonstrates that almost all the critical positive instances are captured well, and the uncertainty map $U$ highlights all the suspicious regions, further increasing the framework's reliability.

#### 4.7.2. Effect of different magnifications

To evaluate the effect of magnification on both bag-level and instance-level classification tasks, we conduct experiments using a 20× magnification. The specific experiment results are reported in Table 11. For our E²-MIL framework and all selected baselines, the performance of all models exhibits negligible variance under different resolution conditions (*i.e.*, 10× and 20×). Specifically, E²-MIL shows only slight performance variations, with 0.2% in bag-level and 0.5% in instance-level classification tasks. Notably, with the introduction of DAM, SRM, and UIC within E²-MIL, our framework consistently outperforms the current best existing methods at both magnifications.

#### 4.7.3. Effect of attention distillation loss

To evaluate the proposed attention distillation loss $L_{AD}$, which can help improve both bag-level and instance-level performances, we introduce it into other MIL-based methods (*i.e.*, ABMIL Ilse et al., 2018 and GraphMIL Chen et al., 2021). The number of randomly split sub-bags in each method is set to five. As reported in Table 12, after introducing our proposed attention distillation loss, the bag-level and instance-level performances have increased across all three metrics. Specifically, for ABMIL, the instance-level performances increase by 18.8% and 21.2% on the metric of ACC and F1, respectively. This further illustrates the effectiveness of our proposed attention distillation loss, which simply utilizes the local complementary sub-bags to help the global network learn more fine-grained attention knowledge.

#### 4.7.4. Effect of the instance-level label

Acquiring a substantial number of pixel-level instance labels is very time-consuming and labor-intensive. Therefore, we explore the

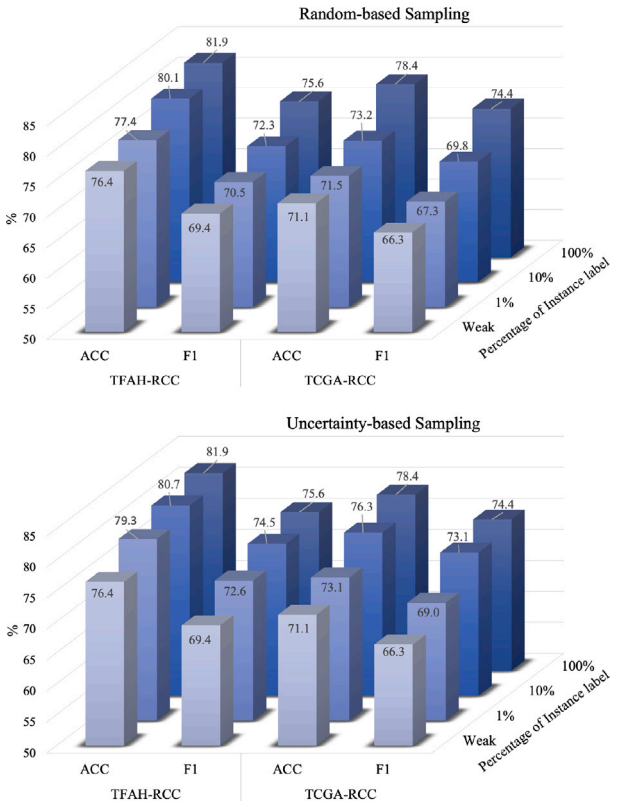


**Fig. 9.** Effect of the patch-level label. The top figure shows the instance-level classification performance by randomly selecting different ratios of patch-level labels; The bottom figure shows the instance-level classification performance by introducing the patch labels with high uncertainty values based on the uncertainty map $U$.

effectiveness of the uncertainty map in guiding the annotation process by comparing two sampling methods (*i.e.*, random sampling and uncertainty-based sampling) on the TFAH-RCC and TCGA-RCC datasets. For the random sampling method, we randomly select different ratios of slides[2] and introduce their patch labels in the training process. For the uncertainty-based sampling method, we calculate each slide's uncertainty value by averaging its uncertainty map $U$. Then, the different ratios of slides with the highest slide uncertainty values are selected to introduce their patch label in the training process. As shown in Fig. 9, after introducing different percentages of instance labels, the instance-level performance has been increased to varying degrees. The uncertainty-based sampling method is more effective in selecting the slides that need to be annotated and significantly reduces the workload of annotation. Specifically, with only one-tenth of instance labels, the model's instance-level classification performance is comparable to the fully annotated scenario in the TFAH-RCC dataset.

#### 4.7.5. Effect of reject rate

To evaluate the effect of the reject rate on the model's performance, we discard different ratios of slides based on the uncertainty map $U$ and calculate the accuracy of the remaining data. Specifically, the uncertainty value of each slide is calculated by averaging the uncertainty map $U$ of all instances. Then, based on the slide uncertainty value, different ratios of slides with high uncertainty are rejected. As shown in Fig. 10, with the increase of reject rate, the performances of most uncertainty estimation methods at the bag level and instance level are

[2] Note that it is rational to introduce all the patch labels in a slide together because pathologists usually annotate all tumor regions in a slide when they annotate that slide.

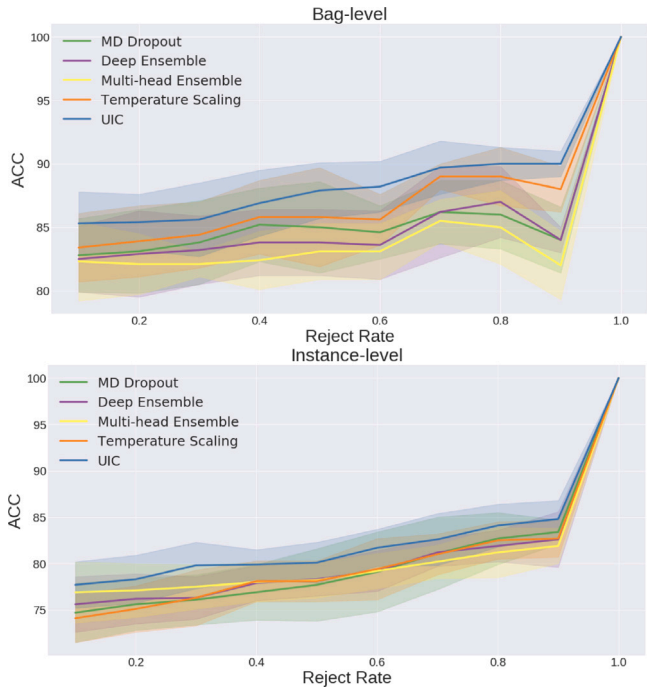


**Fig. 10.** Accuracy after rejecting slides with the unreliable prediction (*i.e.*, high estimated uncertainty). The top and bottom figures are the bag-level and instance-level performance in the TFAH-RCC dataset, respectively. The x-axis is the proportion of rejected slides over the dataset. The y-axis is the corresponding accuracy on the remaining slides.

improved. Our proposed UIC achieves the best performance under all reject rates, indicating its superiority in enhancing the reliability of the model by discarding the most unreliable slides.

#### 4.7.6. Effect of hyper-parameters

To evaluate the effect of hyper-parameters on the instance-level classification performance, we conduct a series of ablation experiments on the TFAH-RCC dataset. The results are summarized in Fig. 11. Specifically, the best value for the number of sub-bags $K$ is five. When the number of sub-bags is too small, the global network cannot learn enough detailed attention knowledge. For the masked bag threshold $m$, the best threshold is 0.5. When the threshold is too low, some normal instances are also masked, causing the model to locate some false-positive instances. When the threshold is too high, only part of the positive instances are masked, affecting the localization ability on critical instances. For the threshold $t_f$ to generate the pseudo-label $Y_{pse}$, the best threshold is 0.7. Higher or lower thresholds may introduce more noise in the generated patch pseudo-labels. For the threshold $t_u$ to filter the noise in the pseudo-label $Y_{pse}$, the best threshold is 0.8. Lower thresholds cannot effectively filter out the noises, while higher thresholds may remove some instances with high-quality patch labels.

## 5. Discussion

Although E$^2$-MIL can obtain accurate prediction results and robust instance-level uncertainty estimation, it still has several limitations. Firstly, in this work, we assume each slide contains only one type of cancer. Under this assumption, we can assign the slide-level classification results to the patches. Therefore, it is sufficient to perform binary patch classification (*i.e.*, cancer vs. non-cancer). However, if a slide contains multiple cancer subtypes (*e.g.*, a WSI of stomach cancer may contain multiple subtypes, such as papillary, tubular, and mucinous), the slide-level classification task should be reformulated as a multi-label classification problem (instead of a multi-class classification problem).

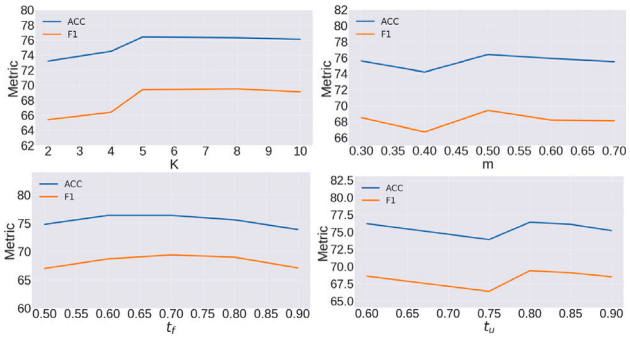


**Fig. 11.** Impact of four important hyper-parameters: the sub-bag numbers $K$, the threshold for building the masked bag $m$, the threshold $t_f$ for generating the pseudo-label $Y_{pse}$, and the threshold $t_u$ to determine whether the pseudo-label $Y_{pse}$ is noisy.

And then, the binary instance label cannot differentiate multiple subtypes. To adapt our framework to complex multi-label scenarios, we plan to incorporate multiple class tokens into the structure-aware attention refined module. Each class token will generate a class-aware attention map corresponding to a specific subtype. Moreover, we will expand the uncertainty-aware classifier to handle multi-class classification tasks by adjusting the number of classes in the classifier head. Secondly, although E$^2$-MIL has significantly improved the instance-level classification performance compared with existing methods, there is still much room for improvement in instance-level performance compared with the bag-level performance. We hope that our proposed large multi-center multi-cancer subtyping datasets, which include fine-grained pixel-level annotation, will serve as a catalyst for further research in this area. In the future, we will explore the potential of leveraging popular large language models (LLMs) and vision-language models (VLMs) to further enhance our model's capabilities in locating the positive instances, which will be achieved by integrating additional external clinical knowledge to provide effective guidance. Thirdly, the detail-aware attention distillation module just needs to be utilized in the training process. Given the partitioning of bags in DAM is random during each forward pass, the local network can continuously identify and focus on the hard-to-classified instances, thereby enhancing the localization ability of the attention map. This attention map is then utilized to generate instance pseudo-labels, providing strong guidance for supervising the training of the uncertainty-aware classifier.

## 6. Conclusion

In this work, we proposed an explainable and evidential multiple instance learning (E$^2$-MIL) framework for whole slide image analysis. The E$^2$-MIL improves the clinical usability of the model from the aspects of interpretability and reliability. Specifically, the detail-aware attention distillation module generates a fine-grained attention map by utilizing the complementary sub-bags to help the global model learn more attention knowledge from the local network. The structure-aware attention refined module generates a structure-aware attention map by utilizing the clustering bags to capture the whole structure of tumor region. These two attention maps are fused and significantly improve the model's interpretability by locating all the positive instances well. Moreover, the uncertainty-aware instance classifier improves the model's reliability by providing robust instance-level predictive uncertainty estimation based on the subjective logic theory. Extensive comparative and ablation experiments on three large multi-center subtyping datasets showed that E$^2$-MIL achieved state-of-the-art results for whole slide image analysis. In the future, we will extend our framework to more cancer types and diagnostic tasks.

## CRediT authorship contribution statement

**Jiangbo Shi:** Writing – review & editing, Writing – original draft, Methodology, Data curation, Conceptualization. **Chen Li:** Writing – review & editing, Writing – original draft, Supervision, Project administration, Funding acquisition, Data curation, Conceptualization. **Tieliang Gong:** Writing – review & editing, Methodology, Data curation, Conceptualization. **Huazhu Fu:** Writing – review & editing, Writing – original draft, Supervision, Methodology, Funding acquisition, Conceptualization.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Data availability

Data will be made available on request.

## Acknowledgments

This work has been supported by the Key Research and Development Program of Ningxia Hui Nationality Autonomous Region under Grant 2022BEG02025 and 2023BEG02023, in part by the Project of China Knowledge Center for Engineering Science and Technology. This research is supported by the National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No: AISG2-TC-2021-003), and Agency for Science, Technology and Research (A*STAR) Central Research Fund ("Robust and Trustworthy AI system for Multi-modality Healthcare").

## References

Bianchi, F.M., Grattarola, D., Alippi, C., 2020. Spectral clustering with graph neural networks for graph pooling. In: Proceedings of the International Conference on Machine Learning. PMLR, pp. 874–883.

Brixtel, R., Bougleux, S., Lézoray, O., Caillot, Y., Lemoine, B., Fontaine, M., Nebati, D., Renouf, A., 2022. Whole slide image quality in digital pathology: Review and perspectives. IEEE Access 10, 131005–131035.

Chai, L.R., 2018. Uncertainty Estimation in Bayesian Neural Networks and Links to Interpretability (Master's Thesis). Massachusetts Institute of Technology.

Chen, R.J., Lu, M.Y., Shaban, M., Chen, C., Chen, T.Y., Williamson, D.F., Mahmood, F., 2021. Whole slide images are 2D point clouds: Context-aware survival prediction using patch-based graph convolutional networks. In: Proceedings of the International Conference on Medical Image Computing and Computer Assisted Intervention. Springer, pp. 339–349.

Chikontwe, P., Sung, H.J., Jeong, J., Kim, M., Go, H., Nam, S.J., Park, S.H., 2022. Weakly supervised segmentation on neural compressed histopathology with self-equivariant regularization. Med. Image Anal. 80, 102482.

Cui, Y., Liu, Z., Liu, X., Liu, X., Wang, C., Kuo, T.-W., Xue, C.J., Chan, A.B., 2023. Bayes-MIL: A new probabilistic perspective on attention-based multiple instance learning for whole slide images. In: International Conference on Learning Representations.

Cui, M., Zhang, D.Y., 2021. Artificial intelligence and computational pathology. Lab. Investig. 101 (4), 412–422.

Di, D., Zhang, J., Lei, F., Tian, Q., Gao, Y., 2022. Big-hypergraph factorization neural network for survival prediction from whole slide image. IEEE Trans. Image Process. 31, 1149–1160.

Dolezal, J.M., Srisuwananukorn, A., Karpeyev, D., Ramesh, S., Kochanny, S., Cody, B., Mansfield, A.S., Rakshit, S., Bansal, R., Bois, M.C., et al., 2022. Uncertainty-informed deep learning models enable high-confidence predictions for digital histopathology. Nature Commun. 13 (1), 6572.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N., 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In: Proceedings of the International Conference on Learning Representations.

Fey, M., Lenssen, J.E., 2019. Fast graph representation learning with PyTorch geometric. In: Proceedings of the International Conference on Learning Representations.

Gal, Y., Ghahramani, Z., 2016. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In: Proceedings of the International Conference on Machine Learning. PMLR, pp. 1050–1059.

Godson, L., Alemi, N., Nsengimana, J., Cook, G.P., Clarke, E.L., Treanor, D., Bishop, D.T., Newton-Bishop, J., Gooya, A., Magee, D., 2024. Immune subtyping of melanoma whole slide images using multiple instance learning. Med. Image Anal. 103097.

Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q., 2017. On calibration of modern neural networks. In: Proceedings of the International Conference on Machine Learning. PMLR, pp. 1321–1330.

Han, Z., Zhang, C., Fu, H., Zhou, J.T., 2022. Trusted multi-view classification with dynamic evidential fusion. IEEE Trans. Pattern Anal. Mach. Intell. 45 (2), 2551–2566.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778.

Ilse, M., Tomczak, J., Welling, M., 2018. Attention-based deep multiple instance learning. In: Proceedings of the International Conference on Machine Learning. PMLR, pp. 2127–2136.

Jiang, H., Zhang, R., Zhou, Y., Wang, Y., Chen, H., 2023. DoNet: Deep de-overlapping network for cytology instance segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15641–15650.

Kingma, D.P., Ba, J., 2015. Adam: A method for stochastic optimization. In: Proceedings of the International Conference on Learning Representations.

Lakshminarayanan, B., Pritzel, A., Blundell, C., 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. In: Proceedings of Advances in Neural Information Processing Systems, vol. 30, pp. 6402–6413.

Laleh, N.G., Muti, H.S., Loeffler, C.M.L., Echle, A., Saldanha, O.L., Mahmood, F., Lu, M.Y., Trautwein, C., Langer, R., Dislich, B., et al., 2022. Benchmarking weakly-supervised deep learning pipelines for whole slide classification in computational pathology. Med. Image Anal. 79, 102474.

Li, B., Li, Y., Eliceiri, K.W., 2021. 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. pp. 14318–14328.

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

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

Linmans, J., Elfwing, S., van der Laak, J., Litjens, G., 2023. Predictive uncertainty estimation for out-of-distribution detection in digital pathology. Med. Image Anal. 83, 102655.

Liu, P., Ji, L., Ye, F., Fu, B., 2024. Advmil: Adversarial multiple instance learning for the survival analysis on whole-slide images. Med. Image Anal. 91, 103020.

Lu, M.Y., Williamson, D.F., Chen, T.Y., Chen, R.J., Barbieri, M., Mahmood, F., 2021. Data-efficient and weakly supervised computational pathology on whole-slide images. Nat. Biomed. Eng. 5 (6), 555–570.

Mehrtens, H.A., Kurz, A., Bucher, T.-C., Brinker, T.J., 2023. Benchmarking common uncertainty estimation methods with histopathological images under domain shift and label noise. arXiv preprint arXiv:2301.01054.

Pocevičiūtė, M., Eilertsen, G., Jarkman, S., Lundström, C., 2022. Generalisation effects of predictive uncertainty estimation in deep learning for digital pathology. Sci. Rep. 12 (1), 8329.

Pocevičiūtė, M., Eilertsen, G., Lundström, C., 2020. Survey of XAI in digital pathology. In: Artificial Intelligence and Machine Learning for Digital Pathology. Springer, pp. 56–88.

Qu, L., Wang, M., Song, Z., Luo, X., 2022. Bi-directional weakly supervised knowledge distillation for whole slide image classification. In: Proceedings of the Advances in Neural Information Processing Systems, vol. 35, pp. 15368–15381.

Schmidt, A., Morales-Álvarez, P., Molina, R., 2023. Probabilistic attention based on Gaussian processes for deep multiple instance learning. IEEE Trans. Neural Netw. Learn. Syst..

Sensoy, M., Kaplan, L., Kandemir, M., 2018. Evidential deep learning to quantify classification uncertainty. In: Proceedings of the Advances in Neural Information Processing Systems, vol. 31, pp. 3183–3193.

Shao, Z., Bian, H., Chen, Y., Wang, Y., Zhang, J., Ji, X., Zhang, Y., 2021. TransMIL: Transformer based correlated multiple instance learning for whole slide image classification. In: Proceedings of the Advances in Neural Information Processing Systems, vol. 34, pp. 2136–2147.

Shao, Z., Chen, Y., Bian, H., Zhang, J., Liu, G., Zhang, Y., 2023. HVTSurv: Hierarchical vision transformer for patient-level survival prediction from whole slide image. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 37, (no. 2), pp. 2209–2217.

Shi, J., Li, C., Gong, T., Zheng, Y., Fu, H., 2024. 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. pp. 11248–11258.

Shi, J., Tang, L., Gao, Z., Li, Y., Wang, C., Gong, T., Li, C., Fu, H., 2023a. MG-Trans: Multi-scale graph transformer with information bottleneck for whole slide image classification. IEEE Trans. Med. Imaging.

Shi, J., Tang, L., Li, Y., Zhang, X., Gao, Z., Zheng, Y., Wang, C., Gong, T., Li, C., 2023b. A structure-aware hierarchical graph-based multiple instance learning framework for pt staging in histopathological image. IEEE Trans. Med. Imaging 42 (10), 3000–3011.

Stacke, K., Eilertsen, G., Unger, J., Lundström, C., 2020. Measuring domain shift for deep learning in histopathology. IEEE J. Biomed. Health Inf. 25 (2), 325–336.

Tang, W., Huang, S., Zhang, X., Zhou, F., Zhang, Y., Liu, B., 2023. 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. pp. 4078–4087.

Van Amersfoort, J., Smith, L., Teh, Y.W., Gal, Y., 2020. Uncertainty estimation using a single deep deterministic neural network. In: Proceedings of the International Conference on Machine Learning. PMLR, pp. 9690–9700.

Wang, M., Lin, T., Wang, L., Lin, A., Zou, K., Xu, X., Zhou, Y., Peng, Y., Meng, Q., Qian, Y., et al., 2023. Uncertainty-inspired open set learning for retinal anomaly identification. Nature Commun. 14 (1), 6757.

Wang, X., Yang, S., Zhang, J., Wang, M., Zhang, J., Yang, W., Huang, J., Han, X., 2022. Transformer-based unsupervised contrastive learning for histopathological image classification. Med. Image Anal. 81, 102559.

Xiang, H., Shen, J., Yan, Q., Xu, M., Shi, X., Zhu, X., 2023. Multi-scale representation attention based deep multiple instance learning for gigapixel whole slide image analysis. Med. Image Anal. 89, 102890.

Xie, Y., Zhang, J., Liao, Z., Verjans, J., Shen, C., Xia, Y., 2021. Intra- and inter-pair consistency for semi-supervised gland segmentation. IEEE Trans. Image Process. 31, 894–905.

Zhang, H., Burrows, L., Meng, Y., Sculthorpe, D., Mukherjee, A., Coupland, S.E., Chen, K., Zheng, Y., 2023. Weakly supervised segmentation with point annotations for histopathology images via contrast-based variational model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15630–15640.

Zhang, H., Meng, Y., Zhao, Y., Qiao, Y., Yang, X., Coupland, S.E., Zheng, Y., 2022. DTFD-MIL: Double-tier feature distillation multiple instance learning for histopathology whole slide image classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18802–18812.

Zhao, B., Deng, W., Li, Z.H.H., Zhou, C., Gao, Z., Wang, G., Li, X., 2024. LESS: Label-efficient multi-scale learning for cytological whole slide image screening. Med. Image Anal. 103109.

Zheng, Y., Gindra, R.H., Green, E.J., Burks, E.J., Betke, M., Beane, J.E., Kolachalama, V.B., 2022. A graph-transformer for whole slide image classification. IEEE Trans. Med. Imaging 41 (11), 3003–3015.

Zou, K., Yuan, X., Shen, X., Wang, M., Fu, H., 2022. TbraTS: Trusted brain tumor segmentation. In: International Conference on Medical Image Computing and Computer Assisted Intervention. Springer, pp. 503–513.