Medical Image Analysis 89 (2023) 102890

Contents lists available at ScienceDirect

Medical Image Analysis

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

Multi-scale representation attention based deep multiple instance learning for gigapixel whole slide image analysis

Hangchen Xiang [a,1], Junyi Shen [b,1], Qingguo Yan [c], Meilian Xu [d,*], Xiaoshuang Shi [a,*], Xiaofeng Zhu [a]

[a] School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China
[b] Division of Liver Surgery, Department of General Surgery, West China Hospital, Sichuan University, Chengdu, 610044, China
[c] Department of Pathology Key Laboratory of Resource Biology and Biotechnology in Western China, Ministry of Education, School of Medicine, Northwest University, 229 Taibai North Road, Xi'an 710069, China
[d] School of Electronic Information and Artificial Intelligence, Leshan Normal University, Leshan, 614000, China

ARTICLE INFO

Keywords:
Whole slide images
Weakly supervised
Convolutional neural network
Multi-scale representation attention

ABSTRACT

Recently, convolutional neural networks (CNNs) directly using whole slide images (WSIs) for tumor diagnosis and analysis have attracted considerable attention, because they only utilize the slide-level label for model training without any additional annotations. However, it is still a challenging task to directly handle gigapixel WSIs, due to the billions of pixels and intra-variations in each WSI. To overcome this problem, in this paper, we propose a novel end-to-end interpretable deep MIL framework for WSI analysis, by using a two-branch deep neural network and a multi-scale representation attention mechanism to directly extract features from all patches of each WSI. Specifically, we first divide each WSI into bag-, patch- and cell-level images, and then assign the slide-level label to its corresponding bag-level images, so that WSI classification becomes a MIL problem. Additionally, we design a novel multi-scale representation attention mechanism, and embed it into a two-branch deep network to simultaneously mine the bag with a correct label, the significant patches and their cell-level information. Extensive experiments demonstrate the superior performance of the proposed framework over recent state-of-the-art methods, in term of classification accuracy and model interpretability. *All source codes are released at*: https://github.com/xhangchen/MRAN/.

1. Introduction

Pathological examination is viewed as the "gold standard" for clinical diagnosis of tumors (Elmore, 2021). With the development of whole slide imaging techniques, pathologists can manually examine the tumors through whole slide images (WSIs). However, each whole slide image (WSI) usually consists of billions of pixels, e.g., $50{,}000 \times 50{,}000$, so that manual examination is laborious, time-consuming, expensive and even error-prone. To reduce the workloads of pathologists and boost the diagnosis accuracy, computer-aided diagnosis (CAD) systems have developed by using computer vision and machine learning techniques (Sirinukunwattana et al., 2017; Chen et al., 2019b; Keikhosravi et al., 2020; Peng et al., 2022). Recently, convolutional neural networks (CNNs) have been widely used in CAD systems and achieved great success on pathological images (Litjens et al., 2017; Li et al., 2021a), because they can automatically extract powerful features. Unfortunately, they often require a large amount of high quality annotated data to obtain satisfactory performance, and they cannot directly handle the input with a large size due to the limited computer memory, i.e., their input size is usually with hundreds multiply hundreds pixels, e.g., 224×224. Thus, many patch-based deep methods (Xu et al., 2017; Araújo et al., 2017; Zhang et al., 2019), which crop each WSI into hundreds or even thousands of patches and then annotate them for model training, have been developed for WSI analysis. Obviously, these methods also require pathologists to annotate a large number of patches, leading to expensive annotation costs. To further reduce pathologists' workloads, WSIs-based deep methods (Cruz-Roa et al., 2014; Hashimoto et al., 2020; Chikontwe et al., 2020), which adopt CNNs to directly tackle with WSIs by only using slide-level labels for tumor diagnosis analysis, have attracted increasing attention.

Based on whether assigning labels to patches in each WSI, existing WSIs-based deep methods can be roughly classified into two categories: (1) patches with pseudo-labels (Hou et al., 2016; Cheng et al.,

* Corresponding authors.
E-mail addresses: xu.meilian05@gmail.com (M. Xu), xsshi2013@gmail.com (X. Shi).
[1] Co-first authors: They have equal contribution.

https://doi.org/10.1016/j.media.2023.102890
Received 27 December 2022; Received in revised form 22 April 2023; Accepted 3 July 2023
Available online 8 July 2023

2020; Wang et al., 2021); (2) multiple instance learning (MIL) (Zheng et al., 2018; Campanella et al., 2019). The methods using patches with pseudo-labels first divide each WSI into a set of patches, and then assign the WSI's label to its patches so as to iteratively train the convolutional neural network (CNN) for automatically selecting the confident patches consistent with its slide-level label, and finally employ a decision fusion model to fuse the prediction results of patches to obtain the predicted category of the WSI. Unfortunately, each WSI is often composed of a large part of trivial patches, especially for positive WSIs that often contain only a small part of positive patches, so that CNNs usually utilize a large number of patches with wrong labels for model training, thereby possibly deteriorating the model performance due to their powerful memory capability (Lu et al., 2019; Dehaene et al., 2020). MIL only provides a general statement of the category for multiple instances (Maron and Lozano-Pérez, 1997; Dietterich et al., 1997), i.e., one bag contains tens or hundreds instances with only the bag label known. Additionally, even if a bag has only one positive instance, its category is positive. Thus, MIL is very suitable for handling WSIs.

Existing deep MIL based methods for WSI analysis view each WSI as a bag and crop it into a set of patches as instances, and then directly utilize the bag label to train the model (Xu et al., 2014; Kandemir and Hamprecht, 2015). These methods have two popular strategies. One of them is to first employ a pre-trained model on ImageNet (Russakovsky et al., 2015) for obtaining feature representations of patches, and then leverage these feature representations, the slide-level label, and a network with multiple fully connected layers to train the model (Gao et al., 2016; Campanella et al., 2019). However, the intrinsic distribution between different data types is often different, and thus this strategy cannot obtain the optimal feature representations of patches. Another intuitive strategy is to resize each patch into a lower resolution one (Liu et al., 2020) or maximumly reduce the number of patches to decrease the memory costs (Chikontwe et al., 2020), so that CNNs can directly analyze each WSI. As a result, this strategy might lose some significant information of each WSI, thereby restricting the model performance. Additionally, the other popular strategy for extracting features of WSIs is to employ unsupervised methods, e.g., contrastive learning (Li et al., 2021b; Mo et al., 2023), domain adaptation (Ren et al., 2019), clustering (Chen et al., 2022). However, these methods make extracted features easily be prone to bias and noise, so that their generalization performance is unstable on data sets with different distributions (Dike et al., 2018). Therefore, it is very necessary and challenging to directly handle the gigapixel WSIs.

In clinical diagnosis, pathologists not only require the CAD system to provide diagnosis results but also supply the evidence to support and interpret the results (Zhang et al., 2019; Pirovano et al., 2020). Hence, to better interpret and locate the diseased region in WSIs, some interpretable deep MIL methods have been proposed by using different attention mechanisms. These methods majorly adopt two strategies: (1) posterior attention mechanism; (2) prior attention mechanism. The posterior attention mechanism is that first employs a pre-trained model or unsupervised learning model to extract the features of patches, and then utilizes feature visualization technologies, e.g., class activation map (CAM) (Zhou et al., 2016; Selvaraju et al., 2017), to obtain the discriminative patches (Bae et al., 2020; Zhang et al., 2022). The prior attention mechanism is that directly embeds the attention mechanism into the deep neural network to mine the significant patches during model training for WSI analysis (Li et al., 2019; Lu et al., 2021). Compared to the methods using the posterior attention mechanism, which severely relies on the pre-trained or unsupervised learning model, the methods using the prior attention mechanism can directly mine and interpret the significant patches in each WSI and often obtain better interpretation results.

Motivated by aforementioned observations, in this paper, we propose a novel end-to-end interpretable weakly supervised deep framework, namely *multi-scale representation attention based network* (MRAN), to directly handle gigapixel WSIs (see Fig. 1). The proposed framework employs the CNN to automatically extract powerful features from patches, leverages MIL for WSI analysis, and embeds multi-scale representation attention into the network to boost model interpretability. Specifically, in order to reduce the effect of noisy labels on model performance, we first divide each WSI into a set of bags, each of which contains tens of patches, and assign the slide-level label to its bags, thereby decreasing the number of wrong labels. Additionally, because the bags contain noisy labels, each bag might consist of a large apart of trivial patches, and cell-level information in each patch is very significant for disease diagnosis (Shi et al., 2017; Hu et al., 2018), we propose a multi-scale representation attention mechanism to simultaneously discover the bag with a correct label, interpret the significance of patches, and mine the significant cell-level information. Finally, extensive experiments on four publicly available WSI datasets demonstrate the effectiveness of the proposed framework on WSI classification and interpretation.

In summary, our major contributions are listed as follows:

- We propose a novel end-to-end interpretable deep MIL framework, namely MRAN, which can directly extract features from all patches of gigapixel WSIs by only using slide-level labels during model training. In other words, our framework can directly handle any size of each WSI, with extracting its all patch-level features for model training.
- We propose a multi-scale representation attention mechanism to simultaneously mine the significant bag-, patch- and cell-level images for WSI analysis.
- Extensive experimental results illustrate that the proposed framework outperforms recent state-of-the-art methods on WSI classification, with better interpretation capability.

The rest of this paper is organized as follows. Section 2 briefly reviews some related popular weak supervision methods. Section 3 introduces the proposed framework. Section 4 shows and analyzes experimental results. Finally, Section 5 concludes this paper and points out the future work.

2. Related work

In this section, we briefly review some popular weakly supervised methods for pathology image analysis, including pseudo-label patch-based methods, MIL-based methods, and attention-based deep MIL methods.

Pseudo-label patch-based methods: which first utilize a certain method to generate pseudo-labels for unlabeled patches, and then employs these patches with their pseudo-labels to train the model in a supervised learning manner. Generally, the training process iteratively conducts the pseudo-label generation and model training until convergence. Because pseudo-label generation is the key stage during the training process, many different strategies have been proposed to generate powerful pseudo-labels. For instances, the expectation–maximization (EM) based strategy (Hou et al., 2016), which automatically locates discriminative patches robustly by utilizing the spatial relationships of patches in each WSI; the cluster-then-label strategy (Peikari et al., 2018; Xu et al., 2022), which utilizes clustering analysis to identify high-density regions in the data space for annotating the unlabeled data; the consistency regularization strategy (Shi et al., 2020a; Pulido et al., 2020; Marini et al., 2021), which adopts the consistency regularization between augmented samples of different epochs to generate strong pseudo-labels; the graph-based strategy (Shin et al., 2022), which considers local and global relationships of patches to refine the initial pseudo-labels. Although these methods generating pseudo-labels for patches can effectively leverage the patch-level information of WSIs, deep neural networks are easily overfitting on the wrong labels, thereby deteriorating the model performance (Zhang et al., 2021).

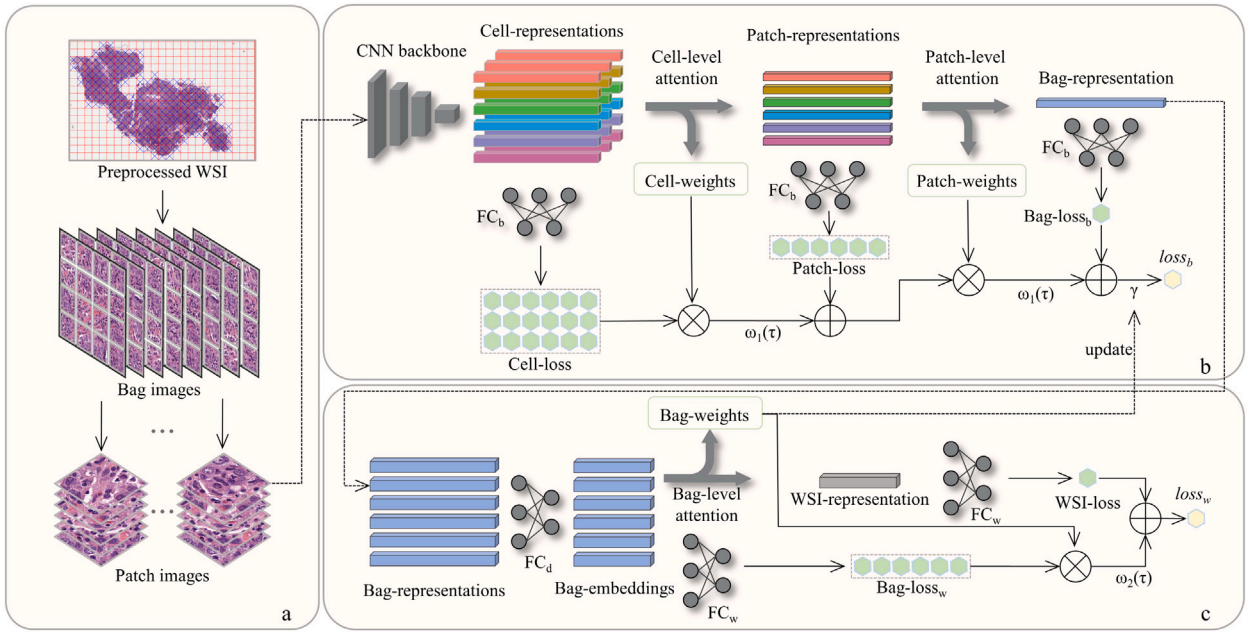


Fig. 1. The structure of the proposed MRAN. **(a)** WSI preprocessing: cropping each WSI into bag- and patch-level images; **(b)** Bag-level image classification: feeding patch-level images of each bag into the first branch to obtain Cell-, Patch- and Bag-representations, and their losses that constitute of the loss_b; **(c)** WSI classification: feeding Bag-representations of each WSI into the second branch to produce Bag-embeddings and WSI-representation, and their losses that make up the loss_w.

MIL-based methods: which view each WSI as a bag consisting of multiple instances, where each instance is one patch cropped from the WSI. In MIL, when the bag (WSI) contains at least one positive instance (patch), then the bag (WSI) is positive. Based on this characteristic, MIL is very suitable for the WSI scenario, and thus there are many MIL-based methods have been proposed for WSI analysis. E.g., a MIL-based deep learning system only using the reported diagnoses as labels has been proposed for WSIs (Campanella et al., 2019). To better leverage the useful information in WSIs, the rectified cross-entropy loss and an upper transition loss have been designed for WSI analysis (Chen et al., 2019a). To jointly learn instance- and bag-level features and reduce intra-class variations, an end-to-end MIL-based deep framework using a center loss has been developed (Chikontwe et al., 2020). In order to integrate the embedding extractor and aggregator together for training, EPL (Xie et al., 2020) utilizes clustering to divide WSI into several groups and sample each group, and employs part learning to perform end-to-end training. Most of these methods fail to take into account the significance of instances, thereby possibly restricting the model classification accuracy and interpretability (Shi et al., 2020b).

Attention-based deep MIL methods: To interpret the significance of instances and meanwhile attain desired bag classification accuracy, attention-based deep MIL methods have been proposed to simultaneously utilize deep neural networks to extract the instance features and employ the attention mechanism to interpret their significance. Specifically, C2C (Sharma et al., 2021) clusters and samples the patches feature obtained through feature extractor, and then fuses their features by using an adaptive attention mechanism for end-to-end training. The method (Ilse et al., 2018) designs two attention mechanisms and individually embeds them to learn the instance weights. TransMIL (Shao et al., 2021) adopts the pre-trained model to obtain the feature representations, and employs the self-attention to calculate the relations among patches so as to explore both morphological and spatial information of WSIs. In addition to the self-attention, a dual-stream architecture with self-supervised contrastive learning has been proposed to better extract the feature representation of patches in an unsupervised manner (Li et al., 2021b). Instead of using the attention mechanism to interpret the significance of patches, dynamic pooling has been applied into the deep MIL framework for selecting key patches (Yan et al., 2018). Because previous methods do not consider the relationship of bags, graph-based MIL with the attention mechanism has been developed to better explore the relations among bags (Tu et al., 2019). Double-tier feature distillation multiple instance learning (DTFD-MIL) (Zhang et al., 2022) generates bags with pseudo-labels by Grad-CAM (Selvaraju et al., 2017), and then utilizes a double-tier MIL framework to mine significant bags and their patches. Although these methods can obtain promising performance on many types of WSIs, they usually utilize the pre-trained model or the unsupervised learning manner to learn patch-level features. Hence, they might fail to extract optimal patch-level features for WSIs.

Different from previous pseudo-label patch-based methods, the proposed framework generates the pseudo-labels for bags in each WSI to largely reduce the number of wrong labels. Compared to previous MIL-based methods, instead of using the pre-trained model or the unsupervised learning manner to extract patch-level features in MIL-based and attention-based deep MIL frameworks, the proposed framework directly feeds the patch-level images into the model so as to better leverage the patch-level information of WSIs. Additionally, our framework can directly extract all patch-level features from each WSI, without using lower resolution WSIs or sampling part of patches. Moreover, previous methods fail to consider the significance of cell-level features, which are very crucial in pathological images, but the proposed framework can interpret the cell-level images and meanwhile remove the trivial ones.

3. Method

3.1. WSIs preprocessing

Given N WSIs $\{\mathbf{X}_n\}_{n=1}^{N}$, where $\mathbf{X}_n \in \mathbb{R}^{C\times H\times W}$, C, H and W denote the number of channels, image height and width, we resize the size of WSIs to $\frac{H}{2}\times\frac{W}{2}$ in order to reduce the computational and memory costs in model training. We first apply a default global binary thresholding algorithm (Malathy et al., 2016) in OpenCV to localize the tissue regions in each WSI $\mathbf{X}_n$, and then divide the region into a set of non-overlapping bag-level images $\{\mathbf{X}_{ni}\}_{i=1}^{M_b}$, where $\mathbf{X}_{ni} \in \mathbb{R}^{C\times H_b\times H_b}$, H_b denotes its height and weight, and $M_b < M = \frac{HW}{H_b^2}$.

Additionally, we divide each bag-level image into multiple patch-level images $\{\mathbf{X}_{nij}\}_{j=1}^{M_p}$, where $\mathbf{X}_{nij} \in \mathbb{R}^{C\times H_p\times H_p}$, H_p is its height and weight, and $M_p = \frac{H_b^2}{H_p^2}$. Moreover, to mine the significant cell-level information, we view each patch-level image consisting of a set of cell-level images $\{\mathbf{X}_{nijk}\}_{k=1}^{M_c}$, where $\mathbf{X}_{nijk} \in \mathbb{R}^{C\times H_c\times H_c}$, H_c is its height and weight, and $M_c = \frac{H_p^2}{H_c^2}$. Although each WSI has a different size in practice, in the proposed framework we empirically set fixed values for H_b, H_p and H_c, respectively, e.g., $H_b = 1024$, $H_p = 128$ and $H_c = 32$. As a result, each bag contains 64 patch-level images and each patch consists of 16 cell-level images.

3.2. Multi-scale representation attention based network

Let $f_\theta(\cdot)$ represent a network consisting of two branches, where θ denotes the network parameters, the first branch contains $L-1$ convolutional layers with the parameters $\{\theta_l\}_{l=1}^{L-1}$ and one fully connected layer with the parameters θ^L for bag classification, and the second branch is composed of two fully connected layers with the parameters $\{\theta_l\}_{l=L+1}^{L+2}$ for WSI classification. In this paper, we adopt a very popular network, ResNet18 (He et al., 2016), as the backbone. By feeding the patch-level images of each bag into the network, it can obtain the features of the bag-level, patch-level and cell-level images. Note that we do not directly divide each patch-level image into multiple cell-level images, whose features are obtained based on the feature map of its corresponding patch-level image. Let $\mathbf{Z}_n^w$ denote the feature representation of the nth WSI $\mathbf{X}_n$, $\mathbf{Z}_{ni}^b$ represent the feature representation of $\mathbf{X}_{ni}$ (which is the ith bag in $\mathbf{X}_n$), $\mathbf{Z}_{nij}^p$ be the feature representation of $\mathbf{X}_{nij}$ (which is the jth patch in $\mathbf{X}_{ni}$), and $\mathbf{Z}_{nijk}^c$ mean the feature representation of $\mathbf{X}_{nijk}$ (which represents the kth cell-level image of $\mathbf{X}_{nij}$). Note that $\mathbf{Z}_{nijk}^c$ is obtained by using the $L-1$ convolutional layers to extract features from the patch-level image $\mathbf{X}_{nij}$. E.g., considering that the input is a single patch with a size of 128×128, and its output after the $L-1$ convolutional layer in the first branch is transposed to $4\times 4\times 512$, which corresponds to the representations of 16 cell-level images in the patch.

Each WSI might be composed of certain bags with wrong labels, each bag might contain some trivial patches and each patch might consist of cell-level noise information. Thus, we propose a multi-scale representation attention based deep framework to simultaneously mine the bag-level images with correct labels, significant patch-level images and cell-level information. The proposed multi-scale representation attention consists of cell-level attention, patch-level attention and bag-level attention. They are on the basis of the $\ell_{2,1}$-norm based attention (Shi et al., 2020c), we describe their respective processes in this subsection.

3.2.1. Cell-level attention

$\{\mathbf{Z}_{nijk}^c\}_{k=1}^{M_c}$ obtained from the $L-1$ convolutional layers denote the cell-level feature representations in the patch-level image $\mathbf{X}_{nij}$. They are the input of the cell-level attention, which is:

$$\mathbf{P}_{nijk}^c = h(\mathbf{Z}_{nijk}^c), \tag{1a}$$

$$\eta_{nijk} = \frac{\sqrt{\sum_{r=1}^R (\mathbf{P}_{nijkr}^c)^2}}{\sum_{t=1}^{M_c}\sqrt{\sum_{r=1}^R (\mathbf{P}_{nijtr}^c)^2}}, \tag{1b}$$

$$\eta_{nijk} = \frac{max(\eta_{nijk} - \frac{\xi}{M_c}, 0)}{\sum_{t=1}^{M_c} max(\eta_{nijt} - \frac{\xi}{M_c}, 0)}, \tag{1c}$$

$$\mathbf{Z}_{nij}^p = \sum_{t=1}^{M_C} \eta_{nijt}\mathbf{Z}_{nijt}^c, \tag{1d}$$

where $\mathbf{P}_{nijk}^c$ and η_{nijk} denote the class vector and the attention weight of the cell-level image $\mathbf{X}_{nijk}$, respectively, θ^L represents the parameters of the fully connected layer, $\xi \in [0, 1]$ and $\frac{\xi}{M_c}$ is a threshold to remove trivial cell-level images in each patch-level image, $\mathbf{Z}_{nij}^p$ is the feature representation of the patch-level image $\mathbf{X}_{nij}$, and R is the number of classes.

3.2.2. Patch-level attention

Based on the cell-level attention, it can obtain the patch-level feature representations $\{\mathbf{Z}_{nij}^p\}_{j=1}^{M_p}$, which are fed into the patch-level attention. It is:

$$\mathbf{P}_{nij}^p = h(\mathbf{Z}_{nij}^p), \tag{2a}$$

$$\delta_{nij} = \frac{\sqrt{\sum_{r=1}^R (\mathbf{P}_{nijr}^p)^2}}{\sum_{t=1}^{M_p}\sqrt{\sum_{r=1}^R (\mathbf{P}_{nitr}^p)^2}}, \tag{2b}$$

$$\delta_{nij} = \frac{max(\delta_{nij} - \frac{\xi}{M_p}, 0)}{\sum_{t=1}^{M_p} max(\delta_{nit} - \frac{\xi}{M_p}, 0)}, \tag{2c}$$

$$\mathbf{Z}_{ni}^b = \sum_{t=1}^{M_p} \delta_{nit}\mathbf{Z}_{nit}^p, \tag{2d}$$

$$\mathbf{P}_{ni}^{b_1} = h(\mathbf{Z}_{ni}^b), \tag{2e}$$

where $\mathbf{P}_{nij}^p$ and δ_{nij} denote the class vector and the attention weight of the patch-level image $\mathbf{X}_{nij}$, respectively, $\frac{\xi}{M_p}$ is a threshold to remove trivial patch-level images in each bag, and $\mathbf{Z}_{ni}^b$ is the feature representation of the bag-level image $\mathbf{X}_{ni}$.

3.2.3. Bag-level attention

Based on the patch-level attention, it can obtain the bag-level feature representations $\{\mathbf{Z}_{ni}^b\}_{i=1}^{M_b}$, which are fed into the bag-level attention in the second branch. The process of bag-level attention is:

$$\mathbf{P}_{ni}^{b_2} = h(\mathbf{Z}_{ni}^b), \tag{3a}$$

$$\gamma_{ni} = \frac{\sqrt{\sum_{r=1}^R (\mathbf{P}_{nir}^b)^2}}{\sum_{t=1}^{M_b}\sqrt{\sum_{r=1}^R (\mathbf{P}_{ntr}^b)^2}}, \tag{3b}$$

$$\gamma_{ni} = \frac{max(\gamma_{ni} - \frac{\xi}{M_b}, 0)}{\sum_{t=1}^{M_b} max(\gamma_{ni} - \frac{\xi}{M_b}, 0)}, \tag{3c}$$

$$\mathbf{Z}_n^w = \sum_{t=1}^{M_b} \gamma_{nt}\mathbf{Z}_{nt}^b, \tag{3d}$$

$$\mathbf{P}_n^w = h(\mathbf{Z}_n^w), \tag{3e}$$

where $\mathbf{P}_{ni}^b$ and γ_{ni} denote the class vector and the attention weight of the bag-level image $\mathbf{X}_{ni}$ in the second branch, respectively, $h(\cdot)$ denotes the two fully connected layer with the parameters $\{\theta_l\}_{l=L+1}^{L+2}$, $\frac{\xi}{M_b}$ is a threshold to remove trivial bags in each WSI, and $\mathbf{Z}_n^w$ is the feature representation of the WSI $\mathbf{X}_n$.

For clarity, we present an example of the multi-scale representation attention used with the backbone network ResNet18 and the two fully connected layers constituted of the second branch in Fig. 2.

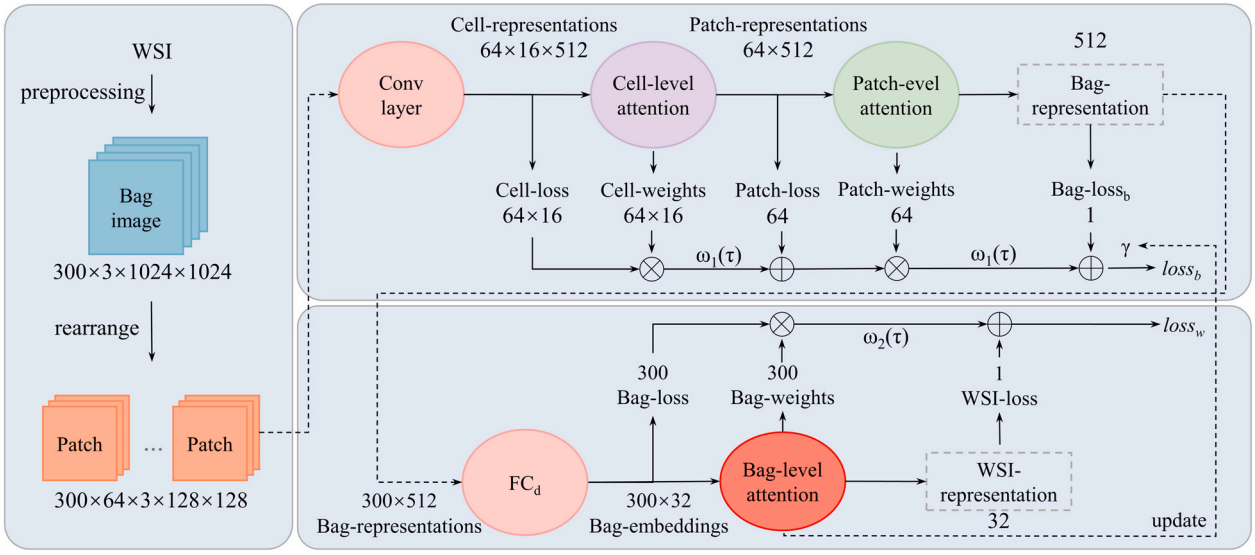


Fig. 2. An example of one WSI with its outputs during the preprocessing stage and the two branches of the proposed framework. During the right two subfigures, the top one shows the output of one bag in the first branch, and the bottom one displays the output of one WSI in the second branch.

3.3. Loss function

Given a batch of single-label WSIs $\mathbf{X}_{n\in\mathcal{B}}$ and their corresponding one-hot label vectors $\mathbf{y}_{n\in\mathcal{B}}$, where $\mathbf{X}_n \in \mathbb{R}^{C\times H\times W}$, $\mathbf{y}_n = \{y_{nr}\}^R$, $y_{nr} \in \{0,1\}$ and R is the number of classes, let $\mathbf{P}_n^w$, $\mathbf{P}_{nij}^p$ and $\mathbf{P}_{nijk}^c$ denote the class vector of $\mathbf{X}_n$, $\mathbf{X}_{nij}$ and $\mathbf{X}_{nijk}$, $\mathbf{P}_{ni}^{b_1}$ and $\mathbf{P}_{ni}^{b_2}$ represent the class vector of $\mathbf{X}_{ni}$ in the first and second branches, respectively. Because the framework consists of two branches, the proposed loss function also contains two loss functions, which are based on the popular cross-entropy loss as follows:

$$loss(\mathbf{y}_n;\mathbf{P}_n^w) = -\sum_{r=1}^{R} y_{nr} log(s(\mathbf{P}_n^w)[r]), \tag{4}$$

where $\mathbf{P}_n^w \in \mathbb{R}^R$ and $s(\cdot)$ denotes the softmax function. Additionally, because $\mathbf{X}_n$ is a single-label image and $\mathbf{y}_n \in \{0,1\}^R$, we can obtain $\sum_{r=1}^R y_{nr} = 1$. Suppose that $\mathbf{X}_n$ belongs to the tth class, Eq. (4) equals

$$loss(\mathbf{y}_n;\mathbf{P}_n^w) = -log((s(\mathbf{P}_n^w))[t]). \tag{5}$$

Since previous literature (Shi et al., 2020b,c) has demonstrated that connecting the attention mechanism and loss function can boost the model classification performance and interpretability, we also connect the bag-level attention with the loss function for WSI classification. Hence, the loss function for the second branch is

$$loss_w(y_{n\in\mathcal{B}};\mathbf{P}_{n\in\mathcal{B}}^w;\mathbf{P}_{n\in\mathcal{B}}^{b_2}) = -\frac{1}{|\mathcal{B}|}\sum_{\mathbf{X}_{n\in\mathcal{B}}}\left(log(s(\mathbf{P}_n^w)[t]) + \omega_2(\tau)\sum_{i=1}^{M_b}\gamma_{ni} log(s(\mathbf{P}_{ni}^{b_2})[t])\right), \tag{6}$$

where $|\mathcal{B}|$ denotes the number of WSIs, the first term is used for WSI classification and the second term is utilized for bag classification, $\omega_2(\tau)$ is an unsupervised weight function to balance the WSI and bag classification, τ is the number of current training epochs, and γ_{ni} obtained by the bag-level attention is the weight of the ith bag in the nth WSI.

Additionally, each class usually has different numbers of WSIs in practice, thereby leading to the popular unbalanced problem, which might significantly degrade the model performance. To alleviate this problem, we add a class weight for each WSI, and thus Eq. (6) becomes

$$loss_w(y_{n\in\mathcal{B}};\mathbf{P}_{n\in\mathcal{B}}^w;\mathbf{P}_{n\in\mathcal{B}}^{b_2}) = -\frac{1}{|\mathcal{B}|}\sum_{\mathbf{X}_{n\in\mathcal{B}}}\beta_t\left(log(s(\mathbf{P}_n^w)[t]) + \omega_2(\tau)\sum_{i=1}^{M_b}\gamma_{ni} log(s(\mathbf{P}_{ni}^{b_2})[t])\right), \tag{7}$$

where $\beta_t = \frac{\frac{N}{N_t}}{\sum_{r=1}^R \frac{N}{N_r}}$ represents the weight of the tth class, $N = \sum_{r=1}^R N_r$ is the total number of WSIs, and N_r is the number of WSIs belonging to the rth class.

Similarly, for the first branch, we connect the patch-level attention and cell-level attention with the loss function, which is simultaneously employed for bag-level, patch-level and cell-level images classification. It is

$$loss_b(\mathbf{y}_{ni\in S};\mathbf{P}_{ni\in S}^{b_1};\mathbf{P}_{ni\in S}^{p};\mathbf{P}_{ni\in S}^{c}) = -\frac{1}{|S|}\sum_{\mathbf{X}_{ni\in S}}\beta_t\left(log(s(\mathbf{P}_{ni}^{b_1})[t]) + \omega_1(\tau)\sum_{j=1}^{M_p}\delta_{nij}\left(log(s(\mathbf{P}_{nij}^{p})[t]) + \omega_1(\tau)\sum_{k=1}^{M_c}\eta_{nijk} log(s(\mathbf{P}_{nijk}^{c})[t])\right)\right), \tag{8}$$

where S is a set containing the indices of bags in different WSIs, $|S|$ denotes the number of bags, $\omega_1(\tau)$ is used to balance the bag-level and patch-level classification, and is used to balance the patch-level and cell-level classification at the same time, δ_{nij} is the weight of the jth patch in the ith bag of the nth WSI, i.e., $\mathbf{X}_{nij}$, and η_{nijk} is the weight of the kth cell-level image in the jth patch of the ith bag in the nth WSI, i.e., $\mathbf{X}_{nijk}$.

Based on the bag-level attention in the second branch, we can reduce the effect of bags with wrong labels by using the attention weights $\{\gamma_{ni}\}_{i=1}^{M_b}$ in the nth WSI, and then Eq. (8) becomes

$$loss_b(\mathbf{y}_{ni\in S};\mathbf{P}_{ni\in S}^{b_1};\mathbf{P}_{ni\in S}^{p};\mathbf{P}_{ni\in S}^{c}) = -\frac{1}{|S|}\sum_{\mathbf{X}_{ni\in S}}\gamma_{ni}\beta_t\left(log(s(\mathbf{P}_{ni}^{b_1})[t]) + \omega_1(\tau)\sum_{j=1}^{M_p}\delta_{nij}\left(log(s(\mathbf{P}_{nij}^{p})[t]) + \omega_1(\tau)\sum_{k=1}^{M_c}\eta_{nijk} log(s(\mathbf{P}_{nijk}^{c})[t])\right)\right), \tag{9}$$

Finally, based on Eqs. (7) and (9), we can obtain the proposed loss function as follows

$$loss = \upsilon loss_w(y_{n\in\mathcal{B}};\mathbf{P}_{n\in\mathcal{B}}^w;\mathbf{P}_{n\in\mathcal{B}}^{b_2}) + (1-\upsilon)loss_b(\mathbf{y}_{ni\in S};\mathbf{P}_{ni\in S}^{b_1};\mathbf{P}_{ni\in S}^{p};\mathbf{P}_{ni\in S}^{c}). \tag{10}$$

where $\upsilon \in \{0,1\}$ is to control which branch is employed to train the model during the training process. For clarity, we present the detailed training process of the proposed framework in Algorithm 1.

Algorithm 1: MRAN

Input: Training WSIs $\{\mathbf{X}_n\}_{n=1}^{N}$, label vectors $\{\mathbf{y}_n\}_{n=1}^{N}$, weight function $\omega_1(\tau)$, $\omega_2(\tau)$, network with parameters θ: $f_\theta(\cdot)$, augmentation function $g(\cdot)$.

Output: Parameters θ.

Preprocess $\mathbf{X}_n$ to obtain $\{\mathbf{X}_{ni}\}_{i=1}^{M_b}$ for $\forall 1 \le n \le N$;
Initialize $\gamma_{ni} = 1$ for $\forall 1 \le n \le N$ and $\forall 1 \le i \le M_b$;
for $\tau \in [1, T]$ **do**
  $v \leftarrow 0$;
  for *ecah minibatch* $\mathcal{S}$ **do**
    Divide $\mathbf{X}_{ni}$ to attain $\{\mathbf{X}_{nij}\}_{j=1}^{M_p}$ for $\forall ni \in \mathcal{S}$;
    $\mathbf{Z}^c_{nijk} \leftarrow f_\theta(g(\mathbf{X}_{nij}))$ for $\forall 1 \le j \le M_p$ and $\forall ni \in \mathcal{S}$;
    // Cell-level attention, where $att(\cdot)$ denotes Eq. (1).
    $\mathbf{Z}^p_{nij}, \{\eta_{nijk}\}_{k=1}^{M_c} \leftarrow att(\{\mathbf{Z}^c_{nijk}\}_{k=1}^{M_c}, \theta^L)$ for $\forall 1 \le j \le M_p$ and $\forall ni \in \mathcal{S}$;
    // Patch-level attention, where $att(\cdot)$ denotes Eq. (2).
    $\mathbf{Z}^b_{ni}, \{\delta_{nij}\}_{j=1}^{M_p} \leftarrow att(\{\mathbf{Z}^p_{nij}\}_{j=1}^{M_p}, \theta^L)$ for $\forall ni \in \mathcal{S}$;
    $loss \leftarrow$ Eq.(10);
    Back-propagate $loss$ to update model parameters θ^l for $1 \le l \le L$;
  end
  $v \leftarrow 1$;
  for *each minibatch* $\mathcal{B}$ **do**
    // Bag-level attention.
    $\mathbf{Z}^w_n, \{\gamma_{ni}\}_{i=1}^{M_b} \leftarrow att(\{\mathbf{Z}^b_{ni}\}_{i=1}^{M_b}, \{\theta^l\}_{l=L+1}^{L+2})$ for $\forall n \in \mathcal{B}$;
    $loss \leftarrow$ Eq.(3);
    Back-propagate $loss$ to update model parameters θ^l for $L+1 \le l \le L+2$;
  end
end

4. Experiments

In this section, we conduct experiments on four popular WSI datasets from The Cancer Genome Atlas (TCGA) and CAMELYON16 to evaluate the proposed framework.

4.1. Experimental setting

4.1.1. Datasets

Lung Squamous Cell Carcinoma (LUSC): This dataset consists of 1100 slides from 495 patients, with 753 positive slides and 347 negative ones, whose average size is around 35,374 × 25,178 pixels. After preprocessing WSIs based on Section 3.1, the average number of bags in each slide is 271. We randomly adopt 40%, 20% and 40% patients for model training, validation and testing, respectively.

Breast Invasive Carcinoma (BRCA): This dataset is composed of 1962 slides from 1088 patients, including 1565 positive slides and 397 negative ones, whose average size is about 89,405 × 29,784 pixels. After preprocessing WSIs, each slide has an average of 701 bags. We randomly employ 60%, 20% and 20% patients for model training, validation and testing, respectively.

Stomach Adenocarcinoma (STAD): This dataset contains 749 slides from 428 subjects, with 626 positive slides and 123 negative ones, whose average size is around 60,198 × 35,254 pixels. After preprocessing WSIs, the average number of bags contained in each slide is 605. We randomly utilize 60%, 20% and 20% subjects for model training, validation and testing, respectively.

CAMELYON16: This dataset contains 398 slides, with 159 positive slides and 239 negative ones, whose average size is around 110,084 × 168,762 pixels. After preprocessing WSIs, the average number of bags contained in each slide is 4241. We randomly utilize 80%, 10% and 10% subjects for model training, validation and testing, respectively.

4.1.2. Comparison methods

To better illustrate the strength of the proposed framework, we compare it with the following popular methods:

MSKCC-MIL (Campanella et al., 2019): which is a two stage method that first trains the model with MIL at patch-level images and selects confident patch-level images by using the predicted scores in each WSI during model training, and then employs random forest for WSI classification.

MSKCC-RNN (Campanella et al., 2019): which utilizes the trained MSKCC-MIL model to extract the feature representations of patch-level images, and then adopts the RNN method to classify WSIs.

DSMIL (Li et al., 2021b): which adopts the self-supervised contrastive learning to extract the feature representations of patch-level images, and then employs MIL for WSI classification with using the self-attention to take into account the instance mutual information.

ABMIL (Ilse et al., 2018): which employs the pre-trained model of ResNet18 on ImageNet (Russakovsky et al., 2015) to extract features of patch-level images, and then adopts the gated attention to interpret the significance of patches for WSI classification.

CLAM-SB (Lu et al., 2021): which adopts the pre-trained model of ResNet18 on ImageNet (Russakovsky et al., 2015) as the feature extractor of patch-level images, and utilizes the SVM-based loss and single-branch attention for WSI classification.

CLAM-MB (Lu et al., 2021): which also utilizes the same pre-trained model as CLAM-SB for feature extraction, but employs the SVM-based loss and multi-branch attention to aggregate patch-level features for WSI classification.

TransMIL (Shao et al., 2021): which adopts the pre-trained model of ResNet18 on ImageNet (Russakovsky et al. 2015) as the feature extractor of patch-level images, and then utilizes the Transformer-based attention MIL method to classify WSIs.

DTFD-MIL (Zhang et al., 2022): which also adopts the pre-trained model of ResNet18 on ImageNet to extract features of patch-level images, and then combine gradient-based probability calculation and Attention-based MIL to classify WSIs.

4.1.3. Implementation details

We conduct experiments on a server with 8 NVIDIA GeForce 3090 (24 GB memory each), and implement the proposed framework by using the PyTorch framework. We employ the Adam (Kingma and Ba, 2014) optimizer to update model parameters. For the first branch, we utilize the OneCycleLR (Smith and Topin, 2019) scheduler to adjust the learning rate, with setting the initial learning rate as 0.0001; For the second branch, we only update the parameters of the two fully connected layers by setting the learning rate as 0.0001 during the first 5 epochs, and then adjust the learning rate as 0.00005. Totally, we run the model 20 epochs, with alternatively training the first branch and the second one. Additionally, we set the batch size as 2 for the first branch, and adopt the batch size as 1 for the second branch. In Eq. (3c), same as Shi et al. (2020c), we set ξ as 0.1 for the cell-level, patch-level and bag-level attention. For the unsupervised weighting function $\omega(\tau)$, we adopt $\omega_i(\tau) = \lambda_i e^{-|1-\frac{\tau}{T}|^2}, i \in \{1, 2\}$, and empirically set $\lambda = 5$, where $\frac{\tau}{T}$ linearly increases from 0 to 1 during the training process and T is total number of training epochs. For data augmentation, we augment each patch-level image by only using the random flip.

For fairness, we employ the same patch size and downsampling rate as the proposed framework for the comparison methods, but we adopt their default augmentation to obtain the best performance of themselves.

Table 1
Classification results on the four datasets, LUSC, BRCA, STAD and CAMELYON16, We bold the best result and underline the second-best result of each setting.

Model	LUSC				
	ACC	AUC	SE	SP	F_1
MSKCC-MIL	0.8579 ± 0.0383	0.9427 ± 0.0167	0.9428 ± 0.0933	0.6781 ± 0.1792	0.9001 ± 0.0300
MSKCC-RNN	0.8962 ± 0.0293	0.9436 ± 0.0244	0.9672 ± 0.0161	0.7437 ± 0.1157	0.9278 ± 0.0184
DSMIL	0.9409 ± 0.0117	0.9846 ± 0.0054	0.9578 ± 0.0127	0.9050 ± 0.0125	0.9564 ± 0.0092
CLAM-SB	0.8366 ± 0.0084	0.9003 ± 0.0069	0.8955 ± 0.0242	0.7149 ± 0.0589	0.8805 ± 0.0061
CLAM-MB	0.8464 ± 0.0149	0.9249 ± 0.0070	0.8623 ± 0.0321	0.8137 ± 0.0337	0.8828 ± 0.0142
ABMIL	0.8139 ± 0.0455	0.8959 ± 0.0111	0.9124 ± 0.0558	0.5942 ± 0.2658	0.8717 ± 0.0196
TransMIL	0.8695 ± 0.0144	0.9156 ± 0.0126	0.9235 ± 0.0099	0.7517 ± 0.0239	0.9065 ± 0.0110
DTFD-MIL	0.8808 ± 0.0161	0.9223 ± 0.0273	0.9054 ± 0.0098	0.8268 ± 0.0555	0.9125 ± 0.0106
MRAN	**0.9719 ± 0.0078**	**0.9954 ± 0.0021**	**0.9825 ± 0.0066**	**0.9483 ± 0.0216**	**0.9796 ± 0.0054**
Model	BRCA				
	ACC	AUC	SE	SP	F_1
MSKCC-MIL	0.9027 ± 0.0424	0.9855 ± 0.0126	**0.9981 ± 0.0017**	0.5202 ± 0.1921	0.9430 ± 0.0237
MSKCC-RNN	0.9584 ± 0.0259	0.9735 ± 0.0213	0.9905 ± 0.0039	0.8317 ± 0.1209	0.9746 ± 0.0155
DSMIL	0.9062 ± 0.0693	0.9728 ± 0.0096	0.9162 ± 0.1044	0.8712 ± 0.1046	0.9366 ± 0.0527
CLAM-SB	0.9252 ± 0.0101	0.9754 ± 0.0039	0.9310 ± 0.0189	0.9055 ± 0.0277	0.9509 ± 0.0068
CLAM-MB	0.8765 ± 0.0149	0.9487 ± 0.0034	0.8894 ± 0.0383	0.8347 ± 0.0658	0.9179 ± 0.0111
ABMIL	0.9432 ± 0.0127	0.9765 ± 0.0127	0.9652 ± 0.0077	0.8525 ± 0.0447	0.9646 ± 0.0076
TransMIL	0.9714 ± 0.0158	0.9887 ± 0.0058	0.9874 ± 0.0088	0.9084 ± 0.0588	0.9822 ± 0.0099
DTFD-MIL	0.9768 ± 0.0092	0.9904 ± 0.0115	0.9792 ± 0.0067	**0.9728 ± 0.0260**	0.9850 ± 0.0059
MRAN	**0.9775 ± 0.0169**	**0.9923 ± 0.0078**	0.9818 ± 0.0275	0.9594 ± 0.0444	**0.9857 ± 0.0112**
Model	STAD				
	ACC	AUC	SE	SP	F_1
MSKCC-MIL	0.8382 ± 0.0137	0.8028 ± 0.0520	0.9708 ± 0.0300	0.2025 ± 0.1224	0.9086 ± 0.0080
MSKCC-RNN	0.7214 ± 0.0323	0.7385 ± 0.0472	0.7623 ± 0.0342	0.5190 ± 0.1171	0.8189 ± 0.0257
DSMIL	0.8024 ± 0.0191	0.7514 ± 0.0154	0.8889 ± 0.0349	0.3836 ± 0.1120	0.8812 ± 0.0145
CLAM-SB	0.8322 ± 0.0204	0.8401 ± 0.0435	0.9710 ± 0.0217	0.2340 ± 0.1489	0.9037 ± 0.0117
CLAM-MB	0.8647 ± 0.0315	0.8599 ± 0.0505	0.9286 ± 0.0367	0.5834 ± 0.2112	0.9176 ± 0.0193
ABMIL	0.6777 ± 0.0354	0.7303 ± 0.0869	0.6852 ± 0.0455	0.6375 ± 0.1746	0.7786 ± 0.0267
TransMIL	0.8431 ± 0.0039	0.7534 ± 0.0165	**0.9939 ± 0.0101**	0.1099 ± 0.0558	0.9131 ± 0.0028
DTFD-MIL	0.7948 ± 0.0722	0.8553 ± 0.0338	0.7898 ± 0.1009	0.8221 ± 0.0936	0.8621 ± 0.0544
MRAN	**0.9520 ± 0.0321**	**0.9849 ± 0.0160**	0.9679 ± 0.0108	**0.8720 ± 0.2208**	**0.9715 ± 0.0181**
Model	CAMELYON16				
	ACC	AUC	SE	SP	F_1
MSKCC-MIL	0.5875 ± 0.0479	0.5607 ± 0.0773	0.5833 ± 0.1836	0.5900 ± 0.1322	0.5064 ± 0.0838
MSKCC-RNN	0.4949 ± 0.0731	0.4856 ± 0.0654	0.4925 ± 0.1965	0.4958 ± 0.2288	0.5548 ± 0.1523
DSMIL	0.6190 ± 0.1154	0.6936 ± 0.0805	**0.6131 ± 0.0595**	0.7294 ± 0.1289	0.5440 ± 0.2602
CLAM-SB	0.5600 ± 0.1306	0.5745 ± 0.1238	0.5125 ± 0.3406	0.5917 ± 0.3935	0.4266 ± 0.2453
CLAM-MB	0.5750 ± 0.1275	0.5026 ± 0.1430	0.3625 ± 0.2666	0.7167 ± 0.1985	0.3748 ± 0.2495
ABMIL	0.5150 ± 0.1040	0.4805 ± 0.1114	0.4933 ± 0.2385	0.5280 ± 0.1507	0.4162 ± 0.1715
TransMIL	0.6150 ± 0.0454	0.6251 ± 0.0308	0.3733 ± 0.0760	0.7600 ± 0.1131	0.4185 ± 0.0335
DTFD-MIL	0.6250 ± 0.1075	0.5685 ± 0.1509	0.4800 ± 0.2883	0.7120 ± 0.2390	0.4559 ± 0.1908
MRAN	**0.7300 ± 0.0694**	**0.7939 ± 0.0872**	0.5733 ± 0.2140	**0.8240 ± 0.1345**	**0.5943 ± 0.1619**

Table 2
Ablation study on the BRCA dataset.

Model	ACC	AUC	F_1
Baseline	0.8737 ± 0.0539	0.9578 ± 0.0262	0.9133 ± 0.0412
B_α	0.9066 ± 0.0765	0.9971 ± 0.0029	0.9467 ± 0.0477
$B_\alpha + P_\alpha$	0.9225 ± 0.0311	0.9865 ± 0.0075	0.9491 ± 0.0212
$B_\alpha + P_\alpha + C_\alpha$	0.9803 ± 0.0226	0.9976 ± 0.0020	0.9880 ± 0.0134

To evaluate the proposed framework and the comparison methods, we adopt five popular metrics, including accuracy (ACC), Area Under Curve (AUC), Sensitivity (SE), Specificity (SP) and F_1 score. Additionally, we repeat experiments five times for all methods on each dataset, and then report their average results.

4.2. Results and analysis

4.2.1. Classification experiments

Table 1 shows the classification results of the proposed framework and eight comparison methods on the four datasets. As we can see, the proposed framework achieves the best results among all the methods on the four datasets, in terms of the three metrics, especially for the dataset STAD, on which the gain of the proposed framework is 8.47%, 7.58% and 4.94% over the best competitor in terms of ACC, AUC, and F_1 score, respectively. This might because (1) the proposed framework trains the model in an end-to-end manner, which can directly utilize convolutional neural networks to extract features of all patch-level images for WSI classification; (2) the proposed framework with multi-scale representation attention can mine the significant multi-scale features, i.e., bag-level, patch-level and cell-level features.

4.2.2. Interpretation experiments

In addition to the classification experiments, we also conduct interpretation experiments to evaluate the model interpretability of the proposed framework.

Because the proposed framework consists of bag-level, patch-level and cell-level attention mechanisms, we first scale the attention map of each level and then superimpose it into the corresponding position of the original image. Next, to demonstrate their effectiveness on interpreting the significance of bag-level, patch-level and cell-level images, respectively, we present their heatmaps in Fig. 3, in which the position with darker color means that has a larger attention weight. Specifically, Fig. 3(a) shows the heatmaps obtained by the bag-level attention on WSIs to mine the significant bags, Fig. 3(b) displays the

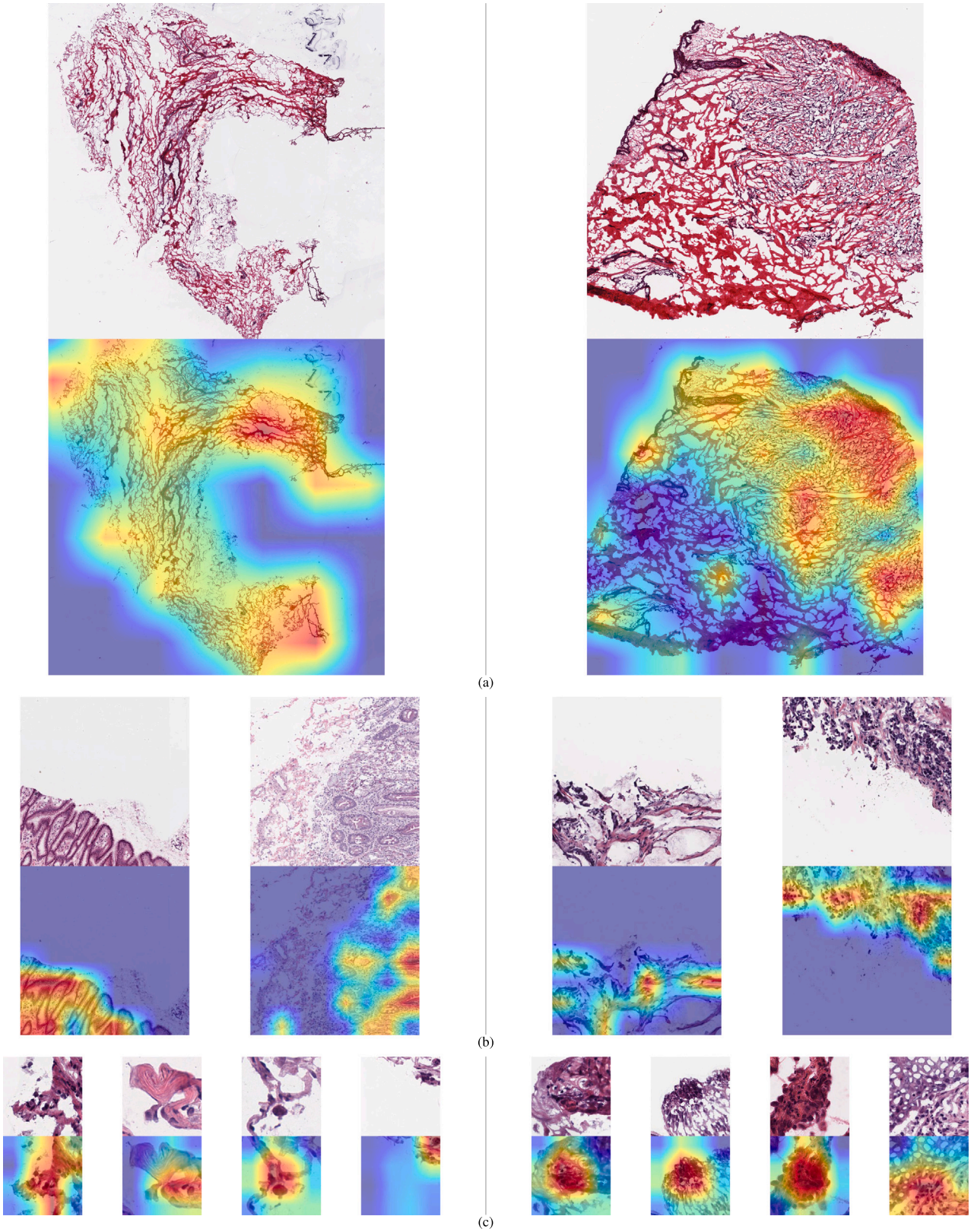
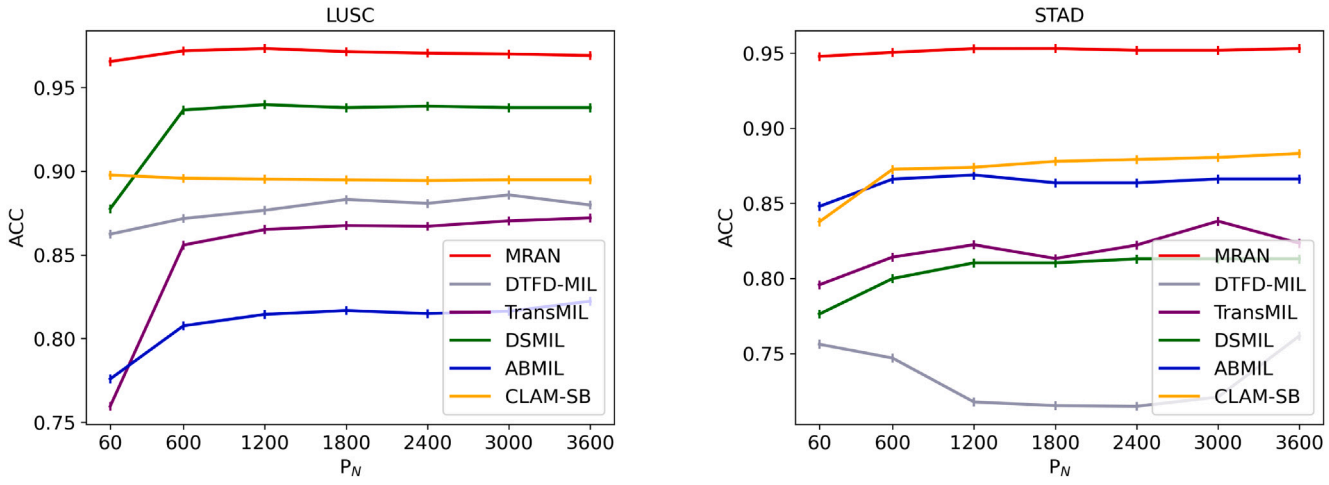
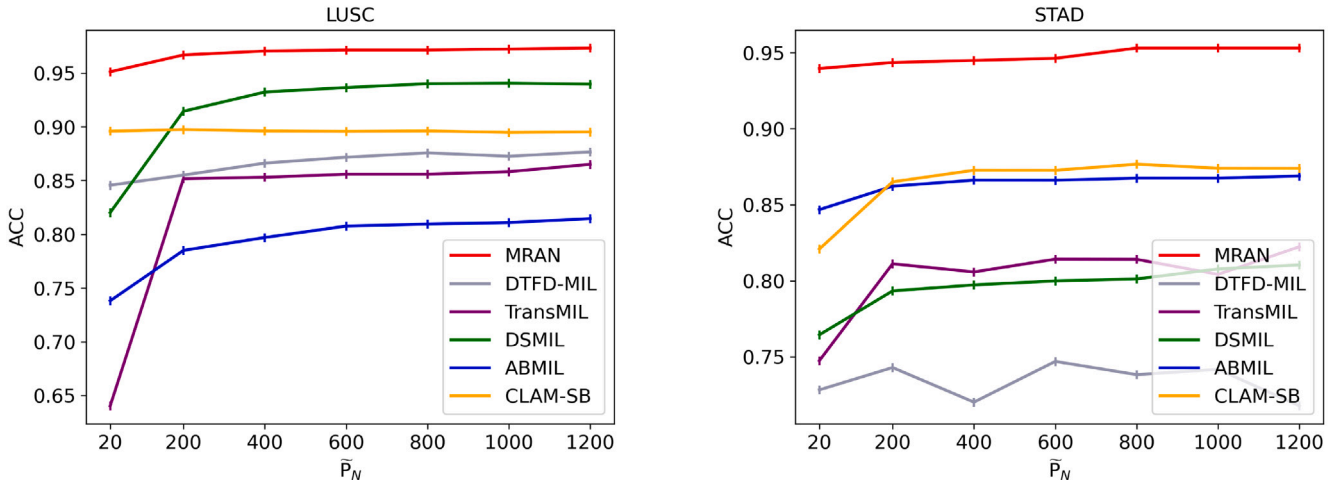


Fig. 3. Heat maps generated at three levels of attention. (a) Heat maps generated with Bag-level attention on WSIs. (b) Heat maps generated with Patch-level attention on Bag images. (c) Heat maps generated with Cell-level attention on Patch images. The left and right half of each row are from negative WSIs and positive WSIs, respectively.



(a)ACC vs. P_N



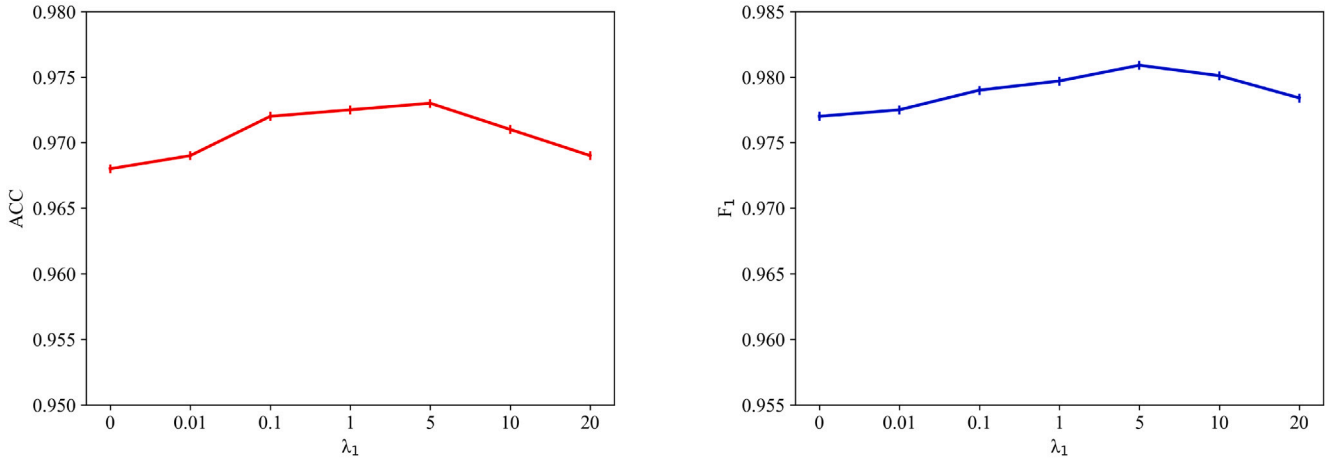
(b)ACC vs. $\widetilde{P}_N$

Fig. 4. The classification accuracy of four methods at different numbers of selected patches on the LUSC and STAD datasets. Both P_N and $\widetilde{P}_N$ are the total number of selected patches. However, in (a), our proposed MRAN selects the top 60 bags for each WSI, with the variable top selected number of patches in each bag; in (b), MRAN selects the top 20 patches for each bag, with the variable top selected number of bags in each WSI.

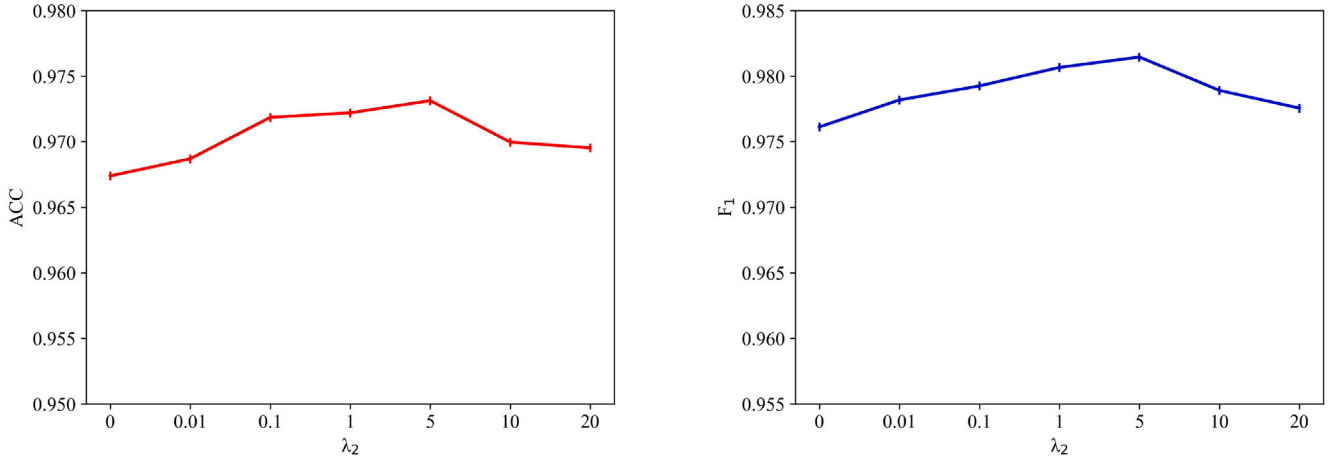
heatmaps attained by the patch-level attention on bag-level images to mine the significant patches, and Fig. 3(c) presents the heatmaps achieved by the cell-level attention on patch-level images to mine the significant cell-level images. As we can see, Fig. 3 suggests that multi-scale representation attention can interpret the significance of different levels of images in WSIs.

To further evaluate the interpretation performance of the proposed framework, because most of previous WSI-based methods focus on interpreting the significance of patch-level images, whose annotation is very time consuming and expensive, similar to previous feature section methods (Ding and Peng, 2005; Nie et al., 2010), we employ classification accuracy as the metric to assess the selected patch-level images of the proposed framework, without using any additional prior knowledge. We compare the proposed framework with five comparative interpretation methods: DSMIL, ABMIL, CLAM-SB, TransMIL and DTFD-MIL. Specifically, we first adopt the selected significant patch-level images in the original testing set of the four methods as a new testing set, and then evaluate their classification performance on the new testing set. Note that for fairness, all methods select the same number of patch-level images in each WSI. Fig. 4 presents the classification accuracy on different numbers of selected patch-level images. It suggests that the proposed framework can achieve the best classification accuracy when selecting the same number of patch-level images. Additionally, the larger number of selected patches usually mean the better model performance. This infers that more patches are beneficial to the model generalizability. However, the proposed framework can obtain the best or sub-best results of itself by using fewer patches compared to the other comparison methods. This illustrates its powerful selection performance on patch-level images.

To more accurately explore the effect of bag-level attention and patch-level attention on model performance, we also conduct experiments to observe that (1) different numbers of selected patches with a fixed number of selected bags (Fig. 6(a)); (2) a fixed number of selected patches with different numbers of selected bags (Fig. 6(b)). As we can see, Fig. 6(a) displays that when each bag has the number of selected patches during $[10, 30]$, the proposed framework achieves the best or sub-best classification accuracy. This suggests that each bag consists of many trivial or even noise patches, and the proposed patch-level attention can mine the significant patches and remove the trivial ones. Additionally, Fig. 6(b) shows that the proposed framework can obtain the sub-best performance with using only 20 bags, which is less than 10% of the total number of bag-level images. This infers that each WSI might contain a set of trivial bag-level images, and the proposed bag-level attention can mine the significant ones.

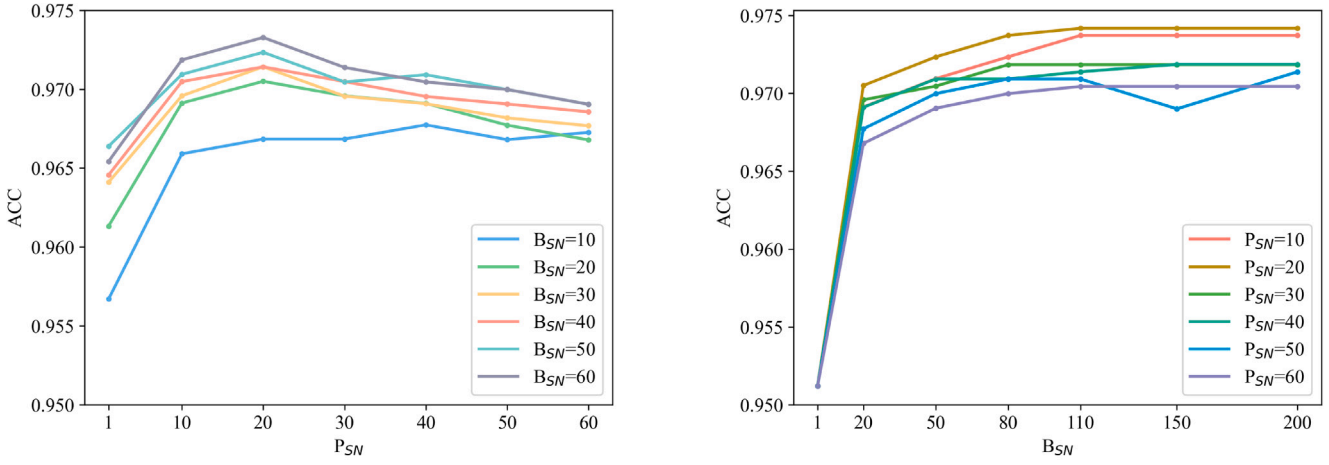


(a)ACC vs. λ_1



(b)ACC vs. λ_2

Fig. 5. (a) The accuracy and F_1 score of the proposed MRAN on the LUSC dataset under different values of λ_1 with fixing $\lambda_2 = 5$; (b) the accuracy and F_1 score of the proposed MRAN on the LUSC dataset under different values of λ_2 with fixing $\lambda_1 = 5$.



(a) ACC vs. P_{SN}

(b) ACC vs. B_{SN}

Fig. 6. The effect of patch-level and bag-level attention mechanisms on model performance on the dataset LUSC. Note that P_{SN} and B_{SN} denote the number of selected patches in each bag and selected bags in each WSI, respectively.

4.2.3. Ablation study

To evaluate the influence of bag-, patch- and cell-level attention on model performance, we conduct experiments by replacing the attention at each level with the mean-pooling operator, and then gradually add bag-, patch- and cell-level attention mechanisms. Specifically, Baseline denotes that the attention at each level is replaced with mean-pooling, i.e., the images at each level have the equivalent weight. B_α denotes replacing the bag-level mean-pooling with the bag-level attention in Baseline, $B_\alpha + P_\alpha$ represents replacing the patch-level mean-pooling with the patch-level attention in B_α, and $B_\alpha + P_\alpha + C_\alpha$ is to replace the cell-level mean-pooling with the cell-level attention in $B_\alpha + P_\alpha$. Table 2 shows the classification results of the proposed framework with different levels of attention mechanisms. It suggests that all the bag-, patch- and cell-level attentions are conducive to boosting model performance. Similar observations can be found on the other three datasets.

4.3. Parameters analysis

We explore the effect of the hyperparameter λ_i in $\omega_i(\tau)$, $i \in \{1, 2\}$, on the performance of the proposed MRAN. Specifically, we conduct experiments with different values of λ_i during $[0, 20]$ respectively, and show the results in terms of accuracy and F_1 score on LUSC in Fig. 5. As we can see, MRAN can obtain the best or sub-best model performance within a large range of $[0.1, 10]$. This indicates that the proposed MRAN is not very sensitive to λ_i. Similar observations can be found on the other three datasets.

4.4. Discussion

Classification experiments on four WSIs datasets demonstrate that the proposed MRAN can obtain better classification performance. One major possible reason is that the comparison methods adopt two stages for WSI classification, e.g., MSKCC-MIL and MSKCC-RNN, unsupervised learning or pre-trained models to extract features of patch-level images, e.g., DSMIL, ABMIL, CLAB-SB, CLAM-MB, TransMIL and DTFD-MIL, while the proposed framework directly utilizes convolutional neural network to extract the feature representation of patch-level images for WSI classification. Additionally, interpretation experiments illustrate that the proposed framework has better model interpretability. Moreover, ablation study infers that mining the significant bag-level, patch-level and cell-level images is very beneficial to boosting model performance. Hence, they might suggest that multi-scale representation attention is another major cause of the proposed framework with better classification performance.

5. Conclusion and future work

In this paper, we propose a novel end-to-end interpretable deep MIL framework, namely MRAN, consisting of convolutional layers, multi-scale representation attention and fully connected layers. Additionally, the proposed multi-scale representation attention is composed of bag-, patch- and cell-level attention to mine the significant bags, patches and cell-level images, respectively. Extensive experiments on four popular and publicly available WSIs datasets illustrate that the proposed framework outperforms the recent state-of-the-art methods on model classification performance and interpretability, and also demonstrate the effectiveness and strength of the proposed multi-scale representation attention.

Although the proposed framework has obtained promising performance, it still has some limitations as follows: (1) failing to consider the relationship among instances; (2) originally designed for binary classification tasks; (3) only taking into account unimodal data, i.e. WSIs. Therefore, in the future, we will boost the proposed framework by considering the relations among instances via self-attention or graph methods, and extend it to multi-class tasks and multimodal data.

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 was supported by the National Key Research and Development Program of China (No. 2022YFA1004100), National Natural Science Foundation of China (No. 62276052 and No. 82100650), and Medico-Engineering Cooperation Funds from University of Electronic Science and Technology of China (No. ZYGX2022YGRH009 and No. ZYGX2022YGRH014).

References

Araújo, T., Aresta, G., Castro, E., Rouco, J., Aguiar, P., Eloy, C., Polónia, A., Campilho, A., 2017. Classification of breast cancer histology images using convolutional neural networks. PLoS One 12 (6), e0177544.

Bae, W., Noh, J., Kim, G., 2020. Rethinking class activation mapping for weakly supervised object localization. In: Proceedings of European Conference on Computer Vision. pp. 618–634.

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

Chen, P., El Hussein, S., Xing, F., Aminu, M., Kannapiran, A., Hazle, J.D., Medeiros, L.J., Wistuba, I.I., Jaffray, D., Khoury, J.D., et al., 2022. Chronic lymphocytic leukemia progression diagnosis with intrinsic cellular patterns via unsupervised clustering. Cancers 14 (10), 2398.

Chen, P.-H.C., Gadepalli, K., MacDonald, R., Liu, Y., Kadowaki, S., Nagpal, K., Kohlberger, T., Dean, J., Corrado, G.S., Hipp, J.D., et al., 2019b. An augmented reality microscope with real-time artificial intelligence integration for cancer diagnosis. Nat. Med. 25 (9), 1453–1457.

Chen, H., Han, X., Fan, X., Lou, X., Liu, H., Huang, J., Yao, J., 2019a. Rectified cross-entropy and upper transition loss for weakly supervised whole slide image classifier. In: Proceedings of International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 351–359.

Cheng, H.-T., Yeh, C.-F., Kuo, P.-C., Wei, A., Liu, K.-C., Ko, M.-C., Chao, K.-H., Peng, Y.-C., Liu, T.-L., 2020. Self-similarity student for partial label histopathology image segmentation. In: Proceedings of European Conference on Computer Vision. pp. 117–132.

Chikontwe, P., Kim, M., Nam, S.J., Go, H., Park, S.H., 2020. Multiple instance learning with center embeddings for histopathology classification. In: Proceedings of International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 519–528.

Cruz-Roa, A., Basavanhally, A., González, F., Gilmore, H., Feldman, M., Ganesan, S., Shih, N., Tomaszewski, J., Madabhushi, A., 2014. Automatic detection of invasive ductal carcinoma in whole slide images with convolutional neural networks. In: Proceedings of Medical Imaging 2014: Digital Pathology, Vol. 9041. 904103.

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

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

Dike, H.U., Zhou, Y., Deveerasetty, K.K., Wu, Q., 2018. Unsupervised learning based on artificial neural network: A review. In: Proceedings of IEEE International Conference on Cyborg and Bionic Systems. pp. 322–327.

Ding, C., Peng, H., 2005. Minimum redundancy feature selection from microarray gene expression data. J. Bioinform. Comput. Biol. 3 (02), 185–205.

Elmore, J., 2021. Abstract SY01-03: The gold standard cancer diagnosis: Studies of physician variability, interpretive behavior, and the impact of AI. Cancer Res. 81 (13_Supplement), SY01–SY03.

Gao, Z., Wang, L., Zhou, L., Zhang, J., 2016. HEp-2 cell image classification with deep convolutional neural networks. IEEE J. Biomed. Health Inf. 21 (2), 416–428.

Hashimoto, N., Fukushima, D., Koga, R., Takagi, Y., Ko, K., Kohno, K., Nakaguro, M., Nakamura, S., Hontani, H., Takeuchi, I., 2020. Multi-scale domain-adversarial multiple-instance CNN for cancer subtype classification with unannotated histopathological images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3852–3861.

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.
Hou, L., Samaras, D., Kurc, T.M., Gao, Y., Davis, J.E., Saltz, J.H., 2016. Patch-based convolutional neural network for whole slide tissue image classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2424–2433.
Hu, B., Tang, Y., Eric, I., Chang, C., Fan, Y., Lai, M., Xu, Y., 2018. Unsupervised learning for cell-level visual representation in histopathology images with generative adversarial networks. IEEE J. Biomed. Health Inf. 23 (3), 1316–1328.
Ilse, M., Tomczak, J., Welling, M., 2018. Attention-based deep multiple instance learning. In: Proceedings of International Conference on Machine Learning. pp. 2127–2136.
Kandemir, M., Hamprecht, F.A., 2015. Computer-aided diagnosis from weak supervision: A benchmarking study. Comput. Med. Imaging Graph. 42, 44–50.
Keikhosravi, A., Li, B., Liu, Y., Conklin, M.W., Loeffler, A.G., Eliceiri, K.W., 2020. Non-disruptive collagen characterization in clinical histopathology using cross-modality image synthesis. Commun. Biol. 3 (1), 1–12.
Kingma, D.P., Ba, J., 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.
Li, B., Keikhosravi, A., Loeffler, A.G., Eliceiri, K.W., 2021a. Single image super-resolution for whole slide image using convolutional neural networks and self-supervised color normalization. Med. Image Anal. 68, 101938.
Li, B., Li, Y., Eliceiri, K.W., 2021b. Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 14318–14328.
Li, J., Li, W., Gertych, A., Knudsen, B.S., Speier, W., Arnold, C.W., 2019. An attention-based multi-resolution model for prostate whole slide image classification and localization. arXiv preprint arXiv:1905.13208.
Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., Van Der Laak, J.A., Van Ginneken, B., Sánchez, C.I., 2017. A survey on deep learning in medical image analysis. Med. Image Anal. 42, 60–88.
Liu, Q., Yu, L., Luo, L., Dou, Q., Heng, P.A., 2020. Semi-supervised medical image classification with relation-driven self-ensembling model. IEEE Trans. Med. Imaging 39 (11), 3429–3440.
Lu, M.Y., Chen, R.J., Wang, J., Dillon, D., Mahmood, F., 2019. Semi-supervised histology classification using deep multiple instance learning and contrastive predictive coding. arXiv preprint arXiv:1910.10825.
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.
Malathy, C., Uddipt, S., Mayuri, N.C., Uma Pratheebha, U., 2016. A new approach for recognition of implant in knee by template matching. Indian J. Sci. Technol. 9 (37).
Marini, N., Otálora, S., Müller, H., Atzori, M., 2021. Semi-supervised training of deep convolutional neural networks with heterogeneous data and few local annotations: An experiment on prostate histopathology image classification. Med. Image Anal. 73, 102165.
Maron, O., Lozano-Pérez, T., 1997. A framework for multiple-instance learning. 10,
Mo, Y., Chen, Y., Lei, Y., Peng, L., Shi, X., Yuan, C., Zhu, X., 2023. Multiplex graph representation learning via dual correlation reduction. Proc. IEEE Trans. Knowl. Data Eng. http://dx.doi.org/10.1109/TKDE.2023.3268069.
Nie, F., Huang, H., Cai, X., Ding, C., 2010. Efficient and robust feature selection via joint ℓ2, 1-norms minimization. 23,
Peikari, M., Salama, S., Nofech-Mozes, S., Martel, A.L., 2018. A cluster-then-label semi-supervised learning approach for pathology image classification. Sci. Rep. 8 (1), 1–13.
Peng, L., Wang, N., Dvornek, N., Zhu, X., Li, X., 2022. FedNI: Federated graph learning with network inpainting for population-based disease prediction. p. 1. http://dx.doi.org/10.1109/TMI.2022.3188728,
Pirovano, A., Heuberger, H., Berlemont, S., Ladjal, S., Bloch, I., 2020. Improving interpretability for computer-aided diagnosis tools on whole slide imaging with multiple instance learning and gradient-based explanations. In: Interpretable and Annotation-Efficient Learning for Medical Image Computing. pp. 43–53.
Pulido, J.V., Guleria, S., Ehsan, L., Fasullo, M., Lippman, R., Mutha, P., Shah, T., Syed, S., Brown, D.E., 2020. Semi-supervised classification of noisy, gigapixel histology images. In: Proceedings of IEEE International Conference on Bioinformatics and Bioengineering. pp. 563–568.
Ren, J., Hacihaliloglu, I., Singer, E.A., Foran, D.J., Qi, X., 2019. Unsupervised domain adaptation for classification of histopathology whole-slide images. Front. Bioeng. Biotechnol. 7, 102.
Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al., 2015. Imagenet large scale visual recognition challenge. Int. J. Comput. Vis. 115 (3), 211–252.
Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D., 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 618–626.
Shao, Z., Bian, H., Chen, Y., Wang, Y., Zhang, J., Ji, X., et al., 2021. Transmil: Transformer based correlated multiple instance learning for whole slide image classification. 34, pp. 2136–2147,
Sharma, Y., Shrivastava, A., Ehsan, L., Moskaluk, C.A., Syed, S., Brown, D., 2021. Cluster-to-conquer: A framework for end-to-end multi-instance learning for whole slide image classification. In: Medical Imaging with Deep Learning. pp. 682–698.
Shi, X., Su, H., Xing, F., Liang, Y., Qu, G., Yang, L., 2020a. Graph temporal ensembling based semi-supervised convolutional neural network with noisy labels for histopathology image analysis. Med. Image Anal. 60, 101624.
Shi, X., Xing, F., Xie, Y., Su, H., Yang, L., 2017. Cell encoding for histopathology image classification. In: Proceedings of International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 30–38.
Shi, X., Xing, F., Xie, Y., Zhang, Z., Cui, L., Yang, L., 2020b. Loss-based attention for deep multiple instance learning. In: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34. pp. 5742–5749.
Shi, X., Xing, F., Xu, K., Chen, P., Liang, Y., Lu, Z., Guo, Z., 2020c. Loss-based attention for interpreting image-level prediction of convolutional neural networks. IEEE Trans. Image Process. 30, 1662–1675.
Shin, H.-K., Uhmn, K.-H., Choi, K., Xu, Z., Jung, S.-W., Ko, S.-J., 2022. Graph segmentation-based pseudo-labeling for semi-supervised pathology image classification. IEEE Access 10, 93960–93970.
Sirinukunwattana, K., Pluim, J.P., Chen, H., Qi, X., Heng, P.-A., Guo, Y.B., Wang, L.Y., Matuszewski, B.J., Bruni, E., Sanchez, U., et al., 2017. Gland segmentation in colon histology images: The glas challenge contest. Med. Image Anal. 35, 489–502.
Smith, L.N., Topin, N., 2019. Super-convergence: Very fast training of neural networks using large learning rates. In: Proceedings of Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications, Vol. 11006. pp. 369–386.
Tu, M., Huang, J., He, X., Zhou, B., 2019. Multiple instance learning with graph neural networks. arXiv preprint arXiv:1906.04881.
Wang, X., Chen, H., Xiang, H., Lin, H., Lin, X., Heng, P.-A., 2021. Deep virtual adversarial self-training with consistency regularization for semi-supervised medical image classification. Med. Image Anal. 70, 102010.
Xie, C., Muhammad, H., Vanderbilt, C.M., Caso, R., Yarlagadda, D.V.K., Campanella, G., Fuchs, T.J., 2020. Beyond classification: Whole slide tissue histopathology analysis by end-to-end part learning. In: Medical Imaging with Deep Learning. pp. 843–856.
Xu, Y., Jia, Z., Wang, L.-B., Ai, Y., Zhang, F., Lai, M., Chang, E.I., et al., 2017. Large scale tissue histopathology image classification, segmentation, and visualization via deep convolutional activation features. BMC Bioinformatics 18 (1), 1–17.
Xu, J., Ren, Y., Tang, H., Yang, Z., Pan, L., Yang, Y., Pu, X., Yu, P.S., He, L., 2022. Self-supervised discriminative feature learning for deep multi-view clustering. http://dx.doi.org/10.1109/TKDE.2022.3193569.
Xu, Y., Zhu, J.-Y., Eric, I., Chang, C., Lai, M., Tu, Z., 2014. Weakly supervised histopathology cancer image segmentation and classification. Med. Image Anal. 18 (3), 591–604.
Yan, Y., Wang, X., Guo, X., Fang, J., Liu, W., Huang, J., 2018. Deep multi-instance learning with dynamic pooling. In: Proceedings of Asian Conference on Machine Learning. pp. 662–677.
Zhang, C., Bengio, S., Hardt, M., Recht, B., Vinyals, O., 2021. Understanding deep learning (still) requires rethinking generalization. Commun. ACM 64 (3), 107–115.
Zhang, Z., Chen, P., McGough, M., Xing, F., Wang, C., Bui, M., Xie, Y., Sapkota, M., Cui, L., Dhillon, J., et al., 2019. Pathologist-level interpretable whole-slide cancer diagnosis with deep learning. Nat. Mach. Intell. 1 (5), 236–245.
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 Conference on Computer Vision and Pattern Recognition. pp. 18802–18812.
Zheng, Y., Jiang, Z., Zhang, H., Xie, F., Ma, Y., Shi, H., Zhao, Y., 2018. Histopathological whole slide image analysis using context-based CBIR. IEEE Trans. Med. Imaging 37 (7), 1641–1652.
Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A., 2016. Learning deep features for discriminative localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2921–2929.

Hangchen Xiang received the B.S. degree in computer science and technology from Shandong University, Shandong, China, in 2021. He is currently working toward the M.S. degree in department of Computer Science and Engineering at the University of Electronic Science and Technology of China. His research interests include deep learning and medical image analysis.

Junyi shen received the B.S. degree in clinical medicine(anesthesiology) from Chongqing medical university, Chongqing, China, in 2014 and the M.S. degree in clinical medicine from Sichuan University, Chengdu, China, in 2017 and the Ph.D. degree in clinical medicine with the School of Sichuan University, Chengdu, China in 2020. Currently, he works in the department of liver surgery in West China Hospital, Sichuan university. His research interests include the mechanism of liver cirrhosis and liver cancer and image study of cancer features.

Qingguo Yan is an associate professor in the School of Medicine at the Northwest University in China. He obtained his Ph.D. degree (2020), master's degree (1997) and

Bachelor's degree (1992) from the Fourth Military Medical University. His research interests including molecular pathology and digital pathology.

Meilian Xu received her B.E., M.Sc. and Ph.D., all in computer science, from East China Normal University, Peking University and the University of Manitoba, respectively. She is now a professor at the School of Electronic Information and Artificial Intelligence, Leshan Normal University. Her research interests include machine learning, nature inspired optimization algorithms, medical imaging, parallel algorithms design and performance improvement on different architectures. She is a member of the ACM society.

Xiaoshuang Shi is a professor in the Department of Computer Science and Engineering at the University of Electronic Science and Technology of China (UESTC). He obtained his Ph.D. degree (2019) from University of Florida, Master degree (2013) from Tsinghua University, and Bachelor degree (2009) from Northwestern Polytechnical University. Before joining UESTC, he worked as a Postdoctoral fellow at the National Institutes of Health (NIH) (2020.01-2021.04), and as a research assistant at Tsinghua University (2013.09-2015.04). His major research interests include large-scale data retrieval, deep learning, and medical image analysis.

Xiaofeng Zhu is a professor of University of Electronic Science and Technology of China, Chengdu, China. His current research interests include large-scale multimedia retrieval, feature selection, sparse learning, data preprocess, and medical image analysis.