Medical Image Analysis 101 (2025) 103468

Contents lists available at ScienceDirect

Medical Image Analysis

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

Dynamic graph based weakly supervised deep hashing for whole slide image classification and retrieval

Haochen Jin [a,1], Junyi Shen [b,1], Lei Cui [c], Xiaoshuang Shi [a,*], Kang Li [d,*], 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 Computer Science and Technology, Northwest University of China, 710075, China
[d] West China Medical Center, Sichuan University, Chengdu, 610041, China

ARTICLE INFO

Keywords:
Whole slide images
Attention-based MIL
Hashing encoding
Dynamic graph

ABSTRACT

Recently, a multi-scale representation attention based deep multiple instance learning method has proposed to directly extract patch-level image features from gigapixel whole slide images (WSIs), and achieved promising performance on multiple popular WSI datasets. However, it still has two major limitations: (i) without considering the relations among patches, thereby possibly restricting the model performance; (ii) unable to handle retrieval tasks, which is very important in clinic diagnosis. To overcome these limitations, in this paper, we propose a novel end-to-end MIL-based deep hashing framework, which is composed of a multi-scale representation attention based deep network as the backbone, patch-based dynamic graphs and hashing encoding layers, to simultaneously handle classification and retrieval tasks. Specifically, the multi-scale representation attention based deep network is to directly extract patch-level features from WSIs with mining the significant information at cell-, patch- and bag-level features. Additionally, we design a novel patch-based dynamic graph construction method to learn the relations among patches within each bag. Moreover, the hashing encoding layers are to encode patch- and WSI-level features into binary codes for patch- and WSI-level image retrieval. Extensive experiments on multiple popular datasets demonstrate that the proposed framework outperforms recent state-of-the-art ones on both classification and retrieval tasks. *All source codes are available at https://github.com/hcjin0816/DG_WSDH.*

1. Introduction

Benefiting from a large amount of high-quality labeled data, deep learning has achieved remarkable success in many scenarios of medical image analysis, even obtaining superior performance over humans (Liu et al., 2020; Yu et al., 2018; Zhang et al., 2019). Hence, due to the characteristics of histopathological whole-slide images (WSIs) (e.g., gigapixel image size, enormous heterogeneity, multiple cancer types), it usually requires pathologists to manually annotate a large number of significant regions from WSIs. However, manual annotation is time-consuming, laborious and even error-prone. In order to reduce pathologists' workload and further boost diagnostic accuracy, computer-aided diagnosis (CAD) systems have been developed for WSI analysis, by using computer vision and machine learning techniques (Sirinukunwattana et al., 2017; Zhang et al., 2019; Keikhosravi et al., 2020; Peng et al., 2022). Generally, CAD systems can be divided into two categories: classifier-based CAD and content-based image retrieval (CBIR). Compared to classifier-based CAD, CBIR systems can not only classify query images but also retrieve and visualize similar images (Zheng et al., 2003; Akakin and Gurcan, 2012). For clarity, we show the differences of classification, retrieval and hybrid models in Fig. 1. Therefore, CBIR techniques have attracted considerable attention in histopathological image analysis (Shi et al., 2017; Erfankhah et al., 2019; Zheng et al., 2022b).

Recently, convolutional neural networks have been widely used in CBIR systems with impressive performance (Litjens et al., 2017; Gour et al., 2020). Due to computer memory limitations, many patch-based methods, which first crop each WSI into hundreds or thousands patches and then annotate them for model training (Xu et al., 2017; Zhang et al., 2019), have been widely used for CBIR. However, patch-based methods still lead to laborious and expensive costs. To further reduce pathologists' workload, weakly supervised learning techniques, which only require the label of each WSI for model training, have attracted a

* Corresponding authors.
E-mail addresses: xsshi2013@gmail.com (X. Shi), likang@wchscu.cn (K. Li).
[1] Co-first authors: They have equal contribution.

https://doi.org/10.1016/j.media.2025.103468
Received 18 March 2024; Received in revised form 1 January 2025; Accepted 9 January 2025
Available online 23 January 2025

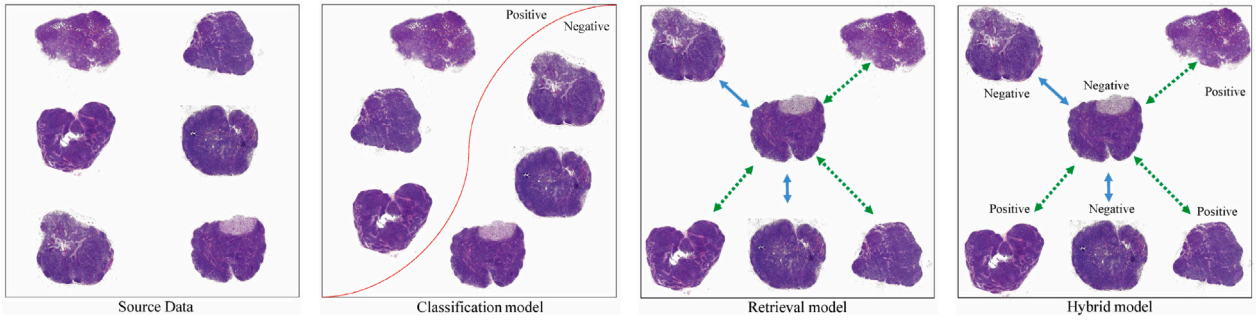


Fig. 1. Differences among the classification, retrieval and hybrid model. Classification model only can provide the class of WSIs; Retrieval model only can retrieve similar images to the query but cannot directly classify the query; Hybrid model can simultaneously retrieve the similar ones and classify the query. For the retrieval model and the hybrid model, the blue line represents similarity between two WSIs, while the green line indicates dissimilarity. The length of the line corresponds to the degree of similarity; shorter lines indicate higher similarity, and longer lines suggest lower similarity.

growing interest in histopathological analysis (Cruz-Roa et al., 2014; Xiang et al., 2023). Most existing weakly supervised learning techniques are based on multiple instance learning (MIL) (Hashimoto et al., 2020; Chikontwe et al., 2020). In MIL-based methods for WSIs, each WSI is regarded as a bag, and small regions or blocks in each WSI are denoted as instances, and the label indicates whether the entire WSI contains cancerous cells or not (Xu et al., 2014; Kandemir and Hamprecht, 2015). Owing to the gigapixel scale of WSIs, a significant challenge arises: in a positive WSI, the actual positive tissue often consists of only a minuscule fraction of the entire image (Wang et al., 2022). This imbalance leads to the generation of numerous trivial patches, on which the model might be overfitted, thereby degrading the model performance. To mitigate the model overfitting, one popular strategy, called attention-based MIL, which is to embed attention mechanisms into various MIL frameworks so as to mine the significant patches, has already been proven effective (Li et al., 2021; Lu et al., 2021; Zhang et al., 2022a). However, existing attention-based MIL methods primarily focus on exploring the significance of instances, and fail to take into account the relevant information among instances. To explore the instance relations, graph convolutional networks (GCNs) have garnered attention in pathology image analysis (Chen et al., 2021a; Chan et al., 2023). However, previous attention-based and GCN-based MIL methods majorly rely on unsupervised learning methods (e.g., contrastive learning (Li et al., 2021), clustering (Chen et al., 2022a)) or a pre-trained model on ImageNet (Russakovsky et al., 2015) for feature extraction, which might be susceptible to noise or lose some semantic information, thereby potentially resulting in sub-optimal performance.

Additionally, previous MIL-based methods mainly focus on classification tasks, but in clinical practice, it is very significant and urgent to retrieve and visualize similar images. To handle retrieval tasks, existing content-based WSI retrieval systems can be broadly classified into two categories: (i) WSI-based retrieval, (ii) Instance-based retrieval. For whole slide image retrieval, most existing methods obtain patch features using unsupervised learning methods or the pre-trained model based on ImageNet (Russakovsky et al., 2015), and then employ the aggregation algorithm to derive the features of WSIs, next, the nearest neighbor algorithm is utilized for effective retrieval (Kalra et al., 2020; Chen et al., 2021b; Wang et al., 2023). Similar to aforementioned MIL-based classification methods, these retrieval methods are also easily susceptible to noise or lose some semantic information, because they cannot directly extract significant semantic features from WSIs. For instance-based retrieval, the popular methods mainly divide each WSI into patches, and subsequently encode patches into a set of binary codes by hash methods, so as to rapidly retrieve these segmented patches with low storage costs (Ma et al., 2016; Shi et al., 2017; Zheng et al., 2017; Ma et al., 2018). However, these methods require to elaborate annotation information for each WSI, thereby consuming high time and economic costs. Therefore, it is very necessary to directly extract the features of significant patches in WSIs by only utilizing the labels of WSIs for WSI- and instance-level image retrieval. Moreover, most existing CBIR methods focus on retrieving similar patches within the same class among different WSIs, instead of considering their relevance or priorly retrieving the similar ones within the same WSI, which is one common demand in clinic (Shi et al., 2018).

Motivated by the aforementioned observations, in this paper, we propose a novel end-to-end **D**ynamic **G**raph based **W**eakly **S**upervised **D**eep **H**ashing framework, namely **DG-WSDH**, to directly and simultaneously classify and retrieve pathological images using only the slide-level label. For clarity, we present the structure pass of the proposed framework in Fig. 2. Specifically, this framework is composed of three principal components: (1) Multi-scale representation attention based deep network; (2) Patch-based dynamic graph; (3) Hashing encode. Specifically, to directly extract significant features from WSIs by only using WSI labels, we first adopt the multi-scale representation attention based deep network, namely MRAN (Xiang et al., 2023), as the backbone of the proposed framework, so that it can directly extract patch-level features from WSIs and meanwhile simultaneously explore the significant cell-, patch- and bag-level image features. Then, to leverage the relations among patches, we propose a novel dynamic graph based construction method to construct dynamic graphs by using patches within each bag. Next, to obtain latent binary codes with considering the relevance among patches and the similarity among WSIs, we embed the hash layer into the second and third stages, respectively, and design a novel pairwise ranking loss with considering the attention weight of patch-level images for model training.

In summary, our main contributions are listed as follows.

- We propose a novel end-to-end MIL-based deep hashing framework, namely DG-WSDH. To the best of our knowledge, this framework is the first work, which can utilize slide-level labels to directly extract significant patch features from gigapixel WSIs for handling both classification and retrieval tasks.
- We propose a novel dynamic graph construction method to model the relations among patches within each bag, so as to better explore their semantic information.
- We design a novel pairwise ranking loss to better learn latent binary codes of patch-level images, and embed it into the final loss function for model training to handle both classification and retrieval tasks.
- Extensive experimental results demonstrate the superior performance of the proposed framework over recent state-of-the-art methods on classification and various retrieval tasks, e.g., WSI retrieval and patch retrieval with considering the class and relevance information.

The rest of this paper is organized as follows. Section 2 briefly reviews some related and popular methods for classification and retrieval tasks. Section 3 introduces the proposed method. 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 related classification and retrieval methods for pathology image analysis, including attention-based deep MIL methods, GCN-based methods, and deep hashing methods.

Attention-based deep MIL methods: Compared to traditional MIL-based methods (Campanella et al., 2019; Chikontwe et al., 2020), attention-based ones have been proposed to measure the importance of each instance in one bag. This allows the model to selectively focus on significant instances while ignoring trivial ones. For example, the attention-based MIL method (Ilse et al., 2018) introduces two distinct attention mechanisms, and directly integrates them into deep multi-instance learning frameworks, which have shown effectiveness across multiple tasks. Additionally, to extract more powerful features with the unsupervised learning manner, DSMIL (Li et al., 2021) proposes a novel two-stream architecture based on self-supervised contrastive learning, focusing on analyzing the relationship between instances via trainable distance vectors. CS-MIL (Deng et al., 2024) proposes a new attention-based "early fusion" paradigm, which aims to capture multi-scale information and exploit inter-scale relationships in a holistic manner. Inspired by transformer (Vaswani et al., 2017), SA-MIL (Liu et al., 2020) introduces a novel self-attention based method to enhance learning processes across multiple instances. And TransMIL (Shao et al., 2021) proposes a transformer-based framework, specifically designed for the exploration of morphological and spatial information among instances. Additionally, HIPT (Chen et al., 2022) proposes a new vit architecture, which exploits the hierarchical structure of WSI and utilizes two levels of self-supervised learning to learn high-resolution image representations. Moreover, to leverage more feature information, a dual-attention multi-instance method (Zhu et al., 2021) is proposed to effectively optimize the learning of both global and local features, and then DTFD-MIL (Zhu et al., 2021), which introduces the concept of pseudo-packets and constructs a dual-layer multi-instance framework, is designed for more effective feature learning of instances (Zhang et al., 2022a). Despite above methods obtain promising performance on various types of WSIs, they usually utilize the pre-trained model or the unsupervised learning manner to learn patch-level features, thereby limiting the overall accuracy and efficiency of WSI analysis. Additionally, compared to HIPI (Chen et al., 2022), which requires hierarchical pre-training and utilizes the designed vit architecture to learn high-resolution image representations, our framework is directly trained on the original WSIs without hierarchical pre-training and employs GCN to leverage the patch-level relations. Compared to CS-MIL (Deng et al., 2024), which focuses on fusing the multi-scale information of pathology images for classification, we focus more on the relationships between different regions to obtain better feature representations for subsequent classification as well as sub-region retrieval tasks. To directly extract significant features from WSIs, MRAN (Xiang et al., 2023) proposes a multi-scale representation attention based deep MIL framework, which can directly extract patch-level image features from WSIs and simultaneously learn the significance of cell-, patch- and bag-level images. However, MRAN is designed only for WSI classification, and cannot be directly applied to retrieval tasks. Additionally, MRAN focuses on exploring the significance of patches without considering their relations, which might be beneficial to learning better bag-level features.

GCN-based methods: Compared to traditional CNNs, GCNs exhibit a distinct advantage in leveraging the structural information of data. Generally, for MIL-based analysis of pathological tissue images, each instance is considered as a node in the graph, and edges are constructed based on positional information or features among instances. Specifically, graph convolutional network (Kipf and Welling, 2016) enables fast convolution on graphs. To mitigate over-smoothing in GCNs, GCNII (Chen et al., 2020) employs initial residual and identity mapping. To reduce the complexity of constructing graphs, MUSTANG (Gallagher-Syed et al., 2023) presents a model to build sparse graphs based on K-Nearest-Neighbour (KNN) for patches, so as to achieve better unsupervised classification performance. Additionally, the work (Adnan et al., 2020) extracts patch features by using the pre-trained model, and then computes the adjacency matrix by learning the global context of patches to represent the relationship among patches. With the prominence of transformer methods in other areas, Graph-Transformer (Zheng et al., 2022a) constructs a graph from the positional information of patches in the WSI, and then utilizes the transformer mechanism to predict the labels of WSIs. Furthering this domain, DMCAH (Zheng et al., 2022b) is presented by constructing a location-aware graph from the location information of sub-regions and encoding the graph for retrieval through graph convolution and self-attention mechanisms. Aforementioned methods have achieved impressive results, however, they often utilize the location information of instances or obtain instance features based on pre-trained models using clustering for static graph construction, so they might lose some semantic information during feature extraction, possibly leading to the model with sub-optimal performance. Additionally, although graph attention network (GAT) (Veličković et al., 2017) can construct dynamic graphs by using the attention mechanism to aggregate features from neighboring nodes, for enhancing the feature representation of each node within the graph, its performance is often restricted by a large amount of trivial instances in WSIs.

Deep hashing methods: Deep hashing methods can maintain the similarity among instances by mapping instances to low-dimensional binary codes for obtaining efficient retrieval performance, and thus they are widely used in massive data retrieval tasks. Based on whether using semantic information to generate binary codes, deep hashing methods can be classified into unsupervised and supervised hashing. In unsupervised hashing, the generation of hash codes is accomplished without using labels. Some popular unsupervised deep hashing methods include DAH (Tang et al., 2022), which introduces autoencoders into deep hash learning to learn compact binary hash codes; DGAH (Dizaji et al., 2018), which presents a framework using contrast methods to generate hash codes; and NRDH (Wang et al., 2021), which takes into account the relationship among instances and employs unsupervised graph convolution to generate hash codes. For supervised deep hashing methods, some popular ones are: PDRH (Shi et al., 2018), which proposes a pairwise supervised hashing framework to not only maintain the similarity among patches but also preserve their ranking order; MTH (Chen et al., 2023), which is a multi-scale triplet hashing method to learn different scales of information from medical images; DenseHashNet (Liu et al., 2022), which presents a DenseNet-based deep hashing algorithm for retrieval of large-scale medical datasets; and DMCAH (Zhang et al., 2022b), which introduces a novel medical cross-modal attention hashing algorithm, so as to learn better global and local features for hash code generation. Although these hashing methods have achieved impressive results in many retrieval tasks, the inherent characteristics of WSIs make them difficult to be directly applied to WSIs.

In contrast to most previous attention-based MIL classification methods, which utilize unsupervised learning methods to learn patch features of each WSI, the proposed framework can directly learn patch features from WSIs by using slide-level labels. Although MRAN can also directly extract patch features, it cannot be directly applied to retrieval tasks and fail to consider the relations among patches, thereby possibly degrading its model performance. Compared to previous GCN-based methods constructing graph to leverage structural information among instances, the proposed framework constructs the dynamic graph by only using the slide-level label information so as to further explore the semantic information. Additionally, given the inherent characteristics of WSIs, the patch-level images inevitably contain a large amount of

noise or trivial ones. Although GAT effectively learns features from neighboring nodes through the attention mechanism, the presence of noise can potentially degrading model performance. To address this challenge, our proposed framework mitigates the effect of noise by constructing a sparse dynamic graph. Moreover, different from most previous unsupervised or supervised deep hashing methods, the proposed framework is a weakly supervised deep hashing method, and can simultaneously preserve the semantic similarity for patches and WSIs, and meanwhile maintain the relevance among patches, i.e., for a query patch, its similar patches from the same WSI should have larger relevance than those from a different WSI.

3. Method

3.1. WSI preprocessing

In this section, we concisely outline the preprocessing procedure for WSIs. Given N WSIs, denoted as $\{\mathbf{X}_n\}_{n=1}^{N}$, where $\mathbf{X}_n \in \mathbb{R}^{C\times H\times W}$ is the nth WSI, and C, H and W denote the number of channels, width and height of the WSI (40x magnification), respectively. And we resize the size of WSIs to $\frac{H}{2}\times\frac{W}{2}$ in order to reduce the computational and memory costs of model training. Firstly, we utilize the global binary thresholding algorithm (Malathy et al., 2016) and GLCM (Haralick et al., 1973) to identify the tissue region of WSIs. Secondly, each tissue region in one WSI is segmented into a series of non-overlapping bag-level images $\{\mathbf{X}_{ni}\}_{i=1}^{N_b}$, where $\mathbf{X}_{ni} \in \mathbb{R}^{C\times H_b\times W_b}$ is the ith bag-level image in $\mathbf{X}_n$, and C, H_b and W_b denote the number of channels, width, and height of the bag-level image, respectively. Additionally, each bag-level image is divided into a set of patches $\{\mathbf{X}_{nij}\}_{j=1}^{N_p}$, where $\mathbf{X}_{nij} \in \mathbb{R}^{C\times H_p\times W_p}$ is the jth patch in $\mathbf{X}_{ni}$, and C, H_p and W_p represent the number of channels, width and height of the patch-level image, respectively. Additionally, we regard that each patch-level image contains a set of cell-level images $\{\mathbf{X}_{nijk}\}_{i=k}^{N_c}$, where $\mathbf{X}_{nijk} \in \mathbb{R}^{C\times H_c\times W_c}$ is the kth cell-level image in $\mathbf{X}_{nij}$, and C, H_c and W_c represent the number of channels, width, and height of the cell-level image, respectively. Similar to MRAN (Xiang et al., 2023), in the proposed framework, we empirically set $H_b = W_b = 1024$, $H_p = W_p = 128$ and $H_c = W_c = 32$.

3.2. Overview of the proposed framework

We present the overview structure of the proposed framework in Fig. 2, which consists of four stages: (i) WSI preprocessing, which aims to preprocess each WSI to obtain patch-level images in the offline phase. After preprocessing WSIs, patch-level images are fed into the second stage, i.e., (ii) Feature extraction and hash codes generation, which consists of a backbone network composed of $L-1$ layers with the parameters $\{\theta_l\}_{l=1}^{L-1}$ to learn cell-level representations, the cell- and patch-level attention to learn cell and patch attention weights ($\{\alpha_{nijk}\}_{k=1}^{N_c}$ and $\{\rho_{nij}\}_{j=1}^{N_p}$) for obtaining patch- and bag-level representations ($\left\{\mathbf{Z}_{nij}^{p}\right\}_{j=1}^{N_p}$ and $\{\mathbf{Z}_{ni}^{b}\}_{i=1}^{N_b}$), respectively, the GCN module to model the relations among patches within each bag to better explore their semantic information and generate the patch embeddings ($\left\{\mathbf{Z}_{nij}^{g_2}\right\}_{j=1}^{N_p}$), a hash layer ($\mathbf{H}_p$) to generate latent binary codes of patch-level images, and a fully-connected layer (FC_d) with the parameters θ^L for cell- and bag-level image classification. After obtaining bag-level feature representations, they will be fed into the third stage, i.e., (iii) WSI classification and hash code generation, which consists of two fully connected layers (FC_b and FC_w) with parameters $\{\theta_l\}_{l=L+1}^{L+2}$ and a bag-level attention to learn attention weights ($\{\kappa_{ni}\}_{i=1}^{N_b}$) of bag-level images and WSI-level representations ($\mathbf{Z}_n^w$) for WSI classification, a hash layer ($\mathbf{H}_w$) to generate latent binary codes of each WSI. Stages (ii) and (iii) are in the training phase. After the model being well trained, we show the process to retrieve the similar images from the database, which contains a large amount of hash codes, for the query patch- and WSI-level images in the fourth stage, i.e., (iv) Image retrieval.

3.3. Multi-scale attention

After WSI preprocessing, each WSI is composed of N_b bags. We assume that each bag has the same label as its corresponding WSI. This will make some bags have wrong labels. Additionally, within each bag, there are numerous trivial patches, and patch-level images often contain cell-level noise information. To effectively mitigate the effect of bag-level wrong labels, cell- and patch-level noise information, same as Xiang et al. (2023), we embed a multi-scale attention mechanism into the learning process, where the attention mechanism is based on Shi et al. (2020). For clarity, we will present the multi-scale attention in the following.

3.3.1. Cell-attention

After obtaining patch-level images $\{\mathbf{X}_{nij}\}_{j=1}^{N_p}$, feeding them into the backbone network, it can obtain cell-level image representations $\left\{\mathbf{Z}_{nijk}^{c}\right\}_{k=1}^{N_c}$. Then, using cell-level attention to obtain patch-level representations, it is:

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

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

$$\alpha_{nijk} = \frac{max(\alpha_{nijk} - \frac{\lambda_c}{N_c}, 0)}{\sum_{t=1}^{N_c} max(\alpha_{nijt} - \frac{\lambda_c}{N_c}, 0)}, \tag{1c}$$

$$\mathbf{Z}_{nij}^{p} = \sum_{k=1}^{N_c} \boldsymbol{\alpha}_{nijk}\mathbf{Z}_{nijk}^{c}, \tag{1d}$$

where $f_b(\cdot)$ denotes the fully connected layer FC_d with the parameters θ^L in the second stage for bag-level image classification, $\mathbf{P}_{nijk}$ is the logit vector of the cell-level image $\mathbf{X}_{nijk}$, R is the number of classes, α_{nijk} is the attention weight of $\mathbf{X}_{nijk}$, and $\mathbf{Z}_{nij}^{p}$ is the feature representation of the patch-level image $\mathbf{X}_{nij}$. Additionally, $\lambda_c \in [0,1]$ is a threshold to remove trivial cell-level images from each patch.

3.3.2. Patch-attention

We first feed the patch-level image feature representations $\left\{\mathbf{Z}_{nij}^{p}\right\}_{j=1}^{N_p}$ into the GCN module to obtain the feature representations $\left\{\mathbf{Z}_{nij}^{g_2}\right\}_{j=1}^{N_p}$, and then feed them into the patch-level attention, which is:

$$\rho_{nij} = \frac{\sqrt{\sum_{r=1}^{R}(\mathbf{Z}_{nijr}^{g_2})^2}}{\sum_{t=1}^{N_p}\sqrt{\sum_{r=1}^{R}(\mathbf{Z}_{nitr}^{g_2})^2}}, \tag{2a}$$

$$\rho_{nij} = \frac{max(\rho_{nij} - \frac{\lambda_p}{N_p}, 0)}{\sum_{t=1}^{N_p} max(\rho_{nit} - \frac{\lambda_p}{N_p}, 0)}, \tag{2b}$$

$$\mathbf{Z}_{ni}^{b} = \sum_{j=1}^{N_p} \rho_{nij}\mathbf{Z}_{nij}^{g_2}, \tag{2c}$$

where ρ_{nij} is the attention weight of $\mathbf{X}_{nij}$, $\lambda_p \in [0,1]$ is a threshold to remove trivial patch-level images from each bag, and $\mathbf{Z}_{ni}^{b}$ is the feature representation of the bag-level image $\mathbf{X}_{ni}$. Note that, different from the patch-level attention in MRAN (Xiang et al., 2023), which utilizes $f_b(\mathbf{Z}_{nij}^{p})$ as the input of Eq. (2a), we employ $\mathbf{Z}_{nij}^{g_2}$ obtained from the GCN module as the input of Eq. (2a).

3.3.3. Bag-attention

Based on the patch-level attention, we can obtain bag-level feature representations $\{\mathbf{Z}_{ni}^{b}\}_{i=1}^{N_b}$, and then feed them into the fully connected layer FC_b in the third stage to obtain bag-level embeddings, which are represented by $\{\mathbf{Z}_{ni}^{e}\}_{i=1}^{N_b}$, next, feed them the into the bag-level attention, which is:

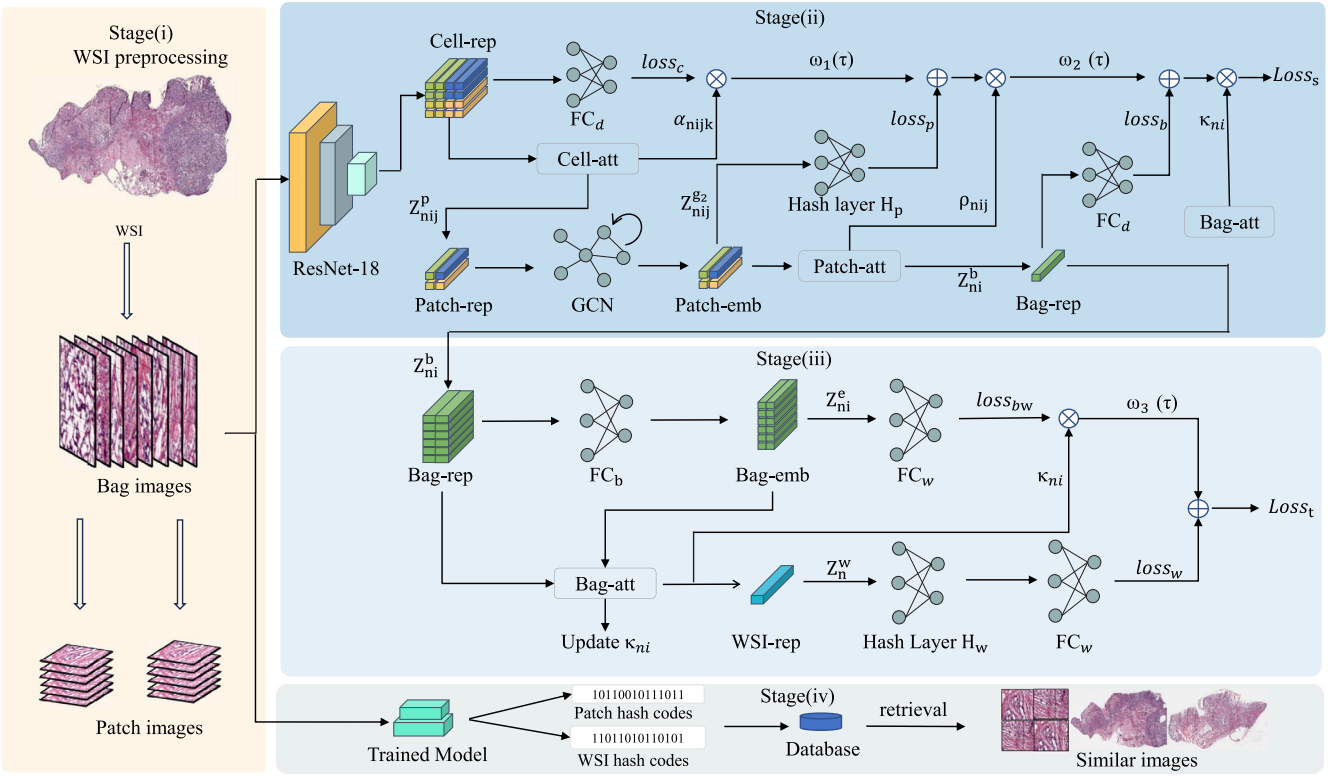


Fig. 2. The overall structure of the proposed framework, DG-WSDH.

$$\mathbf{P}_{ni} = f_w(\mathbf{Z}^e_{ni}), \tag{3a}$$

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

$$\kappa_{ni} = \frac{max(\kappa_{ni} - \frac{\lambda_b}{N_b}, 0)}{\sum_{t=1}^{N_b} max(\kappa_{nt} - \frac{\lambda_b}{N_b}, 0)}, \tag{3c}$$

$$\mathbf{Z}^w_n = \sum_{i=1}^{N_b} \kappa_{ni}\mathbf{Z}^e_{ni}, \tag{3d}$$

where $\mathbf{P}_{ni}$ denotes the logit vector of $\mathbf{X}_{ni}$, $f_w(\cdot)$ denotes the fully connected layer FC$_w$ in the third stage for WSI classification, κ_{ni} is the attention weight of $\mathbf{X}_{ni}$, $\lambda_b \in [0, 1]$ is a threshold to remove trivial bag-level images from each WSI, and $\mathbf{Z}^w_n$ is the feature representation of the WSI $\mathbf{X}_n$.

3.4. Dynamic graph construction

In this subsection, we will present the construction of dynamic graphs in the proposed framework. This is because many literature (Adnan et al., 2020; Zheng et al., 2022a,b) have demonstrated that building the graph to leverage the relationship among patches in WSIs can boost the model performance, where the graph utilizes the adjacency matrix to describe the relations of graph nodes and contribute crucially to the message passing within the graph. Unfortunately, previous GCN-based methods usually adopt the pre-defined or static graph, which might lose some semantic information during feature extraction. To overcome this issue, similar to GAT (Veličković et al., 2017), we utilize the attention mechanism to dynamically construct the adjacency matrix. However, due to the characteristics of pathological tissue images, i.e., each bag might contain a large number of trivial patches, which often degrade the model performance, we construct a sparse adjacency matrix to mine the critical patches and improve the model generalization capability. For clarity, we show the defined adjacency matrix $\mathbf{A}$ as follows:

$$\mathbf{Z}^q_{nij} = LeakyReLU(\mathbf{Z}^p_{nij}\mathbf{W}_p), \tag{4a}$$

$$\mathbf{Q}_{nij} = \mathbf{Z}^q_{nij}\left[\mathbf{Z}^q_{ni1}\mathbf{Z}^q_{ni2}\cdots\mathbf{Z}^q_{niN_p}\right]^T, \tag{4b}$$

$$\xi = \frac{1}{N_p^2}\sum_{j=1}^{N_p}\sum_{t=1}^{N_p}\mathbf{Q}_{nijt}, \tag{4c}$$

$$\mathbf{A}_{nij} = \frac{max(\mathbf{Q}_{nij} - \gamma \times \xi, 0)}{\sqrt{\sum_{t=1}^{N_p} max(\mathbf{Q}_{nijt} - \gamma \times \xi, 0)^2}}, \tag{4d}$$

$$\mathbf{A}_{ni} = \left(\begin{bmatrix}\mathbf{A}_{ni1}\\ \mathbf{A}_{ni2}\\ \vdots\\ \mathbf{A}_{niN_p}\end{bmatrix} + \begin{bmatrix}\mathbf{A}_{ni1}\\ \mathbf{A}_{ni2}\\ \vdots\\ \mathbf{A}_{niN_p}\end{bmatrix}^T\right)/2, \tag{4e}$$

where $\left\{\mathbf{Z}^p_{nij}\right\}_{j=1}^{N_p}$ denote the patch-level feature representations obtained by the cell-attention, and they are from one bag in each batch, $\mathbf{Z}^p_{nij} \in \mathbb{R}^{1\times M}$ means that the dimension of patch-level feature representation is M, $\mathbf{W}_p \in \mathbb{R}^{M\times D_p}$ denotes a trainable parameter matrix to project the high-dimensional features into a low-dimensional space, we empirically set $D_p = 128$. Additionally, $\left\{\mathbf{Q}_{nij}\right\}_{j=1}^{N_p}$ denotes the degree of correlation among $\mathbf{Z}^p_{nij}$ and the other patch features within one bag, where $\mathbf{Q}_{nij} \in \mathbb{R}^{1\times N_p}$. To mitigate the effect of noise, we employ an attention mechanism to obtain a sparse adjacency vector $\mathbf{A}_{nij} \in \mathbb{R}^{1\times N_p}$ for the patch-level image $\mathbf{X}_{nij}$, ξ is a scalar and γ is a threshold parameter designed to regulate the edge ratio between patches within each bag. Moreover, we empirically initialize the parameters of $\mathbf{W}_p$ with a uniform value of one and set $\gamma = 1$. Then, similar to Kipf and Welling (2016), we apply two graph convolutional layers to encode the relations among different patches. Specifically, the graph convolution process for the two layers are formulated as follows:

$$\mathbf{Z}^{g_1}_{nij} = \mathbf{D}^{-\frac{1}{2}}\mathbf{A}_{ni}\mathbf{D}^{-\frac{1}{2}}\mathbf{Z}^p_{nij}\mathbf{W}_{g_1}, \tag{5}$$

$$\mathbf{Z}^{g_2}_{nij} = \mathbf{D}^{-\frac{1}{2}}\mathbf{A}_{ni}\mathbf{D}^{-\frac{1}{2}}\mathbf{Z}^{g_1}_{nij}\mathbf{W}_{g_2}, \tag{6}$$

where $\mathbf{D}$ is a diagonal matrix with its jth diagonal element $d_j = \sum_t a_{jt}$, $a_{jt} \in \mathbf{A}_{ni}$, $\mathbf{W}_{g_1}$ and $\mathbf{W}_{g_2}$ denote trainable parameters of the two layers. Additionally, we provide examples of the learned graphs in the supplementary material.

3.5. Pairwise ranking hashing

Given patch-level images $\{\mathbf{X}_{nij}\}_{j=1}^{N_p}$, assume that they are categorized into R classes. Note that patch-level images are assigned to the same labels as its corresponding WSI. We define the relationship between patches $\mathbf{X}_{nij}$ and $\mathbf{X}_{nik}$ as:

$$\mathbf{S}_{jk} = \begin{cases} r(\mathbf{X}_{nij}, \mathbf{X}_{nik}), & (\mathbf{X}_{nij}, \mathbf{X}_{nik}) \in \mathcal{G}, \\ -\lambda_s, & (\mathbf{X}_{nij}, \mathbf{X}_{nik}) \in \mathcal{V}, \end{cases} \tag{7}$$

where $\mathcal{G}$ denotes the set containing similar pairs within the same class, $\mathcal{V}$ is the set with dissimilar pairs belonging to different classes. Additionally, $r(\mathbf{X}_{nij}, \mathbf{X}_{nik})$ represents the relevance between the patches $\mathbf{X}_{nij}$ and $\mathbf{X}_{nik}$. The relevance is to measure the degree of similarity among patches. For example, suppose that $\mathbf{X}_{nij}$, $\mathbf{X}_{nik}$ and $\mathbf{X}_{nit}$ are within the same class, but $\mathbf{X}_{nij}$ and $\mathbf{X}_{nik}$ are from the same WSI, $\mathbf{X}_{nij}$ and $\mathbf{X}_{nit}$ are from two different WSIs, then it has $r(\mathbf{X}_{nij}, \mathbf{X}_{nik}) > r(\mathbf{X}_{nij}, \mathbf{X}_{nit})$. In this paper, we set $r(\mathbf{X}_{nii}, \mathbf{X}_{nij}) \in (0, m]$, where m is the length of hash codes. More specifically, we empirically set $\lambda_s = 1$, when $(\mathbf{X}_{nii}, \mathbf{X}_{nij}) \in \mathcal{G}$, we set $r(\mathbf{X}_{nii}, \mathbf{X}_{nij}) = 6$ if they are from different WSIs, and set $r(\mathbf{X}_{nii}, \mathbf{X}_{nij}) = 8$ if they are from the same WSI. Then, we introduce the hash layer to encode high-dimensional data into latent binary codes. Similar to Shi et al. (2018), the used hash layer is represented as follows:

$$\mathbf{H}_{nij} = \tanh(\mathbf{Z}_{nij}^{g_2}\mathbf{W}_h + \mathbf{b}_h), \tag{8}$$

where $\mathbf{Z}_{nij}^{g_2}$ obtained from the GCN module denotes the patch-level feature representation of $\mathbf{X}_{nij}$, $\mathbf{W}_h$ is a trainable parameter matrix, $\mathbf{b}_h$ denotes a trainable bias vector, $\mathbf{H}_{nij} \in [-1, 1]^m$ denotes the latent hash codes of $\mathbf{X}_{nij}$, and m is the length of hash codes. Note that because the function $sgn(\cdot)$ is non-differential, it is often to replace $sgn(\cdot)$ with the differentiable function $tanh(\cdot)$ during the model training. After obtaining the latent hash codes, we will update $\mathbf{H}_{nij} = sgn(\mathbf{H}_{nij}) \in \{-1, 1\}^m$ and store them in a database for querying.

For patch-level images, a ranking loss is employed for effective patch retrieval. The primary objective of hash functions is to encode high-dimensional data into compact hash codes while preserving the similarity among the data. This preservation of similarity relationships is vital for efficient patch retrieval. Here, similar to Shi et al. (2018), based on Eqs. (7) and (8), the objective function can be defined as:

$$loss_p = \min_{\mathbf{H}} \lambda_h \left\| \frac{r_{max}}{m}\mathbf{H}\mathbf{H}^T - \mathbf{S} \right\|_F^2, \tag{9}$$

where $\mathbf{H} \in [-1, 1]^{N_h \times m}$ is a matrix composed by the latent hash codes of patch-level images, N_h denotes the number of patches within each batch in the second stage, $\mathbf{S} \in \mathbb{R}^{N_h \times N_h}$ is a matrix to represent their relations, r_{max} is the largest element in $\mathbf{S}$, and λ_h is a regularization hyperparameter.

However, because each bag contains trivial patches and patch-level attention can significantly reduce the effect of the trivial ones. Hence, in this paper, we define the ranking loss for patch-level image retrieval as:

$$loss_p = \min_{\mathbf{H}} \lambda_h \left\| \rho_{nij} \left(\frac{r_{max}}{m}\mathbf{H}_{nij}\mathbf{H}^T - \mathbf{S}_j \right) \right\|_F^2, \tag{10}$$

where ρ_{nij} is the attention weight for patch-level image $\mathbf{X}_{nij}$, $\mathbf{S}_j \in \mathbb{R}^{1 \times N_h}$ is a row vector and $\mathbf{S}_j \subset \mathbf{S}$.

3.6. Loss function

The proposed framework aims to handle classification and retrieval tasks. Thus, its objective loss function consists of two major parts: a cross-entropy loss function and a patch-based ranking loss function for classification and retrieval, respectively. Additionally, because the second and third stages have different goals, they have distinct objective functions. For clarity, we show them in the following.

3.6.1. Loss in the second stage

Given a set of cell-level images $\mathbf{X}_{nijk \in S}$ and their labels $\mathbf{y}_{nijk \in S}$, where $\mathbf{X}_{nijk} \in \mathbb{R}^{C \times H_c \times W_c}$ and $\mathbf{y}_{nijk} = \{y_{nijkr}\}_{r=1}^{R}$ is a one-hot label vector with $y_{nijkr} \in \{0, 1\}$, the cross-entropy loss function for cell-level image classification is defined as:

$$loss_c = -\sum_{r=1}^{R} \mathbf{y}_{nijkr} log\Big(s\left(\mathbf{P}_{nijk}\right)[r] \Big), \tag{11}$$

where $s(\cdot)$ is the softmax function, and $\mathbf{P}_{nijk}$ represents the logit vector of $\mathbf{X}_{nijk}$. Because $\mathbf{X}_{nijk}$ is assumed to be with the same label as its corresponding WSI, which is single-label, it has $\mathbf{y}_{nijkr} \in \{0, 1\}$ and $\sum_{r=1}^{R} y_{nijkr} = 1$. Thus, when $\mathbf{X}_{nijk}$ belongs to the rth class, it has $\mathbf{y}_{nijkr} = 1$ and $\sum_{t=1, t \neq r}^{R} y_{nijkt} = 0$, and then Eq. (11) equals:

$$loss_c = -log\Big(s\left(\mathbf{P}_{nijk}\right)[r] \Big). \tag{12}$$

Similar to Eq. (12), the cross-entropy loss function for bag-level image prediction in the second stage is defined as:

$$loss_b = -log\Big(s\left(\mathbf{P}_{ni}^b\right)[r] \Big), \tag{13}$$

where $\mathbf{P}_{ni}^b = f_b(\mathbf{Z}_{ni}^b)$ represents the logit vector of the bag-level image $\mathbf{X}_{ni}$ in the second stage.

Based on previous work (Shi et al., 2020) and similar to Xiang et al. (2023), we connect the cell- and patch-level attention with the loss function for cell-image classification, so as to boost the performance on mining significant cell-level images, and then combine them with the loss functions for patch-level image retrieval and bag-level image classification. Thus, based on Eq. (10), Eq. (12) and (13), the loss function for the second stage is defined as:

$$loss_s = \frac{1}{|S|} \sum_{\mathbf{X}_{ni} \in S} \kappa_{ni}\beta_r \Bigg(loss_b + \omega_1(\tau) \sum_{j=1}^{N_p} \Big(loss_p + \rho_{nij}\omega_2(\tau) \sum_{k=1}^{N_c} \alpha_{nijk} loss_c \Big) \Bigg), \tag{14}$$

where α_{nijk}, ρ_{nij} and κ_{ni} denote the weights obtained by cell-, patch- and bag-level attention for the $\mathbf{X}_{nijk}$, $\mathbf{X}_{nij}$ and $\mathbf{X}_{ni}$ images, respectively. S denotes a set containing bags and $|S|$ represents the number of bags. In Eq. (14), $loss_b$ is its main objective loss for bag-level image classification, while $loss_p$ and $\sum_{k=1}^{N_c} \alpha_{nijk} loss_c$ are regularization terms to connect the patch- and cell-level attention with patch retrieval and cell image prediction, respectively. The function $\omega_1(\tau)$ and $\omega_2(\tau)$ are unsupervised weight functions used to balance the three terms, and τ is the number of current training epochs. It is worth noting that the proposed model employs only slide-level labels, and thus $\beta_r = \frac{\frac{N}{N_r}}{\sum_{i=1}^{R} \frac{N}{N_i}}$ is the weight allocated to the rth class in WSIs, so as to alleviate the issue of class imbalance, and $N = \sum_{r=1}^{R} N_r$ is the total number of WSIs, N_r is the number of WSIs belonging to the rth class.

3.6.2. Loss in the third stage

Similar to the second stage, for the third stage, we connect the bag-level attention with the loss function for bag-level image classification, and then combine it with the loss function for WSI classification and retrieval. Similar to Eq. (12), the loss function for bag- and WSI-level images prediction in the third stage is defined as:

$$loss_t = \frac{1}{|\mathcal{B}|} \sum_{\mathbf{X}_n \in \mathcal{B}} \beta_r \Bigg(loss_w + \omega_3(\tau) \sum_{i=1}^{N_b} \kappa_{ni} loss_{bw} \Bigg), \tag{15}$$

where $\mathcal{B}$ denotes a set containing WSIs and $|\mathcal{B}|$ is the number of WSIs, $\omega_3(\tau)$ is an unsupervised weight function used to balance the two terms, $loss_{bw} = -log\Big(s(\mathbf{P}_{ni}^e)[r] \Big)$ and $loss_w = -log\Big(s(\mathbf{P}_n)[r] \Big)$, where $loss_w$ is the main objective loss for WSI classification and $\sum_{i=1}^{N_b} \kappa_{ni} loss_{bw}$ is a regularization term to connect the bag-level image attention and

Algorithm 1 DG-WSDH

Input: Training data $\{\mathbf{X}_n\}_{n=1}^{N}$, pairwise matrix $\mathbf{S}$, label vectors $\{\mathbf{y}_n\}_{n=1}^{N}$, weight functions $\{\omega_i(\tau)\}_{i=1}^{3}$, network parameters θ, augmentation function $g(\cdot)$, the number of training epochs T.

Output: Network parameters θ, GCN parameters W_{g_1} and W_{g_2}, hash layer parameters W_h, b_h, W_w, b_w.

Preprocess $\mathbf{X}_n$ to obtain $\{X_{ni}\}_{i=1}^{N_b}$, for $\forall 1 \le n \le N$
Initialize $\boldsymbol{\kappa}_{ni} = 1$ for $\forall 1 \le n \le N$ and for $\forall 1 \le i \le N_b$
for $\tau \in [1, T]$ **do**
  $\upsilon \leftarrow 0$;
  for *each minibatch in* $\mathcal{S}$ **do**
    Cut $\mathbf{X}_{ni}$ to obtain $\{\mathbf{X}_{nij}\}_{j=1}^{N_p}$ for $\forall ni \in \mathcal{S}$.
    $\mathbf{Z}^c_{nijk} \leftarrow f_\theta(g(\mathbf{X}_{nij}))$, for $\forall ni \in \mathcal{S}$ and for $\forall 1 \le j \le N_p$
    $\mathbf{Z}^p_{nij}, \{\alpha_{nijk}\}_{k=1}^{N_c} \leftarrow CellAtt\left(\{\mathbf{Z}^c_{nijk}\}_{k=1}^{N_c}, \theta^L\right)$
    // Cell-attention, where $CellAtt(\cdot)$ is Eq. (1a)–(1d)
    $\mathbf{A}_{ni} \leftarrow Graph\left(\{\mathbf{Z}^p_{nij}\}_{j=1}^{N_p}\right)$
    // Graph construction, where $Graph(\cdot)$ is Eq. (4a)–(4e)
    $\{\mathbf{Z}^{g_2}_{nij}\}_{j=1}^{N_p} \leftarrow GCN\left(\{\mathbf{Z}^p_{nij}\}_{j=1}^{N_p}, \mathbf{A}_{ni}\right)$
    // The GCN module, where GCN consists of two layers, which are represented by Eqs. (5)–(6)
    $\mathbf{H}_{nij} \leftarrow Hash(\mathbf{Z}^{g_2}_{nij})$
    // Hash function, where $Hash(\cdot)$ denotes Eq. (8)
    $\mathbf{Z}^b_{ni}, \{\rho_{nij}\}_{j=1}^{N_p} \leftarrow PatchAtt\left(\{\mathbf{Z}^{g_2}_{nij}\}_{j=1}^{N_p}, \theta^L\right)$
    // Patch-attention, where $PatchAtt(\cdot)$ is Eq. (2a)–(2c)
    $loss_s \leftarrow$ Eq. (18)
    Update the parameters θ^l for $1 \le l \le L$, W_{g_1} and W_{g_2}, and W_h and b_h.
  end for
  $\upsilon \leftarrow 1$;
  for *each minibatch in* $\mathcal{B}$ **do**
    $\mathbf{Z}^w_n, \{\kappa_{ni}\}_{i=1}^{N_b} \leftarrow BagAtt\left(\{\mathbf{Z}^e_{ni}\}_{k=1}^{N_b}, \{\theta^L\}_{l=L+1}^{L+2}\right)$
    // Bag-attention, where $BagAtt(\cdot)$ is Eq. (3a)–(3d)
    $\mathbf{H}_n \leftarrow Hash(\mathbf{Z}^w_n)$
    // Hash function, where $Hash(\cdot)$ denotes Eq. (16)
    $loss_t \leftarrow$ Eq. (18)
    Update the parameters θ^l for $L+1 \le l \le L+2$, W_w and b_w
  end for
end for

prediction, so as to boost the precision and recall of significant bags. Additionally, based on Eq. (3a), it can have $\mathbf{P}^e_{ni} = f_w(\mathbf{Z}^e_{ni})$, which is the logit vector of $\mathbf{X}_{ni}$ in the third stage, and $f_w(\cdot)$ denotes the fully connected layer FC_w in the third stage. Moreover, in order to enable simultaneously retrieve and classify WSIs, $\mathbf{P}_n$ is obtained by the following equation:

$$\mathbf{H}_n = \tanh\left(\mathbf{Z}^w_n \mathbf{W}_w + \mathbf{b}_w\right), \tag{16}$$

$$\mathbf{P}_n = f_w\left(\mathbf{H}_n\right), \tag{17}$$

where $\mathbf{Z}^w_n$ is the feature representation of the WSI $\mathbf{X}_n$, $\mathbf{H}_n$ denotes its latent hash codes, $\mathbf{W}_w$ is a trainable projection matrix and $\mathbf{b}_w$ is a trainable bias vector. Specifically, in the third stage, we first feed the WSI representation $\mathbf{Z}^w_n$ achieved by the bag-level attention into the hash layer to obtain the latent hash codes $\mathbf{H}_n$ of $\mathbf{X}_n$, and then feed $\mathbf{H}_n$ into the fully connected layer FC_w to attain $\mathbf{P}_n$.

3.6.3. Overall loss

Because the second and third stages are alternatively conducted, based on their loss functions $loss_s$ and $loss_t$, the final objective function is:

$$loss = (1 - \upsilon) loss_s + \upsilon loss_t, \tag{18}$$

where $\upsilon \in \{0, 1\}$ is to control which branch is used to train the model during the training process. Specifically, when $\upsilon = 0$, update the model parameters in the second stage; when $\upsilon = 1$, update the model parameters in the third stage. For clarity, we show the detailed training procedure of the proposed framework in Algorithm 1.

4. Experiments

In this section, we introduce the used datasets and the detailed experimental settings in our experiments, and compare the proposed framework with state-of-the-art methods on classification and retrieval tasks, respectively.

4.1. Datasets

Stomach Adenocarcinoma (STAD): The dataset is from the The Cancer Genome Atlas Program (TCGA) (Tomczak et al., 2015) and it has 749 slides from 428 patients, including 626 positive slides and 123 negative slides. In our experiments, these patients are randomly divided into training, validation, and testing sets following a ratio of 3:1:1. This dataset is widely used for the WSI classification task.

CAMELYON16: This dataset (Bejnordi et al., 2017) contains 398 slides, consisting of 159 positive and 239 negative slides. In our experiments, these WSIs are randomly allocated into training, validation, and testing sets according to a ratio of 8:1:1. Additionally, this dataset provides pixel-level annotations for each slide, and thus it can be used for both WSI classification and patch retrieval tasks.

CAMELYON17: This dataset (Bandi et al., 2018) consists of 1,000 slides obtained from 200 patients across five distinct medical centers. However, only 500 of these slides are accompanied by slide-level annotation information, and a subset of 50 slides possesses pixel-level annotations. Consequently, we randomly divide the patients with slide-level annotations into training, validation and testing sets for WSI classification and retrieval, according to a ratio of 3:1:1. Additionally, for the patch retrieval task, we divide the slides with pixel-level annotations into training and testing sets using a ratio of 4:1, where the training set is only used to construct a database containing hash codes.

4.2. Comparison methods

In this subsection, we present the comparison methods used for classification and retrieval tasks, respectively.

4.2.1. Classification methods

MSKCC-MIL (Campanella et al., 2019): which is a weakly supervised two-stage classification method, i.e., it first trains the model with MIL on patch-level images and then selects confident patches in each WSI based on their prediction scores for WSI classification.

MSKCC-RNN (Campanella et al., 2019): which first employs the trained MSKCC-MIL model to extract patch-level image features and then classifies WSIs by using the RNN method.

DSMIL (Li et al., 2021): which first utilizes self-supervised contrast learning to extract features from patch-level images, and then models the relations of the instances with a trainable distance measurement and classifies WSIs with MIL.

CLAM-SB (Lu et al., 2021): which utilizes a pre-trained model of ResNet-18 on ImageNet (Russakovsky et al., 2015) to extract features from patches, and then, based on the attention scores obtained from

Gated-Attention (Ilse et al., 2018), employs a SVM-based loss to classify patches and single-branch attention to classify WSIs, respectively.

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

TransMIL (Shao et al., 2021): which adopts a pre-trained model of ResNet-18 on ImageNet (Russakovsky et al., 2015) to extract features from patch-level images, and then classifies WSIs by using the Transformer-based attention MIL method.

DTFD-MIL (Zhang et al., 2022a): which adopts the pre-trained model of ResNet-18 on ImageNet (Russakovsky et al., 2015) to extract features from patch-level images, and then employs the gradient-based calculation of instance probabilities and the attention-based MIL to classify WSIs.

MRAN (Xiang et al., 2023): which proposes an end-to-end deep MIL framework with a multi-scale representation attention mechanism, so as to directly extract patch-level features from WSIs and simultaneously mine the significant information at different scales for WSI classification.

4.2.2. Retrieval methods

In the following, we briefly introduce the comparison methods, FISH, RetCCL and HSHR, for WSI retrieval, and CDC and HSDH for patch-level image retrieval.

FISH (Chen et al., 2021b): which employs multiple clustering to select patches as well as a pre-trained model VQ-VAE on TCGA to extract features, and a Van Emde Boas tree with an uncertainty-based ranking algorithm to retrieve similar WSIs.

RetCCL (Wang et al., 2023): which adopts CCL-based feature extractor and a ranking aggregation algorithm for WSI retrieval.

HSHR (Li et al., 2023): which utilizes self-supervised learning at the slide level to train the hash encoder, and then leverages a high-order correlation-guided method to retrieve WSIs.

CDC (Li et al., 2022): which applies BYOL (Grill et al., 2020) to initialize feature representations of patches, and then utilizes only slide-level labels and contrastive dynamic clustering to obtain the feature representations for patch-level image retrieval.

HSDH (Alizadeh et al., 2023): which combines a Siamese structure of shared weight parameters with a hash method for patch-level image retrieval, only using slide-level labels.

Note that within the aforementioned retrieval methods, the comparison methods FISH and RetCCL do not utilize hash encoding for WSI retrieval. This absence stems from the lack of popularity and development of hashing techniques within the WSI retrieval domain, leading to a scarcity of hashing methods on WSI retrieval. Thus, it suggests the substantial research significance of our proposed method.

4.3. Implementation details

We implement the proposed method with the PyTorch framework on a server with 8 NVIDIA GeForce 3090 (each one has 24 GB memory). We adopt the Adam (Kingma and Ba, 2014) optimizer with default parameters, and then totally run 25 epochs to alternatively train the second and the third stages. For the second stage, we adjust the learning rate by using the OneCycleLR scheduler (Smith and Topin, 2019), with an initialized learning rate of 0.0001. For the third stage, we initialize the learning rate as 0.0001 and then adjust the learning rate to 0.00005 after 5 epochs. Additionally, we set the batch size as 2 for the second stage, which means that two bags are randomly selected from all the bags in the training dataset, and for the third stage, the batch size is set to 1, which means that all the bags contained in one WSI are selected for model training. We empirically set $\lambda_c = 0.3$, $\lambda_p = 0.2$ and $\lambda_b = 0.1$ for cell-, patch-, and bag-attention, respectively. For the unsupervised weighting function $\omega(\tau)$, we adopt $\omega_i(\tau) = \lambda_i e^{-\left|1-\frac{\tau}{T}\right|^2}$, $i \in \{1,2,3\}$, and empirically set λ_i as 5, 2 and 2.5 in sequence, where T denotes the total number of training epochs and $\tau \in [0, T]$ is the number of current training epochs.

For fairness, we employ the same patch size and downsampling rate as the proposed framework for comparison methods. However, we adopt their default augmentation so as to obtain the best performance of themselves. Additionally, we repeat experiments five times on each dataset for all methods, and then report their average results and standard deviation.

4.4. Evaluation metrics

In this subsection, we briefly introduce the evaluation metrics used in our experiments. In the classification task, we utilize five popular metrics for performance assessment: accuracy (ACC), area under the curve (AUC) score, sensitivity (SE), specificity (SP), and F_1 score. For retrieval tasks, i.e., querying an image will return a series of similar images, we adopt the mean average precision (mAP) and normalized discounted cumulative gain (NDCG) as the evaluation metrics. mAP is derived as the mean of the average precision (AP), which is defined as:

$$AP@K = \frac{\sum_{k=1}^{K} P(k)\delta(k)}{\sum_{k=1}^{K} \delta(k)}, \tag{19}$$

where $P(k)$ is the precision at cut-off k in the returned list, $\delta(k) = 1$ if the image ranked at the kth position is relevant; otherwise, $\delta(k) = 0$. Additionally, for one query image, its NDCG score is calculated as:

$$NDCG@K = \frac{1}{Z} \sum_{k=1}^{K} \frac{2^{r_k} - 1}{log(k+1)}, \tag{20}$$

where r_k is the relevance of the kth nearest neighbors to the query image, Z is a constant to make the maximum of NDCG to be one.

4.5. Results on WSI classification

In this subsection, we present the performance of our proposed method and the other comparison methods on WSI classification. Table 1 displays the classification results of nine methods on three publicly available datasets in term of the five metrics. As we can see, the proposed method obtains the best performance among nine methods on the three datasets in most of cases. Specifically, on the dataset STAD, the gain of the proposed framework is 2.97%, 1.26% and 1.70% over the best competitor in terms of ACC, AUC and F_1 score, respectively. Moreover, on the CAMELYON16 dataset, the proposed framework demonstrates superior performance over the second-best model, achieving enhancements of 2.05%, 1.57% and 7.74% in terms of ACC, AUC and F_1 score, respectively. Similarly, on the CAMELYON17 dataset, the proposed framework outperforms the sub-optimal model with improvements of 1.58%, 0.38% and 2.61% in terms of ACC, AUC and F_1 score, respectively.

It is worth noting that on the CAMELYON16 dataset, MRAN obtains the inferior performance to DSMIL, and this is contrary to the results presented in the paper (Xiang et al., 2023). This discrepancy might be caused by that in the paper (Xiang et al., 2023), MRAN extracting tissue regions within the CAMELYON16 dataset inadvertently produces a large amount of noise regions. However, in this paper, based on MRAN, we further improve the extraction method by using the GLCM (Haralick et al., 1973) from the skimage feature library to largely reduce the amount of noise regions, so that the performance of most of methods is boosted and DSMIL obtains better performance than MRAN. Addi-

Table 1
Classification results of different methods on the three datasets, STAD, CAMELYON16 and CAMELYON17. We bold the best result and underline the second-best result at each setting.

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
DSML	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
TransMIL	0.8431 ± 0.0039	0.7534 ± 0.0615	**0.9939 ± 0.0101**	0.1099 ± 0.0558	0.9131 ± 0.0028
DFTD-MIL	0.7948 ± 0.0722	0.8553 ± 0.0338	0.7898 ± 0.1009	0.8221 ± 0.0936	0.8624 ± 0.0544
MRAN	0.9520 ± 0.0321	0.9849 ± 0.0160	0.9679 ± 0.0108	0.8720 ± 0.2208	0.9715 ± 0.0181
DG-WSDH	**0.9803 ± 0.0123**	**0.9973 ± 0.0039**	0.9794 ± 0.0144	**0.9846 ± 0.0344**	**0.9880 ± 0.0075**
Model	CAMELYON16				
	ACC	AUC	SE	SP	F_1
MSKCC-MIL	0.6308 ± 0.1677	0.7332 ± 0.1107	0.8125 ± 0.1250	0.5043 ± 0.3554	0.6583 ± 0.0976
MSKCC-RNN	0.5897 ± 0.0314	0.5446 ± 0.0679	0.4125 ± 0.1854	0.7130 ± 0.1494	0.4298 ± 0.1346
DSMIL	0.7744 ± 0.0734	0.8973 ± 0.0413	0.6500 ± 0.1746	**0.8609 ± 0.1803**	0.6981 ± 0.0864
CLAM-SB	0.6050 ± 0.1006	0.5870 ± 0.0700	0.5000 ± 0.1326	0.6750 ± 0.1896	0.5013 ± 0.0868
CLAM-MB	0.5800 ± 0.0737	0.5724 ± 0.0868	0.2375 ± 0.2516	0.8083 ± 0.1900	0.2489 ± 0.2576
TransMIL	0.6769 ± 0.0292	0.7174 ± 0.0679	0.4632 ± 0.1515	0.8244 ± 0.0951	0.5387 ± 0.1031
DFTD-MIL	0.7513 ± 0.0444	0.7963 ± 0.0446	0.6703 ± 0.1503	0.8076 ± 0.1064	0.6830 ± 0.0680
MRAN	0.7590 ± 0.0590	0.8190 ± 0.0774	0.6570 ± 0.1220	0.8174 ± 0.1880	0.6961 ± 0.0414
DG-WSDH	**0.7949 ± 0.0363**	**0.9130 ± 0.0135**	**0.8625 ± 0.0523**	0.7478 ± 0.0567	**0.7755 ± 0.0370**
Model	CAMELYON17				
	ACC	AUC	SE	SP	F_1
MSKCC-MIL	0.5476 ± 0.0865	0.4654 ± 0.0118	0.3333 ± 0.3336	0.6712 ± 0.3155	0.2941 ± 0.1991
MSKCC-RNN	0.5635 ± 0.1309	0.5370 ± 0.1081	0.5766 ± 0.2890	0.5551 ± 0.3508	0.4693 ± 0.1034
DSMIL	0.6739 ± 0.0354	0.7078 ± 0.0348	0.3730 ± 0.2344	**0.8455 ± 0.0881**	0.4085 ± 0.2435
CLAM-SB	0.5740 ± 0.0770	0.4974 ± 0.0828	0.4111 ± 0.3703	0.6656 ± 0.3178	0.3332 ± 0.2285
CLAM-MB	0.5360 ± 0.0989	0.5168 ± 0.0312	**0.5944 ± 0.2458**	0.5031 ± 0.2909	0.4658 ± 0.0611
TransMIL	0.6012 ± 0.0663	0.5555 ± 0.0343	0.4000 ± 0.0444	0.7160 ± 0.0959	0.4239 ± 0.0569
DFTD-MIL	0.6567 ± 0.0572	0.6658 ± 0.0362	0.5777 ± 0.1275	0.7018 ± 0.1529	0.5471 ± 0.0391
MRAN	0.6837 ± 0.0457	0.7169 ± 0.0551	0.5459 ± 0.0904	0.7622 ± 0.0834	0.5549 ± 0.0611
DG-WSDH	**0.6995 ± 0.0395**	**0.7207 ± 0.0573**	0.5784 ± 0.0967	0.7687 ± 0.0749	**0.5810 ± 0.0597**

Table 2
Classification results with confidence intervals of different methods on CAMELYON16.

Model	CAMELYON16				
	ACC	AUC	SE	SP	F_1
DSMIL	0.7744	0.8973	0.6500	0.8609	0.6981
(95% CI)	(0.6850-0.8638)	(0.8417-0.9529)	(0.4485-0.8515)	(0.6562-1.0656)	(0.6034-0.7928)
DFTD-MIL	0.7513	0.7963	0.6703	0.8076	0.6830
(95% CI)	(0.6962-0.8064)	(0.7409-0.8517)	(0.4837-0.8569)	(0.6755-0.9397)	(0.5986-0.7674)
MRAN	0.7590	0.8190	0.6570	0.8174	0.6961
(95% CI)	(0.6934-0.8246)	(0.7409-0.8971)	(0.5092-0.8048)	(0.6071-1.0277)	(0.6416-0.7506)
DG-WSDH	0.7949	0.9130	0.8625	0.7478	0.7755
(95% CI)	(0.7498-0.8400)	(0.8962-0.9298)	(0.7976-0.9274)	(0.6774-0.8182)	(0.7296-0.8214)

tionally, it also suggests that MRAN might be more suitable for the WSI classification with a large amount noise information than DSMIL.

Additionally, we have selected the best competitors (e.g., DSMIL, DFTD-MIL, MRAN) based on their performance metrics (ACC, AUC, and F_1) from classification results on WSIs in the CAMELYON16 dataset. Then, we conduct significance testing of the four methods and show their results in Table 2. As we can see, the proposed framework also has superior performance over the best competitors. Similar findings can be observed on the other two datasets.

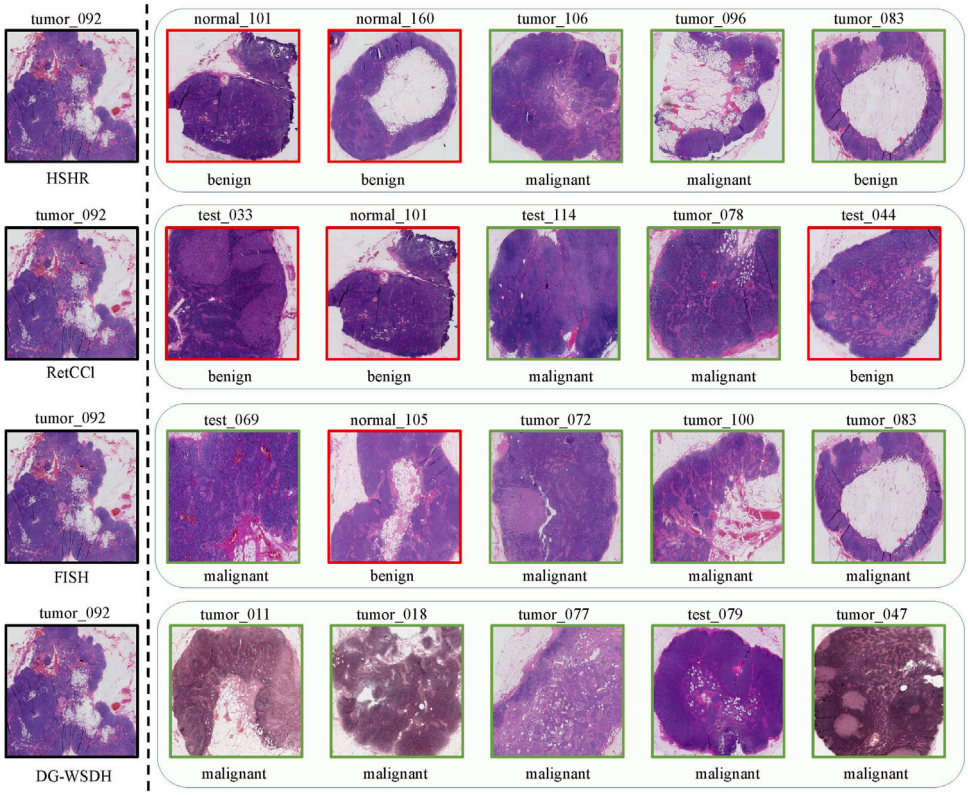
4.6. Results on retrieval tasks

In this subsection, we show the performance of the proposed method and other comparison methods on retrieval tasks. Table 3 displays the WSI retrieval results of four methods on the two publicly available datasets. As we can see, on the CAMELYON16 dataset, the proposed framework demonstrates a notable performance advantage over the best competitors, achieving improvements of 6.11% and 15.16% in terms of MAP@5 and MAP@50, respectively. Similarly, on the CAMELYON17 dataset, our method obtains 2.8% and 6.09% higher performance than the best competitor in terms of MAP@5 and MAP@50, respectively.

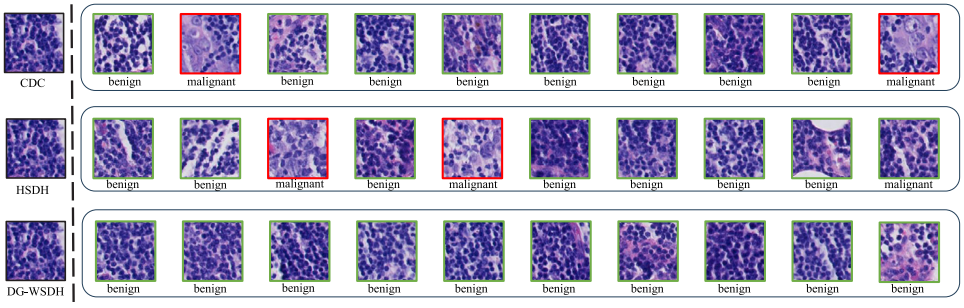
Table 3
WSI retrieval results of different methods on the two public datasets. We bold the best result and underline the second-best result at each setting.

Model	CAMELYON16		CAMELYON17	
	MAP@5	MAP@50	MAP@5	MAP@50
HSHR	0.7050	0.6240	0.7029	0.6398
RetCCl	0.7823	0.6357	0.7067	0.6533
FISH	0.7363	0.6644	0.6924	0.6327
DG-WSDH	**0.8434**	**0.8160**	**0.7347**	**0.7142**

Table 4 presents the similarity retrieval results of three methods for patch-level images on the two public datasets, with only considering whether the patch-level images belong to the same class or not. As we can see, our method achieves 1.56% and 1.69% higher performance



(a) WSI retrieval



(b) Patch retrieval

Fig. 3. Retrieval results of different methods on the CAMELYON16 dataset. The first column denotes the query one and the remaining columns represent the retrieved ones. Additionally, the image with a red box denotes the wrong retrieved one, and the images with green boxes are the corrected retrieved ones.

Table 4
Similarity retrieval results of different hashing methods for patch-level images on the two public datasets. We bold the best result and underline the second best result at each setting.

Model	CAMELYON16		CAMELYON17	
	MAP@50	MAP@500	MAP@50	MAP@500
CDC	0.9120	0.9101	0.9037	0.8971
HSDH	<u>0.9204</u>	<u>0.9172</u>	<u>0.9073</u>	<u>0.9006</u>
DG-WSDH	**0.9360**	**0.9341**	**0.9300**	**0.9238**

on CAMELYON16 in terms of MAP@50 and MAP@500, respectively. Similarly, on the CAMELYON17 dataset, the proposed method also obtains superior performance over the other two methods.

Additionally, Table 5 presents the ranking retrieval results of three methods for patch-level images on the two public datasets, with considering their relevance, i.e., the query with the similar retrieved patches from the same WSI has larger relevance than that with the similar retrieved ones from different WSIs. As we can see, on the CAMELYON16 dataset, the proposed framework achieves the superior performance over the best competitor, achieving improvements of 1.89% and 1.43% in terms of NDCG@10 and NDCG@50, respectively. Similarly, our method also obtains better performance than the other two methods on CAMELYON17.

Moreover, Table 6 shows the similarity retrieval results of different hashing methods for patch-level images within each query WSI on the two public datasets. Note that the query WSI represents the positive WSI in the testing set. As we can see, on the CAMELYON16 dataset, our method consistently performs better than the best competitor, with improvements of 3.09% and 3.26% in terms of MAP@50 and MAP@500, respectively. Similar findings can be observed on the CAMELYON17 dataset.

Table 5
Ranking retrieval results of different methods for patch-level images on the two public datasets. We bold the best result and underline the second best result at each setting.

Model	CAMELYON16		CAMELYON17	
	NDCG@10	NDCG@50	NDCG@10	NDCG@50
CDC	0.9129	0.9052	<u>0.9271</u>	<u>0.9068</u>
HSDH	<u>0.9195</u>	<u>0.9139</u>	0.9059	0.8965
DG-WSDH	**0.9384**	**0.9282**	**0.9450**	**0.9297**

Table 6
Similarity retrieval results of different hashing methods for patch-level images within each query WSI on the two public datasets. We bold the best result and underline the second best result at each setting.

Model	CAMELYON16		CAMELYON17	
	MAP@50	MAP@500	MAP@50	MAP@500
CDC	0.9022	0.8974	0.9420	0.9389
HSDH	<u>0.9114</u>	<u>0.9044</u>	<u>0.9532</u>	<u>0.9496</u>
DG-WSDH	**0.9423**	**0.9370**	**0.9707**	**0.9690**

For clarity, Fig. 3 presents the visualization results of different hashing methods for WSI and patch-level image retrieval on the CAMELYON16 dataset. All Tables 3–6 and Fig. 3 demonstrate the superior performance of our method over the comparison methods on WSI and patch-level image retrieval tasks.

4.7. Ablation study

In this subsection, we show our ablation experiments conducted to evaluate the critical components of our framework, i.e., multi-scale attention and the GCN module with two graph convolutional layers. We summarize the classification and retrieval results on WSIs from the CAMELYON16 dataset in Table 7. Specifically, Baseline denotes that each attention is replaced by mean-pooling, without utilizing the GCN module. Baseline+*Graph* denotes that the integration of the two convolutional graph layers into the Baseline. C_α, P_α and B_α indicate whether cell-, patch- and bag-attention are integrated into the Baseline, respectively. $C_{\alpha+}P_{\alpha+}B_{\alpha+}Graph$ denotes the proposed framework. As we can see from Table 7, both multi-scale attention ($C_{\alpha+}P_{\alpha+}B_\alpha$) and *Graph* can boost the classification and retrieval performance of the model. This illustrates the effectiveness of the multi-scale attention and the dynamic graph module.

Moreover, we compare the proposed dynamic graph (DG) module with a popular graph method, GATv2 (Brody et al., 2021), by replacing the dynamic graph module with GATv2 in our proposed framework. Then, we present their results in Table 8. As we can see, the proposed dynamic graph module has superior classification performance over GATv2. This might be because DG can extract key patches from each bag and use them to represent noisy patches, thus largely reducing the impact of noisy patches on the overall bag to boost the model performance. For clarity, we show examples in the supplemental material.

4.8. Parameters analysis

There are seven significant hyperparameters including λ_1, λ_2, λ_3, λ_c, λ_p and λ_b, and three hyperparameters (H_b, H_p and H_c) to set the size of bag-, patch- and cell-level images. Among them, the settings of λ_1, λ_2 and λ_3 are based on the parameter analysis in the paper MRAN (Xiang et al., 2023). Additionally, the parameter settings including λ_c, λ_p and λ_b are based on the parameter analysis in the paper Loss-Attention (Shi et al., 2020). Thus, we mainly analyze the hyperparameter γ and the three hyperparameters H_b, H_p and H_c in this subsection.

The hyperparameter γ is mainly used to control the sparsity of the adjacency matrix, so as to remove the trivial or irrelevant patches. We explore its effect on the performance of the proposed DG-WSDH and show the results at different values of γ in Fig. 4. As we can see, when $\gamma = 1$, the proposed framework obtains the best accuracy of itself, and thus the hyperparameter γ is very significant in our method.

Additionally, we explore the effect of varied sizes of the bag-, patch-, and cell-level images on the model performance, and show the results in Fig. 5, where we test $H_b = W_b \in \{1024, 2048, 4096\}$, $H_p = W_p \in \{64, 128, 256\}$, and $H_c = W_c \in \{16, 32, 64\}$ for the bag-, patch- and cell-level images, respectively. Note that when testing the effect of a certain size at the bag-, patch-, or cell-level, we maintain the default settings for the other two levels. As we can see from Fig. 5, when $H_b = W_b = 1024$, $H_p = W_p = 128$ and $H_c = W_c = 32$, the proposed method obtains the best performance of itself. Thus, we set them as the default settings. Similar observations can be found on the other two datasets.

4.9. Discussion

Experiments on the three public WSIs datasets illustrate that our proposed DG-WSDH can achieve superior classification and retrieval performance over recent state-of-the-art methods. For WSI classification and retrieval tasks, the major possible reasons are: (i) unlike previous methods like MSKCC-MIL, MSKCC-RNN, DSMIL, CLAM-SB, CLAM-MB, TransMIL, DTFD-MIL, FISH, RetCCL and HSHR, which utilize two-stage processes or rely on unsupervised learning and pre-trained models for feature extraction from patch-level images, our method directly employs convolutional neural networks to extract significant patch-level features from each WSI. (ii) Compared to MRAN, which can also directly extract patch-level features from WSIs, we further take into account the inter-patch relations by constructing dynamic graphs, thereby obtaining more powerful feature extraction capability.

For patch-level image retrieval tasks, (i) When only considering whether the patch-level images belong to the same class or not (shown in Table 4), the superior performance of the proposed method might be because our method employs multi-scale attention mechanisms to obtain more powerful patch-level feature representations. (ii) When considering the relevance of patch-level images (shown in Table 5), our method takes into account the different relevance of similar patches, thereby being beneficial to better ranking performance. (iii) When retrieve similar patches within each WSI (shown in Table 6), our method employs dynamic graph construction to model the relations among patches, thereby possibly being contributed to this task.

5. Conclusion and future work

In this paper, we propose a novel end-to-end MIL-based deep hashing framework, namely DG-WSDH, which can directly extract patch-level features using only slide-level labels to simultaneously handle WSI classification, WSI and patch-level image retrieval tasks. Specifically, this framework consists of a multi-scale representation attention based

Table 7
Ablation study on the dataset CAMELYON16.

Model	ACC	AUC	F_1	MAP@5	MAP@50
Baseline	0.6282 ± 0.1117	0.7541 ± 0.0770	0.6456 ± 0.0724	0.6390	0.6224
Baseline+ *Graph*	0.7051 ± 0.0148	0.8865 ± 0.0494	0.6881 ± 0.0158	0.7459	0.7215
Baseline+ B_α + *Graph*	0.7179 ± 0.0554	0.9100 ± 0.0321	0.6667 ± 0.0559	0.7592	0.7161
Baseline+ $B_\alpha + P_\alpha$ + *Graph*	0.7466 ± 0.0444	0.9015 ± 0.0822	0.7194 ± 0.0422	0.7889	0.7713
Baseline+ $B_\alpha + P_\alpha + C_\alpha$ + *Graph*	$\mathbf{0.7949 \pm 0.0363}$	$\mathbf{0.9130 \pm 0.0135}$	$\mathbf{0.7755 \pm 0.0370}$	**0.8434**	**0.8160**

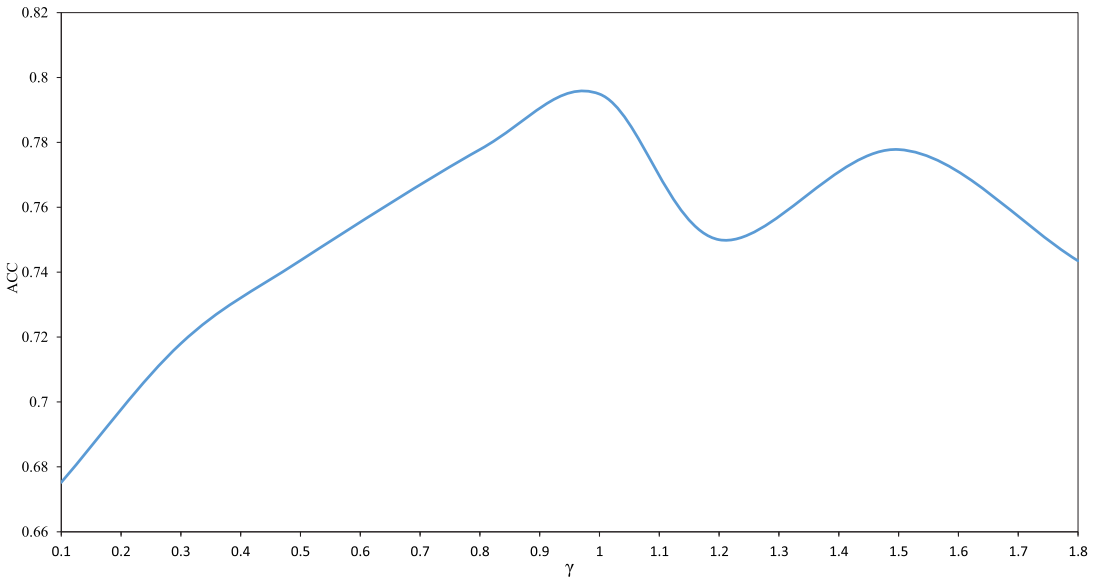


Fig. 4. The classification accuracy of the proposed DG-WSDH on CAMELYON16 at different values of γ in Eq. (4d).

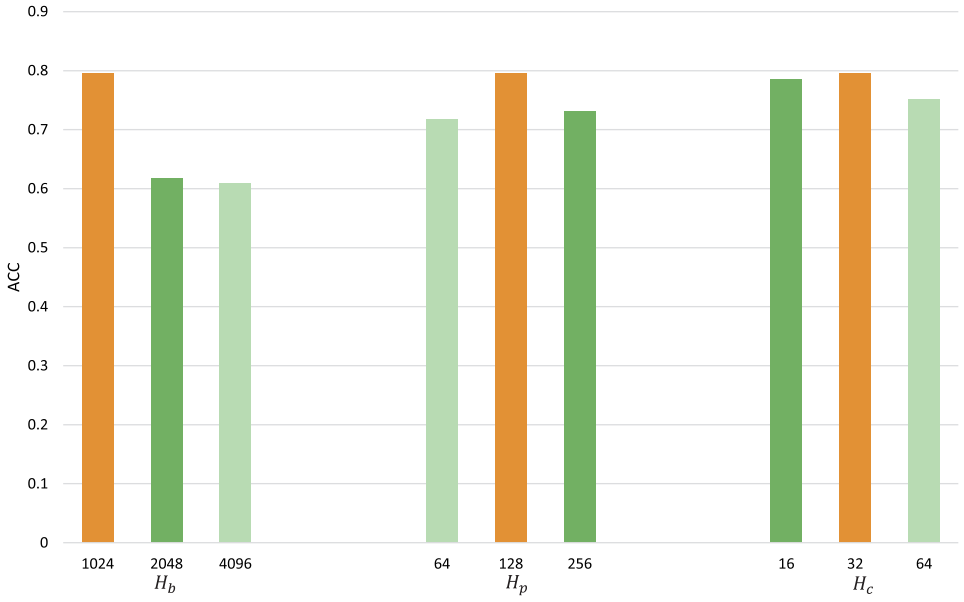


Fig. 5. The classification accuracy of the proposed DG-WSDH on the CAMELYON16 dataset under different sizes of bag-, patch-, and cell- level images.

Table 8
Different message dissemination results on the dataset CAMELYON16.

Model	ACC	AUC	F_1
GATv2(head=1)	0.6325 ± 0.0296	0.6830 ± 0.0265	0.4454 ± 0.1602
GATv2(head=8)	0.6410 ± 0.0444	0.7115 ± 0.1264	0.4209 ± 0.2216
DG	$\mathbf{0.7949 \pm 0.0363}$	$\mathbf{0.9130 \pm 0.0135}$	$\mathbf{0.7755 \pm 0.0370}$

deep network to directly extract patch-level features from WSIs and mine significant information at different scales by only using slide-level labels, dynamic graphs to integrate the relations of patch-level images, and the hashing layer to encode patch- and WSI-level features into binary codes. Additionally, we design a novel objective function to connect attention mechanisms with classification and retrieval tasks. Extensive experiments on three widely public WSI datasets demonstrate the superiority of the proposed model over recent state-of-the-art methods on both classification and retrieval tasks.

Although the proposed framework has achieved promising performance, it is originally designed for binary classification and retrieval tasks, and only applied to the unimodal data. Thus, in the future, we will further extend and boost the proposed framework to handle multi-class and multi-modal tasks. Moreover, the proposed framework employs the traditional convolutional neural network for feature extraction, its performance might be further boosted by using large language models (LLMs). Therefore, in the future, we will integrate LLMs with the proposed framework for obtaining better performance.

CRediT authorship contribution statement

Haochen Jin: Writing – original draft. **Junyi Shen:** Writing – original draft. **Lei Cui:** Writing – review & editing. **Xiaoshuang Shi:** Writing – review & editing. **Kang Li:** Writing – review & editing. **Xiaofeng Zhu:** Writing – review & editing.

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.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 62276052), and Medico-Engineering Cooperation Funds from University of Electronic Science and Technology of China (No. ZYGX2022YGRH014).

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.media.2025.103468.

Data availability

Data will be made available on request.

References

Adnan, M., Kalra, S., Tizhoosh, H.R., 2020. Representation learning of histopathology images using graph neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. IEEE, pp. 988–989.

Akakin, H.C., Gurcan, M.N., 2012. Content-based microscopic image retrieval system for multi-image queries. IEEE Trans. Inf. Technol. Biomed. 16, 758–769.

Alizadeh, S.M., Helfroush, M.S., Müller, H., 2023. A novel siamese deep hashing model for histopathology image retrieval. Expert Syst. Appl. 225, 120169.

Bandi, P., Geessink, O., Manson, Q., et al., 2018. From detection of individual metastases to classification of lymph node status at the patient level: the camelyon17 challenge. IEEE Trans. Med. Imaging 38, 550–560.

Bejnordi, B.E., Veta, M., Van Diest, P., et al., 2017. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. Jama 318, 2199–2210.

Brody, S., Alon, U., Yahav, E., 2021. How attentive are graph attention networks? arXiv preprint arXiv:2105.14491.

Campanella, G., Hanna, M.G., Geneslaw, L., Miraflor, A., et al., 2019. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nature Med. 25, 1301–1309.

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

Chen, R.J., Chen, C., Li, Y., Chen, T.Y., Trister, A.D., et al., 2022. Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16144–16155.

Chen, P., El Hussein, S., Xing, F., Aminu, M., et al., 2022a. Chronic lymphocytic leukemia progression diagnosis with intrinsic cellular patterns via unsupervised clustering. Cancers 14, 2398.

Chen, R.J., Lu, M.Y., Shaban, M., Chen, C., et al., 2021a. Whole slide images are 2D point clouds: Context-aware survival prediction using patch-based graph convolutional networks. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021. Springer, pp. 339–349.

Chen, C., Lu, M.Y., Williamson, D.F., Chen, T.Y., Schaumberg, A.J., Mahmood, F., 2021b. Fast and scalable image search for histology. arXiv preprint arXiv:2107.13587.

Chen, Y., Tang, Y., Huang, J., Xiong, S., 2023. Multi-scale triplet hashing for medical image retrieval. Comput. Biol. Med. 155, 106633.

Chen, M., Wei, Z., Huang, Z., Ding, B., Li, Y., 2020. Simple and deep graph convolutional networks. In: International Conference on Machine Learning. PMLR, pp. 1725–1735.

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

Cruz-Roa, A., Basavanhally, A., González, F., Gilmore, H., et al., 2014. Automatic detection of invasive ductal carcinoma in whole slide images with convolutional neural networks. In: Medical Imaging 2014: Digital Pathology, vol. 9041, SPIE, 904103.

Deng, R., Cui, C., Remedios, L.W., Bao, S., Womick, R.M., et al., 2024. Cross-scale multi-instance learning for pathological image diagnosis. Med. Image Anal. 94, 103124.

Dizaji, K.G., Zheng, F., Sadoughi, N., Yang, Y., Deng, C., Huang, H., 2018. Unsupervised deep generative adversarial hashing network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3664–3673.

Erfankhah, H., Yazdi, M., Babaie, M., Tizhoosh, H.R., 2019. Heterogeneity-aware local binary patterns for retrieval of histopathology images. IEEE Access 7, 18354–18367.

Gallagher-Syed, A., Rossi, L., Rivellese, F., Pitzalis, C., et al., 2023. Multi-stain self-attention graph multiple instance learning pipeline for histopathology whole slide images. arXiv preprint arXiv:2309.10650.

Gour, M., Jain, S., Sunil Kumar, T., 2020. Residual learning based CNN for breast cancer histopathological image classification. Int. J. Imaging Syst. Technol. 30, 621–635.

Grill, J.-B., Strub, F., Altché, F., Tallec, C.o., 2020. Bootstrap your own latent-a new approach to self-supervised learning. Adv. Neural Inf. Process. Syst. 33, 21271–21284.

Haralick, R.M., Shanmugam, K., Dinstein, I.H., 1973. Textural features for image classification. IEEE Trans. Syst. Man Cybern. 610–621.

Hashimoto, N., Fukushima, D., Koga, R., et al., 2020. Multi-scale domain-adversarial multiple-instance CNN for cancer subtype classification with unannotated histopathological images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3852–3861.

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

Kalra, S., Tizhoosh, H.R., Shah, S., Choi, C., et al., 2020. Pan-cancer diagnostic consensus through searching archival histopathology images using artificial intelligence. NPJ Digit. Med. 3, 31.

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, 414.

Kingma, D.P., Ba, J., 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Kipf, T.N., Welling, M., 2016. Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907.

Li, J., Jiang, Z., Zheng, Y., Zhang, H., et al., 2022. Weakly supervised histopathological image representation learning based on contrastive dynamic clustering. In: Medical Imaging 2022: Digital and Computational Pathology, vol. 12039, SPIE, pp. 14–19.

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, S., Zhao, Y., Zhang, J., Yu, T., Zhang, J., Gao, Y., 2023. High-order correlation-guided slide-level histology retrieval with self-supervised hashing. IEEE Trans. Pattern Anal. Mach. Intell..

Litjens, G., Kooi, T., Bejnordi, B.E., et al., 2017. A survey on deep learning in medical image analysis. Med. Image Anal. 42, 60–88.

Liu, C., Ding, W., Cheng, C., Tang, C., Huang, J., Wang, H., 2022. DenseHashNet: A novel deep hashing for medical image retrieval. IEEE J. Radio Freq. Identif. 6, 697–702.

Liu, Y., Jain, A., Eng, C., Way, D.H., et al., 2020. A deep learning system for differential diagnosis of skin diseases. Nature Med. 26, 900–908.

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

Ma, Y., Jiang, Z., Zhang, H., Xie, F., Zheng, Y., Shi, H., Zhao, Y., 2016. Breast histopathological image retrieval based on latent Dirichlet allocation. IEEE J. Biomed. Health Inf. 21, 1114–1123.

Ma, Y., Jiang, Z., Zhang, H., Xie, F., et al., 2018. Generating region proposals for histopathological whole slide image retrieval. Comput. Methods Programs Biomed. 159, 1–10.

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.

Peng, L., Wang, N., Dvornek, N., Zhu, X., Li, X., 2022. Fedni: Federated graph learning with network inpainting for population-based disease prediction. IEEE Trans. Med. Imaging.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., et al., 2015. Imagenet large scale visual recognition challenge. Int. J. Comput. Vis. 115, 211–252.

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. Adv. Neural Inf. Process. Syst. 34, 2136–2147.

Shi, X., Sapkota, M., Xing, F., Liu, F., Cui, L., Yang, L., 2018. Pairwise based deep ranking hashing for histopathology image classification and retrieval. Pattern Recognit. 81, 14–22.

Shi, X., Xing, F., Xu, K., Chen, P., Liang, Y., Lu, Z., Guo, Z., 2020. Loss-based attention for interpreting image-level prediction of convolutional neural networks. IEEE Trans. Image Process. 30, 1662–1675.

Shi, X., Xing, F., Xu, K., Xie, Y., Su, H., Yang, L., 2017. Supervised graph hashing for histopathology image retrieval and classification. Med. Image Anal. 42, 117–128.

Sirinukunwattana, K., Pluim, J.P., Chen, H., Qi, X., 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: Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications, vol. 11006, SPIE, pp. 369–386.
Tang, Y., Chen, Y., Xiong, S., 2022. Deep semantic ranking hashing based on self-attention for medical image retrieval. In: 2022 26th International Conference on Pattern Recognition. ICPR, IEEE, pp. 4960–4966.
Tomczak, K., Czerwińska, P., Wiznerowicz, M., 2015. Review the cancer genome atlas (TCGA): an immeasurable source of knowledge. Contemp. Oncology/ WspÓ l Czesna Onkol. 2015, 68–77.
Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., et al., 2017. Attention is all you need. Adv. Neural Inf. Process. Syst. 30.
Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., Bengio, Y., 2017. Graph attention networks. arXiv preprint arXiv:1710.10903.
Wang, X., Du, Y., Yang, S., Zhang, J., et al., 2023. RetCCL: Clustering-guided contrastive learning for whole-slide image retrieval. Med. Image Anal. 83, 102645.
Wang, Y., Song, J., Zhou, K., Liu, Y., 2021. Unsupervised deep hashing with node representation for image retrieval. Pattern Recognit. 112, 107785.
Wang, X., Yang, S., Zhang, J., Wang, M., et al., 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.
Xu, Y., Jia, Z., Wang, L.-B., Ai, Y., Zhang, F., Lai, M., Chang, E.I.-C., 2017. Large scale tissue histopathology image classification, segmentation, and visualization via deep convolutional activation features. BMC Bioinformatics 18, 1–17.
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, 591–604.
Yu, K.-H., Beam, A.L., Kohane, I.S., 2018. Artificial intelligence in healthcare. Nat. Biomed. Eng. 2, 719–731.
Zhang, Z., Chen, P., McGough, M., Xing, F., et al., 2019. Pathologist-level interpretable whole-slide cancer diagnosis with deep learning. Nat. Mach. Intell. 1, 236–245.
Zhang, H., Meng, Y., Zhao, Y., Qiao, Y., et al., 2022a. 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.
Zhang, Y., Ou, W., Shi, Y., Deng, J., You, X., Wang, A., 2022b. Deep medical cross-modal attention hashing. World Wide Web 25, 1519–1536.
Zheng, Y., Gindra, R.H., Green, E.J., Burks, E.J., et al., 2022a. A graph-transformer for whole slide image classification. IEEE Trans. Med. Imaging 41, 3003–3015.
Zheng, Y., Jiang, Z., Shi, J., Xie, F., et al., 2022b. Encoding histopathology whole slide images with location-aware graphs for diagnostically relevant regions retrieval. Med. Image Anal. 76, 102308.
Zheng, Y., Jiang, Z., Zhang, H., Xie, F., Ma, Y., Shi, H., Zhao, Y., 2017. Size-scalable content-based histopathological image retrieval from database that consists of WSIs. IEEE J. Biomed. Health Inf. 22, 1278–1287.
Zheng, L., Wetzel, A.W., Gilbertson, J., Becich, M.J., 2003. Design and analysis of a content-based pathology image retrieval system. IEEE Trans. Inf. Technol. Biomed. 7, 249–255.
Zhu, W., Sun, L., Huang, J., Han, L., Zhang, D., 2021. Dual attention multi-instance deep learning for Alzheimer's disease diagnosis with structural MRI. IEEE Trans. Med. Imaging 40, 2354–2366.

Haochen Jin received the B.S. degree in cyberspace security from the University of Electronic Science and Technology of China, Chengdu, China, in 2022. He is currently working toward the M.S. degree in the 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.

Lei Cui is a faculty member in the Department of Computer Science and Technology at the Northwest University of China (NWU). He obtained his PhD degree (2019) from Northwest University, and Bachelor degree (2011) from Northwest University. His major research interests include deep learning and medical image analysis.

Xiaoshuang Shi is a faculty member in the Department of Computer Science and Engineering at the University of Electronic Science and Technology of China (UESTC). He obtained his PhD 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 largescale data retrieval, deep learning, medical image analysis.

Kang Li received the Ph.D. degree in Mechanical Engineering from University of Illinois at Urbana Champaign, Champaign, IL, USA, in 2009. He is now a full professor of the Biomedical Big Center at West China Hospital. Before joining West China Hospital, He was an associate professor with the Department of Orthopaedics, New Jersey Medical School (NJMS), Rutgers University, Newark, NJ, USA, and an assistant professor with Department of Industrial and Systems Engineering, Rutgers University. His research interests include AI in healthcare, musculoskeletal biomechanics, medical imaging, design and biorobotics, human reliability, and human factors/ergonomics.

Xiaofeng Zhu is a faculty member 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.