Computer Methods and Programs in Biomedicine 247 (2024) 108099

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

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

HSG-MGAF Net: Heterogeneous subgraph-guided multiscale graph attention fusion network for interpretable prediction of whole-slide image

Meiyan Liang [a,*], Xing Jiang [a], Jie Cao [b,*], Shupeng Zhang [a], Haishun Liu [c], Bo Li [d], Lin Wang [e], Cunlin Zhang [f], Xiaojun Jia [a]

[a] School of Physics and Electronic Engineering, Shanxi University, Taiyuan 030006, China
[b] School of Optics and Photonics, Beijing Institute of Technology, Beijing 100081, China
[c] Department of Automation, Tsinghua University, Beijing 100084, China
[d] Department of Rehabilitation Treatment, Shanxi Rongjun Hospital, Taiyuan 030000, China
[e] Department of Pathology, Shanxi Bethune Hospital, Shanxi Academy of Medical Sciences, Tongji Shanxi Hospital, Third Hospital of Shanxi Medical University, Taiyuan 030032, China
[f] Department of physics, Capital Normal University, Beijing 100048, China

ARTICLE INFO

Keywords:
Heterogeneous subgraph
Multiscale
Whole slide image
Interpretability
Graph
Hypergraph

ABSTRACT

Background and objective: Pathological whole slide image (WSI) prediction and region of interest (ROI) localization are important issues in computer-aided diagnosis and postoperative analysis in clinical applications. Existing computer-aided methods for predicting WSI are mainly based on multiple instance learning (MIL) and its variants. However, most of the methods are based on instance independence and identical distribution assumption and performed at a single scale, which not fully exploit the hierarchical multiscale heterogeneous information contained in WSI.

Methods: Heterogeneous Subgraph-Guided Multiscale Graph Attention Fusion Network (HSG-MGAF Net) is proposed to build the topology of critical image patches at two scales for adaptive WSI prediction and lesion localization. The HSG-MGAF Net simulates the hierarchical heterogeneous information of WSI through graph and hypergraph at two scales, respectively. This framework not only fully exploits the low-order and potential high-order correlations of image patches at each scale, but also leverages the heterogeneous information of the two scales for adaptive WSI prediction.

Results: We validate the superiority of the proposed method on the CAMELYON16 and the TCGA- NSCLC, and the results show that HSG-MGAF Net outperforms the state-of-the-art method on both datasets. The average ACC, AUC and F_1 score of HSG-MGAF Net can reach 92.7 %/0.951/0.892 and 92.2 %/0.957/0.919, respectively. The obtained heatmaps can also localize the positive regions more accurately, which have great consistency with the pixel-level labels.

Conclusions: The results demonstrate that HSG-MGAF Net outperforms existing weakly supervised learning methods by introducing critical heterogeneous information between the two scales. This approach paves the way for further research on light weighted heterogeneous graph-based WSI prediction and ROI localization.

1. Introduction

Pathology whole slide images (WSI) is considered to be the "gold standard" for confirming the presence of cancer [1]. To date, the pathological diagnostic conclusions are obtained based primarily on manual inspections of specialists. This process is not only labor-intensive and time-consuming, but also relies seriously on subjective interpretation, which is a challenging task for precision medicine [2–4]. Therefore, developing artificial intelligence systems for automatically diagnosing and analyzing whole slide images are the prospective trend in computational pathology [5–8]. However, most WSIs have hundreds of millions of pixels (e.g. the typical size is 40,000 × 30,000), which lacks of pixel-wise annotations [9,10]. Thus, it is of great significance to develop an effective weakly-supervised learning method for WSI prediction and ROI localization.

Currently, the existing computer-aided methods for predicting WSI

* Corresponding authors.
E-mail addresses: meiyanliang@sxu.edu.cn (M. Liang), caojie@bit.edu.cn (J. Cao).

https://doi.org/10.1016/j.cmpb.2024.108099
Received 17 December 2023; Received in revised form 12 February 2024; Accepted 22 February 2024
Available online 23 February 2024

are mainly based on multiple instance learning (MIL) [11,12] and its variants. In MIL algorithm, it treats WSI as a bag consisting a multiple of cropped patches called instances [13]. The slide-level label prediction is obtained by aggregating the features of partial instance in the WSI [14]. However, these approaches have the following drawbacks: (1) They do not reasonably consider the contextual information existing between instances in WSI based on the instance independence and identical distribution (*i.i.d*) assumption. (2) Most of the MIL-based methods are based on single-scale and lack consideration of cross-scale contextual heterogeneous information of WSI. (3) The massive computational cost in processing the whole slide images. In real scenarios, whole slide images can provide multi-scale heterogeneous information on macroscopic and microscopic features of tissue phenotypes, which plays an important role in pathological diagnosis and analysis. The heterogeneous features between different magnifications are shown in Fig. 1.

1.1. Related works

In the field of whole slide image prediction and ROI localization, the mainstream methods are multi-instance learning and its variants. These methods generally divide pathology slides into patches named instance, and extract the feature embedding of each patch for further processing. The existing MIL-based deep learning frameworks mainly include CNN+MIL [15–18] and GCN+MIL architectures [19–22].

1.1.1. CNN-based mil approaches

Many CNN+MIL frameworks are based on the hypothesis that instances follow an independent and identical distributions (*i.i.d*) supplemented by instance-level constraints. For instance, top-k instances are selected as an effective restriction to assist the CNN+MIL model in highlighting ROIs and making slide-level label predictions [23–25]. However, it ignores the fine-grained spatial relationships which are implied between instances in the slide. Only a subset of instances in each slide participates in the training process, which will result in performance degradation. Therefore, Transformer-based MIL architectures [26–28], multi-scale MIL [29–31], and other features representation [12,32,33] methods are proposed to exploit contextual information within the slide for further prediction.

Transformer-based MIL Architectures. To fully consider the global correlation between instances in WSI, transformer-based MIL methods [26–28,34] are surged as a promising branch to capture contextual information of the instance in the slide. For instance, Huang et al. [26] proposed SeTranSurv for survival prediction, which uses Transformer to adaptively aggregate the patch features in the slide based on their spatial information and correlations. In 2021, Shao et al. [34] design a TransMIL architecture that breaks the *i.i.d.* assumptions to establish morphological and spatial feature correlations among all instances in a slide. In practice, these approaches can exploit the context-aware information of slides to improve prediction performance to a certain extent. Unfortunately, not all instances in the slide are related to each other. Computing the attention scores for these irrelevant instances not only leads to limited performance improvements, but also places undesirable burden on computing devices.

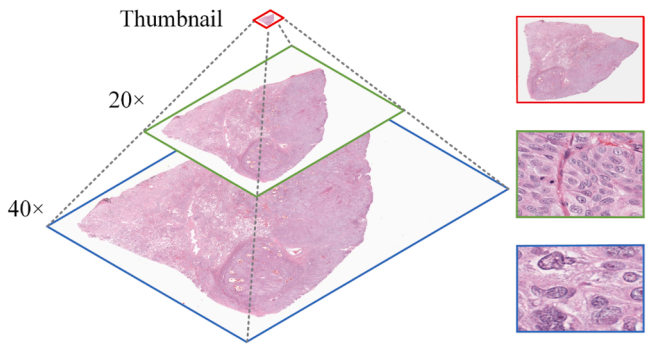


Fig. 1. The representation of WSI pyramid. The patches at lower resolution and the patches at higher resolution are treated as heterogeneous node representations for a slide (right).

Multi-Scale MIL Architectures. Inspired by the diagnostic workflow of pathologists, the utilization of multiscale features has been proved beneficial in MIL-based approaches [29–31,35]. For instance, Li et al. [35] proposed a DSMIL based on a dual-stream architecture, which adopts a pyramid fusion mechanism for concatenate features of two scales and combines trainable distance measurements to improve classification and localization accuracy. To reduce the computational cost of MIL methods based on multiple scales, Thandiackal et al. [30] proposed a ZoomMIL to obtain the slide-level representation by aggregating contextual information of the pathological slide at multiple magnifications through a multi-scale local zooming strategy. However, these multi-scale MIL methods only concatenate aggregated features from two scales for final prediction. They do not consider the heterogeneous information implicited at both scales and the contribution of each scale.

Feature representation methods. Some features representation approaches based on MIL have been applied for WSI prediction [12,23,32,33]. Campanella et al. [23] proposed MIL-RNN to aggregate features of selected top-ranked patches to obtain sliding-level predictions more effectively. However, only using top-k instances for feature representation has large bias and also leads to poor interpretability. In 2022, Zhang et al. [33] proposed DTFD-MIL to effectively utilize the intrinsic features of instances by dividing each slide into multiple pseudo-bags. However, the performance of these models is also limited as they do not fully consider and exploit the global context-aware information of slide.

Overall, CNN+MIL frameworks, either based on multi-scale architecture or single-scale with versatile instance-level constraints, are insufficient to describe the refined feature correlations of instances in the slides, thus affecting prediction accuracy and interpretability. Therefore, GCN+MIL is proposed to leverage the graph to simulate context-aware topology of pathological slides to obtain a global representation. The instances in graph could be cell nuclei or patches.

1.1.2. GCN-based mil approaches

Cell-based graphs. Cell-based graphs [36–39] are proposed to simulate the contextual topology of whole slide images, which use cells as graph nodes and construct edges based on the Euclidean distance of the nodes. Here, the cell features and edge connections are fed into GCN to perform WSI classification. The representative studies are as follows: Sureka et al. [39] exploited a spatial hop counts of cell nuclei to build cell graph and utilized an attention-based graph network to learn contextual information between cells. Lu et al. [37] use a graph representation to describe cell-level context-aware topology of pathology slides. However, these methods are based on the precise localization of the cell nucleus. Each cell can only establish relationships with surrounding cells, which lacks the expression of long-distance dependence among cells. Besides, the performance of this framework is limited because the properties of heterogeneous cell nuclei cannot be fully expressed.

Patch-based graphs. Compared with cell-based graphs, the whole slide image can be cropped to multiple of patches. Each patch in the slide is treated as a node to form a patch-based graph based on its neighbors [21,28,40,41]. For instance, Chen et al. [40] construct a Patch-GCN framework to simulate spatial context-aware information of WSI for survival outcome prediction. The results showed that there is a significant improvement in performance compared to the CNN+MIL architecture. To increase flexibility and reduce computational pressure, Liu et al. [41] used sparsely sampled patches of the slide to construct a dynamic graph network, which not only adaptively adjusts the potential correlation between patches but also appropriately represents pathology slides for survival prediction.

In clinical scenarios, the relations between instances in slide are more complex and cannot be effectively represented by the GCN+MIL

framework alone. In practice, low-order, complex high-order correlations and multi-scale heterogeneous information were coexisted in the instances of a WSI. The accurate representation and description of these information can provide fine-grained diagnostic information, which can further improve the predictive performance of the model. In spired by the workflow of pathologist, Wang et al. [42] proposed a H^2-MIL to learn hierarchical representation from a heterogeneous graph with different resolutions in WSI. However, this type of heterogeneity is based on all instances at multiple scales for a specific slide, which can lead to a huge computation burden. Besides, most of the existing MIL methods that utilize multi-scale features are based on concatenation operation, which only stacks the obtained features of multiple scales with equal weight, is insufficient in considering the contribution of each scale. To solve the above challenges, Heterogeneous Subgraph-Guided Multiscale Graph Attention Fusion Network (HSG-MGAF Net) is proposed to build the topology of top ranked image patches at multiple scales for adaptive WSI prediction and lesion localization. This framework not only exploits the low-order and potential high-order correlations of image patches at each scale through graphs and hypergraphs, but also hierarchically leverages the most significant heterogeneous information at both scales for adaptive WSI prediction. Meanwhile, Self-Supervised contrastive learning is applied across the two scales to maximize the consistency of predictions on both scales. The results show that HSG-MGAF Net outperforms the state-of-the-art (SOTA) methods on both datasets.

1.2. Contributions

The main contributions of our framework are as follows:

- Inspired by the workflow of pathologists, a HSG-MGAF Net is proposed to adaptively establish low-order and potential high-order spatial relationships of patches at two scales for interpretable prediction of WSI using weak labels.
- Top-2k heterogeneous subgraph (Top-2k HSG) is proposed to guide multi-scale graph network for adaptive slide prediction and ROI localization. This method exploits the heterogeneous topological relationships of image patches at two scales without introducing too much computational burden.
- Self-supervised contrastive learning constraint is introduced across the two scales of HSG- MGAF Net to maximize the consistency of predictions at both scales in the optimization process.
- Interpretable heatmaps with complementary features can be obtained by using heterogeneous relationships of the top ranked patches at two scales.

2. Methods

2.1. MIL formulation

In deep learning, multiple-instance learning (MIL) is a form of weakly supervised learning paradigm for addressing weakly annotated data. Specifically, instead of receiving a set of instances which are individually annotated, the learner receives a set of labeled bags, each consisting of multiple instances. Thus, MIL is applicable for processing gigapixel whole slide images, where each WSI is considered as a bag and the image patches within the bag are regarded as instances. In binary classification scenarios, a bag is labeled negative if all the instances in it are negative. While a bag is labeled positive if there is at least one instance in it which is positive.

2.2. Data preprocessing

Given a WSI dataset $D = \{X, Y\}$ where X and Y denote the input whole slide image and the assigned slide-level label. In our case, $Y \in \{0, 1\}$. Firstly, automatic background removal is performed by Otsu's binarization in HSV space at scale 20 × and scale 40 × , respectively. Then, the foreground of slide is cropped into non-overlapping patches of fixed size 256 × 256 at both scales. Therefore, the slide X can be represent as $X^s = \{x_j\}_{j=1}^{P_s}$. Where x_j denote the jth instance in slide X at scale s. Here, each instance x_j implies a binary pseudo-label y_j, which is not defined exactly. P_s is the number of instances in for scale $s(s=1,2)$, which varies widely for each slide. Meanwhile, the x-y coordinates of these instances are preserved in the process. According to MIL assumption, the relationship between y_j and Y is given by

$$Y = \begin{cases} 0, & if \sum_j y_j = 0, \\ 1, & otherwise. \end{cases} \tag{1}$$

2.3. Graph and hypergraph construction

Inspired by the workflow of pathologists, a dual-stream architecture is proposed that uses graph and hypergraph networks to formulate low-order and complex high-order correlations of the slides at scale 20 × and 40 × , respectively. Fig. 2(a) shows the graph and hypergraph construction process of the whole slide images. Here, the graph network is used to formulate the spatial correlation between patches in the slide at low magnification, which attempts to capture the features of nodes in the local neighborhood. While the hypergraph network is applied to simulate the high-order correlations of patches in the slide at high magnification, which allows the model to obtain the feature correlations across the whole slide. Therefore, HSG-MGAF Net can more comprehensively depict the contextual topology of patches in the slide through multi-scale architectures. In the framework, the pre-trained ResNet50 is applied for feature extraction at scale 20 × and 40 × , respectively. Thus, the feature representations of the slide X at scales 20 × and 40 × can be denoted as $F^s = \{f_j^d\}_{j=1}^{P_s} (f_j^d \in \mathbb{R}^{1\times 1024}, s = 1, 2)$.

Graph Construction. In scale 20 × , graph is constructed based on the instances which are treated as nodes in the graph. The K-nearest neighbor (K-NN) algorithm is used to construct the spatial connections of each node. That is, the algorithm simulates a 3 × 3 receptive field by establishing spatial connections between each image patch and the surrounding 8 image patches. Therefore, the WSI is represented as $G = (F^1, A)$ based on the concept of the graph. Where A is the adjacency matrix of input slide X at scale 20 × in branch 1, which can be expressed as:

$$A^{i,j} = \begin{cases} 1, & if\ x_i is\ adjacent\ to\ x_j, \\ 0, & otherwise. \end{cases} \tag{2}$$

Here, $j, j' \in \{1, 2, \cdots, P_1\}$.

Hypergraph Construction. Likewise, hypergraph is constructed based on the patches at scale 40 × , which are also treated as nodes/hypernodes in the hypergraph. With the feature representation of hypernodes F^2, the hypergraph $\widetilde{G}$ is constructed by connecting highly correlated hyper nodes using hyper edges according to their Euclidean distance in latent space. This correlation is denoted by incidence matrix H. Here, three combinations of hyper edges are included in the hypergraph. These hyper edges could connect 2, 3 and 4 hypernodes in the iteration. Therefore, hypergraph provides an effective way to capture higher-order dependencies between image patches in a slide. When the hyperedge is incident with hyper nodes, the corresponding element of H can be given by $h_{v,e}=1$, otherwise 0. Note that, the nodes in graph (in branch 1) and the corresponding nodes in hypergraph are heterogeneous to each other.

Therefore, the hypergraph of the slide is represented as $\widetilde{G} = (F^2, H)$, H is the incidence matrix of input slide X at scale 40 × in branch 2. Each entry of H can be expressed as:

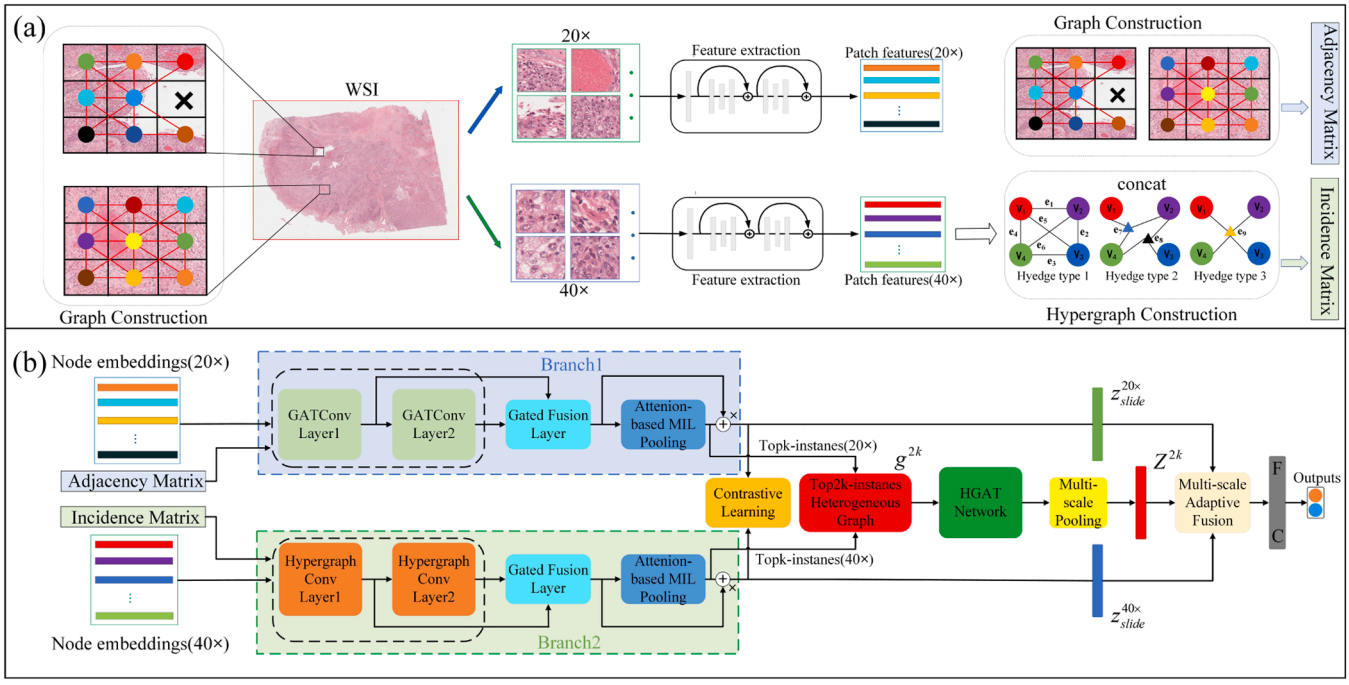


Fig. 2. HSG-MGAF Net. (a) Preprocessing of WSI and construction of graph and hypergraph at two scales. (b) The block diagram of HSG-MGAF Net.

$$h_{v,e} = \begin{cases} 1, & if\ v \in e \\ 0, & if\ v \notin e \end{cases} \tag{3}$$

Here, v and e denote the hyper node and hyper edge in the constructed hypergraph in branch 2, respectively.

2.4. The framework of HSG-MGAF Net

Fig. 2(b) is the schematic diagram of HSG-MGAF Net. As shown in Fig. 2(b), the framework mainly consists of two branches. The framework in branch 1 is a gated graph attention network, which is proposed to formulate low-order correlations of image patches at low magnification based on attention mechanism. While the framework in branch 2 is a gated hypergraph convolutional network, which is applied to exploit the potential high-order correlations of fine-grained image patches at higher magnifications. In each branch, it includes two GAT/hypergraph convolutional layers, a gated fusion layer and an attention-based pooling module. Afterwards, a multi-scale adaptive fusion module is applied to fuse the obtained features of the two branches according to their weights. Note that, the weights are obtained by slide-level representation of a specific scale and the aggregate feature of Top2k heterogeneous subgraph (Top2k-HSG). Here, Top2k heterogeneous subgraph across the two scales is also obtained by attention-based pooling module in each branch during the iterative process, which contains the critical features of the WSI at these scales. For Top2k-HSG, its features are extracted and aggregated by the HGAT network and multi-scale pooling operation, respectively. Therefore, it could fully utilize the multi-scale context-aware heterogeneous information of the highly ranked instances to further improve the performance of the slide prediction and ROI localization.

2.4.1. Gated graph attention network

As shown in Fig. 2(b), the gated graph attention network consists of two graph convolutional layers, a gated fusion layer, and an attention-based pooling module. In this branch, whole slide image prediction can be treated as a graph classification problem after slide-level graph construction.

For graph G, the instance x_j is transformed into a d-dimensional feature vector by a truncated ResNet50 and denoted as $h_j \in \mathbb{R}^{1\times d}$. Therefore, the graph mapping function from l-th to $(l+1)$-th layer can be represented as: $F^{(l)} = \{h_1^{(l)}, \cdots h_j^{(l)}, \cdots h_{P_s}^{(l)}\} \in \mathbb{R}^{P_1\times d_{in}} \rightarrow F^{(l+1)} = \{h_1^{(l+1)}, \cdots h_j^{(l+1)}, \cdots h_{P_s}^{(l+1)}\} \in \mathbb{R}^{P_1\times d_{out}}$. Here, $\{d_{in} \rightarrow d_{out}\}$ are $\{1024 \rightarrow 512\}$ and $\{512 \rightarrow 256\}$ for the two graph convolutional layers.

To enhance the expressive ability of node features, the shared linear transformation parameterized by the weight matrix $w \in \mathbb{R}^{d_{out}\times d_{in}}$ transforms the features of the input nodes into higher-level features. Meanwhile, self-attention mechanism is used to calculate the attention coefficient $e_{j,j'}$ between node j and its neighbor nodes $x_{j'} \in Ne(x_j)$.

$$e_{j,j} = a_{att}\left(wh_j, wh_j\right) \tag{4}$$

$e_{j,j'}$ indicates the contribution for the feature of node $x_{j'}$ to node x_j.

Then, $e_{j,j'}$ is normalized to obtain attention score of each neighbor to better describe the node contributions:

$$a_{j,j} = softmax\left(e_{j,j}\right) \tag{5}$$

The attention scores of the graph node x_j and its 8-neighbor nodes are calculated as follows:

$$a_{j,j} = \frac{\exp\left\{LeakyReLU\left(\left(a_{att}\right)^{\mathrm{T}}\left[wh_j \parallel wh_j\right]\right)\right\}}{\sum_{x_j\in Ne\left(x_j\right)}\exp\left\{LeakyReLU\left(\left(a_{att}\right)^{\mathrm{T}}\left[wh_j \parallel wh_j\right]\right)\right\}} \tag{6}$$

Where $a_{att} \in \mathbb{R}^{1\times 2d_{out}}$ denote a weight vector. T represents the matrix transpose operation, and $\parallel$ is the concatenation operation. The normalized coefficient $a_{j,j'}$ is then used to update feature representation of node x_j:

$$h_j^{l+1} = \sigma\left(\sum_{x_j\in Ne\left(x_j\right)} a_{j,j} wh_j^l\right) \tag{7}$$

Where $\sigma(\cdot)$ denote a activation function.

In the framework, each graph convolution layer is followed by layer normalization and dropout operation to improve generalization ability, accelerate convergence and prevent the gradient over-smoothing issues of the model.

2.4.2. Gated hypergraph convolution network

Gated hypergraph convolution network is mainly composed of two hypergraph convolution layers [43], a gated fusion layer, and an attention-based pooling module. Compared with graph architecture, hypergraph network is more flexible and can better describe potential high-order relationships between the nodes. Therefore, a gated hypergraph convolutional network is constructed at 40 × to depict the correlations between the fine-grained features of the image patches.

We learn a hypergraph mapping function for the slide: $\widetilde{G} = (F^2, H)$:$(F^2)^{(l)} \in \mathbb{R}^{P_2 \times d_{in}'} \rightarrow (F^2)^{(l+1)} \in \mathbb{R}^{P_2 \times d_{out}'}$. $(F^2)^l$ and $(F^2)^{(l+1)}$ indicate the node feature representation in the lth layer and $(l+1)$ layer, respectively. Here, $\{d_{in}' \rightarrow d_{out}'\}$ are {1024 → 512} and {512 → 256} for the two hypergraph convolutional layers.

Here, we use hypergraph Laplacian operator to realize hypergraph convolution, which can be defined as:

$$\widetilde{H} = D_v^{-1/2} H W D_e^{-1} H^{\mathrm{T}} D_v^{-1/2} \tag{8}$$

$\widetilde{H}$ denotes hypergraph laplacian operator. $H \in \mathbb{R}^{P_2 \times E}$ is the incidence matrix of the hypergraph $\widetilde{G}$, and T represents the transpose operation. $D_v \in \mathbb{R}^{P_2 \times P_2}$ and $D_e \in \mathbb{R}^{E \times E}$ are degree matrices of vertices and hyperedges, respectively. $W \in \mathbb{R}^{E \times E}$ is the weight matrix of hyperedges. Therefore, the message passing process of hypergraph convolution layer can be expressed as:

$$F^{(l+1)} = \sigma\left(\widetilde{H}(F^2)^{(l)} P^{(l)}\right) \tag{9}$$

Where σ(·) is the LeakyReLu function, and $P^{(l)}$ denotes the weight matrix of the first two layers.

2.4.3. Gated fusion layer

Node representation evolves on the graph as the number of layers' increases. Therefore, it is necessary to fuse low-level features and high-level semantic features in these layers. As shown in Fig. 3, Gated Fusion Layer (GFL) is proposed to selectively fuse the features from former layer and semantic layer to improve the effectiveness of feature representation. Moreover, GFL can also increase the stability of the model and avoid overfitting issues.

Here, the obtained features in former layer and deep layer are represented as $F_1 \in \mathbb{R}^{P_s \times 512}$ and $F_2 \in \mathbb{R}^{P_s \times 256}$. These two features are prefused to obtain the feature $F_C \in \mathbb{R}^{n \times 256}$ through concatenation and convolution operations (Conv1). Therefore, the gated map is obtained by

$$G_C = Sigmoid(conv2(F_C)) \tag{10}$$

Where $F_C = conv1(F_1; F_2)$, $conv2$ also denote convolution operation. The gated map $G_C \in [0,1]^{P_s \times 256}$.

Node representations at deeper layers have larger receptive field and can effectively capture contextual information of the slide within a larger scope. Therefore, it is reasonable to assign larger weight to deep features rather than shallow features. Thus, the fused feature after gated fusion layer can be defined as:

$$F_G = (1 + G_C) \odot F_2 + (1 - G_C) \odot F_1 \tag{11}$$

⊙is element-wise multiplication. Here, the gated fusion layer can fully retain the meaningful features and effectively suppress undistinctive features of the two layers.

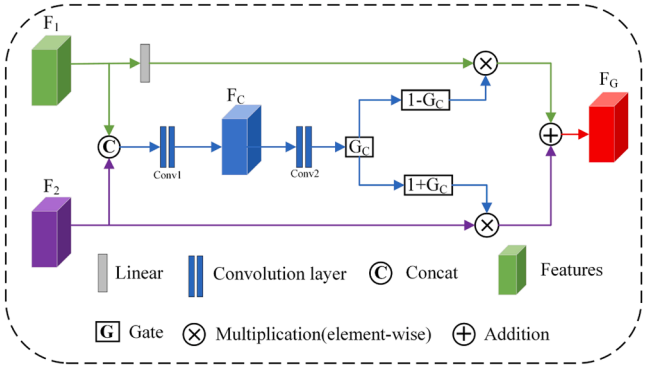


Fig. 3. Gated fusion layer.

2.4.4. Attention-based MIL pooling

In each branch, patch features at two magnifications are aggregated into slide-level representations through gated attention-based MIL pooling operation [44]. This operation provides each node with learnable attention weights, which not only assists the model to selectively aggregate the features of the patches to form an efficient representation of the slide, but also obtains interpretable heatmaps based on the attention scores. Therefore, slide-level feature representations can be expressed as:

$$z_{slide}^s = \sum_{j=1}^{P_s} \widehat{a}_j^s \widehat{h}_j \quad \left(z_{slide}^s = \left\{z_{slide}^{20\times} \; or \quad z_{slide}^{40\times}\right\}, z_{slide}^s \in \mathbb{R}^{1\times 256}\right) \tag{12}$$

Where $\widehat{h}_j$is the final feature vector of the jth instance in slide X.

$$\widehat{a}_j^s = \frac{\exp\left\{w'^T\left(\tanh\left(V^T \widehat{h}_j^T\right) \odot \mathrm{sigm}\left(U\widehat{h}_j^T\right)\right)\right\}}{\sum_{j=1}^{P_s} \exp\left\{w'^T\left(\tanh\left(V\widehat{h}_j^T\right) \odot \mathrm{sigm}\left(U\widehat{h}_j^T\right)\right)\right\}} \tag{13}$$

Where $\widehat{a}_j^s$ denote the attention score of the jth patch in the slide X under scale s.$w' \in \mathbb{R}^{256\times 1}$,$V \in \mathbb{R}^{256\times 128}$and $U \in \mathbb{R}^{256\times 128}$ are trainable parameters. tanh (·) and sigm(·) are nonlinear activation function.

The attention-based MIL pooling can assign attention weights to each instance to locate top-ranked patches and regions of interest (ROIs) at each scale. It also provides interpretable heatmaps for each slide according to the x-y coordinates of the patches.

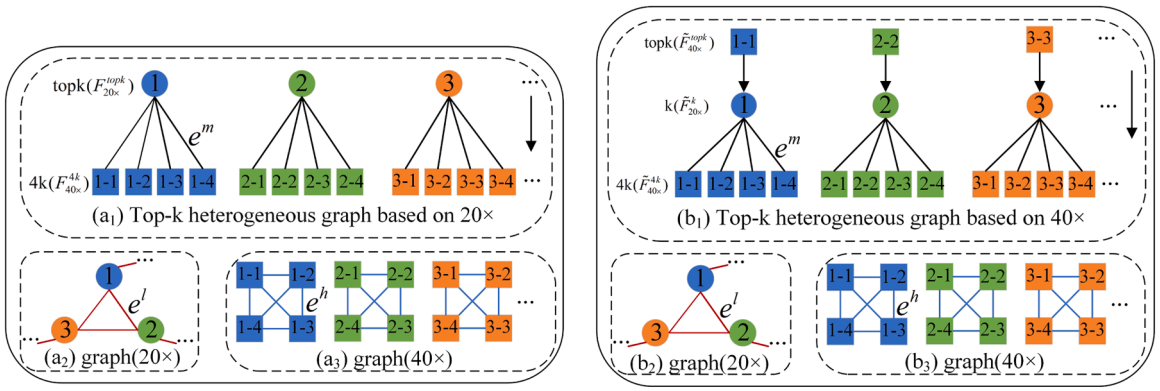
2.4.5. Top2k heterogeneous subgraph (Top2k-HSG)

In clinical scenarios, pathologists draw comprehensive diagnostic conclusions based on observing macroscopic information of pathological slides at low resolution and microscopic features at high resolution. This is due to the fact that patches in high magnification have heterogeneous correlations with their low-magnification counterparts, especially in critical regions. Therefore, a top2k heterogeneous subgraph (Top2k-HSG) is introduced across two magnification scales to assist the prediction of the dual-stream model. The architecture of Top2k-HSG is illustrated in Fig. 4. As shown in Fig. 4, Top2k-HSG consists of two top-k heterogeneous subgraphs, which are obtained according to the attention scores of patches in two parallel branches.

Fig. 4(a) is the first top-k heterogeneous subgraph, which is established based on the top-ranked instances at scale 20 × . Specifically, top-k instances are obtained and highlighted as the positive areas according to the attention scores of branch 1(20 ×). Then the corresponding patches at 40 × are localized using the downward mapping operation $L_{xy\downarrow}$:

$$\left\{F_{40\times}^{4k}\right\} = L_{xy\downarrow}\left(\left\{F_{20\times}^{topk}\right\}\right) \tag{14}$$

Where $\{F_{20\times}^{topk}\}$ denote the feature representations of highlighted top-k image patches at 20 × . Here, $F_{20\times}^{topk} \in \mathbb{R}^{k\times 256}$. $L_{xy\downarrow}$(·)indicates a downward mapping operation, which returns the features corresponding to 4k image patches at 40 × magnification according to the x-y coordinates. Therefore, $\{F_{40\times}^{4k}\}$ are the feature representations of image patches at 40 × . Note that the number of patches in 40 × is 4k via one-to-four



(a) Top-k heterogeneous subgraph based on the nodes at 20×.

(b) Top-k heterogeneous subgraph based on the nodes at 40×.

Fig. 4. Top2k-HSG. Squares indicate the nodes at 40 × , while circles denote the nodes at 20 × .e^l(red line), e^h(blue line) and e^m(black line) denote three types of edges in top2k-HSG.

mapping function. Thus, each element can be written as $F^{4k}_{40\times} = \{F^k_{40\times}, F^k_{40\times}, F^k_{40\times}, F^k_{40\times}\}, F^k_{40\times} \in \mathbb{R}^{k\times 256}$.

Afterwards, all the obtained image patches at scales 20 × and 40 × are utilized to form the heterogeneous subgraph, which includes two types of nodes and three types of edges. Here, the feature representations of image patches $\{F^{topk}_{20\times}\}$ and $\{F^{4k}_{40\times}\}$ are treated as the heterogeneous nodes of the topk subgraph.

Therefore, the total number of nodes in the first top-k heterogeneous subgraph is 5*k*. As shown in Fig. 4(a), the heterogeneous subgraph also includes three types of edges e^l, e^h and e^m. Here, e^l and e^h denote the connections of homogeneous nodes at 20 × and 40 × , respectively. They can also be viewed as correlations between the homogeneous patches at the macro-histological level and the micro-histological level, respectively. Whereas e^m indicates the cross-scale connections between heterogeneous nodes. Thus, the adjacency matrices at each scale can be expressed as $A^l_1 \in \{0,1\}^{k\times k}$and $A^h_1 \in \{0,1\}^{4k\times 4k}$, which are used to represent the connections of top-k instances in 20 × and connections of mapped 4*k* instances in 40 × through *KNN* algorithm, respectively. Furthermore, heterogeneous adjacency matrix $A^m_1 \in \{0,1\}^{k\times 4k}$ is also applied to denote the contextual relationship of patches across scales. Therefore, the top-k heterogeneous graph based on 20 × can be expressed as: $g^k \in \{F^{topk}_{20\times}, F^{4k}_{40\times}, A^l_1, A^h_1, A^m_1\}$.

Meanwhile, to fully exploit the heterogeneous information from the two branches, the top-k instances in scale 40 × are also selected and localized according to the attention scores of gated hypergraph convolution network in branch 2. As shown in Fig. 4(b_1), the squares in the top row indicate the obtained top-ranked patches in the branch 2. The circles represent the corresponding parent patches at 20 × . These patches are localized through an upward mapping operation:

$$\{\tilde{F}^k_{20\times}\} = L_{xy\uparrow}(\{\tilde{F}^{topk}_{40\times}\}) \tag{15}$$

Where $\tilde{F}^{topk}_{40\times} \in \mathbb{R}^{k\times 256}$ denote the feature of the top-k instances in the gated hypergraph convolutional network.$L_{xy\uparrow}(\cdot)$ is also a location mapping operation, which returns the feature of parent node at 20 × according to the *x*-*y* coordinates. Therefore, $\{\tilde{F}^k_{20\times}\}$ indicate the features of the parent patches at 20 × . The subsequent process is exactly similar as that of g^1. The patches at 40 × are localized again using the downward mapping operation $L_{xy\downarrow}$:

$$\{\tilde{F}^{4k}_{40\times}\} = L_{xy\downarrow}(\{\tilde{F}^k_{20\times}\}) = \{\tilde{F}^k_{40\times}, \tilde{F}^k_{40\times}, \tilde{F}^k_{40\times}, \tilde{F}^k_{40\times}\} \tag{16}$$

$\{\tilde{F}^{4k}_{40\times}\}$ denote the feature representations of the upscaled image patches at 40 × , the total number of which is also 4*k*. Note that these 4*k* nodes include the original topk node features located based on scale 40 × .

Likewise, the adjacent matrix is also obtained based on the approach in Fig. 4(a), which can be denoted as A^l_2, A^h_2andA^m_2. Therefore, the second top-k heterogeneous graph based on 40 × can be expressed as $g^{k'} \in \{\tilde{F}^k_{20\times}, \tilde{F}^{4k}_{40\times}, A^l_2, A^h_2, A^m_2\}$.

To avoid information redundancy, these two top-k subgraphs are combined to form a Top2k-HSG and denote as $g^{2k} \in \{g^k, g^{k'}\} = \{F^{2k}_{20\times}, F^{8k}_{40\times}, A^l, A^m, A^h\}$. Here,$F^{2k}_{20\times} = \{F^{topk}_{20\times}, \tilde{F}^k_{20\times}\}$ and $F^{8k}_{40\times} = \{F^{4k}_{40\times}, \tilde{F}^{4k}_{40\times}\}$ are treated as heterogeneous nodes in the Top2k-HSG.

It is worth noting that Top2k-HSG not only establishes correlations between key patches at each scale, but also introduces significant spatial heterogeneous relationships between these patches to guide the HSG-MGAF Net for adaptive prediction and ROI localization.

Heterogeneous Subgraph Attention Network For the constructed Top2k-HSG, a two-layer Heterogeneous Subgraph Attention Network (HGAT) is applied to simulate multiple types of nodes and edges in g^{2k}, as it leverages the heterogeneous graphs [45] to comprehensively learn context-aware correlations of the heterogeneous nodes. Here, the hierarchical attention mechanism can be divided into node-level attention and semantic-level attention.

Firstly, node-level attention is introduced to learn the correlation between the nodes based on a specific edge type, and then aggregates the features of neighborhoods to form node embeddings. Specifically, the feature representation of the node in g^{2k} can be denoted as $\tilde{h}_j \in \mathbb{R}^{1\times 256}$. Then, the feature embedding of the node can be obtained based on various edge types $\tilde{e}$. Therefore, it can be expressed as:

$$\tilde{h}^{\tilde{e}}_j = \sigma\left(\sum_{j\in Ne^{\tilde{e}}(\tilde{h}_j)} a^{\tilde{e}}_{jj}\tilde{h}_j\right) \tag{17}$$

$$a^{\tilde{e}}_{jj} = \frac{\exp\left(\sigma\left(a^{\mathrm{T}}_{\tilde{e}}\left[\tilde{h}_j \parallel \tilde{h}_j\right]\right)\right)}{\sum_{j\in Ne^{\tilde{e}}(\tilde{h}_j)}\exp\left(\sigma\left(a^{\mathrm{T}}_{\tilde{e}}\left[\tilde{h}_j \parallel \tilde{h}_j\right]\right)\right)} \tag{18}$$

Where $a^{\tilde{e}}_{jj}$ denotes the weight coefficient of a node pair $(\tilde{h}_j, \tilde{h}_j)$ based on the edge type $\tilde{e}$. $\tilde{h}_j, \tilde{h}_j \in \mathbb{R}^{1\times 256}$ are the feature representations of the node and its neighbor in Top2k-HSG. $Ne^{\tilde{e}}(\tilde{h}_j)$denotes the neighbors of node $\tilde{h}_j$based on the edge type $\tilde{e}$. ∥ represents the concatenation operation. $a_{\tilde{e}}$ is the node-level attention vector based on a specific edge type $\tilde{e}$. In our Top2k-HSG, each node contains two types of edges denoted as $\tilde{e} = \{\tilde{e}_1, \tilde{e}_2\}$ ($\tilde{e}_1 \in \{e^l\}$or $\{e^h\}$, $\tilde{e}_2 \in \{e^m\}$).Thus, the node embedding in Top2k-HSG is updated and denoted as $\tilde{h}'_j = \{\tilde{h}^{\tilde{e}_1}_j, \tilde{h}^{\tilde{e}_2}_j\}$. Here, $\tilde{h}^{\tilde{e}_1}_j$ and

$\widetilde{h}_j^{\widetilde{e}_2}$ stand for the node feature representations based on edge type $\widetilde{e}_1$ and $\widetilde{e}_2$, respectively. Since node features obtained from heterogeneous edges are feature embeddings in different feature spaces, a space transformation matrix M is designed to project edge type-specific features into the same feature space.

$$\widetilde{h}_j^{'\widetilde{e}_1} = M_{\widetilde{e}_1} \cdot \widetilde{h}_j^{\widetilde{e}_1}, \quad \widetilde{h}_j^{'\widetilde{e}_2} = M_{\widetilde{e}_2} \cdot \widetilde{h}_j^{\widetilde{e}_2} \tag{19}$$

Where $M_{\widetilde{e}_1}$ and $M_{\widetilde{e}_2}$ are space transformation matrices.

Then, semantic-level feature embedding is performed to adaptively fuse the features of $\widetilde{h}_j^{'\widetilde{e}_1}$ and $\widetilde{h}_j^{'\widetilde{e}_2}$ based on the learned weights of the two edge types ($\widetilde{e}_1$ and $\widetilde{e}_2$). The weights of node feature embeddings based on each edge type can be expressed as

$$\left[\omega_j^{\widetilde{e}_1}, \omega_j^{\widetilde{e}_2}\right] = Att\left(\widetilde{h}_j^{'\widetilde{e}_1}, \widetilde{h}_j^{'\widetilde{e}_2}\right) \tag{20}$$

Finally, the semantic-level node feature embedding can be given by

$$\widehat{h}_j = \omega_j^{\widetilde{e}_1} h_j^{'\widetilde{e}_1} + \omega_j^{\widetilde{e}_2} h_j^{\widetilde{e}_2} \tag{21}$$

Where$Att(\cdot)$is the attention function, $\omega_j^{\widetilde{e}_1}$ and $\omega_j^{\widetilde{e}_2}$ are the attention weights of the feature $h_j^{\widetilde{e}_1}$ and $h_j^{\widetilde{e}_2}$, respectively.

Multi-scale pooling module. Top2k-HSG contains discriminative features of image patches at both scales after message passing mechanism of heterogeneous nodes. Therefore, aggregating the node features of Top2k-HSG at multiple scales is meaningful for the our HSG-MGAF Net. Here, multi-scale pooling is module introduced to aggregate the heterogeneous node features of Top2k-HSG at two scales into an overall feature representation, which is shown in Fig. 5(a).

Multi-scale pooling module mainly consists of two components: multi-scale attention module (Fig. 5(b)) and global average pooling (*GAP*) operation. As is shown in Fig. 5(b), multi-scale attention is used to handle the latent semantic gap between low-magnification and high-magnification patches, allowing patch features on the two scales to be fused more smoothly. Specifically, the transformed feature of image patches at 20 × is treated as a Query (Q), and the corresponding four image patch representations at 40 × are treated as Keys (K) and Values (V). Then, the contribution of each patch at 40 × to the image patch at 20 × are obtained by attention mechanism. Finally, the image patches at 40 × are aggregated into a feature representation F_{att}^{2k} at 20 × based on their contributions to the parent image patch. This process can be expressed mathematically as

$$Q = F_{20\times}^{2k} W_q, K = F_{40\times}^{8k} W_k, V = F_{40\times}^{8k} W_v \tag{22}$$

$$F_{att}^{2k} = soft\max\left(\frac{Q(K)^T}{\sqrt{d_c}}\right)\cdot(V) = soft\max\left(\frac{\left(F_{20\times}^{2k} W_q\right)\cdot\left(F_{40\times}^{8k} W_k\right)^T}{\sqrt{d_c}}\right)\cdot\left(F_{40\times}^{8k} W_v\right) \tag{23}$$

Where d_c denotes the feature dimension of K. Here, d_c=128. W_q, W_k and W_v are feature transformation matrices, which are trainable parameters of the multi-attention module.

To maintain the spatial relationships of the heterogeneous nodes, *GAP* is performed to aggregate patch features of each scale. After that, the features at two scales are concatenated to obtain the final output feature $Z^{2k} \in \mathbb{R}^{1\times 256}$. Here,

$$Z^{2k} = \left[GAP\left(F_{20\times}^{2k}\right) \parallel GAP\left(F_{att}^{2k}\right)\right] \tag{24}$$

2.4.6. Multi-scale adaptive feature fusion module

The multi-scale adaptive feature fusion module is proposed to fully aggregate the slide-level features at two scales according to the attention score of each branch. As is shown in Fig. 6, we uses the aggregated feature of Top2k-HSG to adaptively guide HSG-MGAF Net to capture

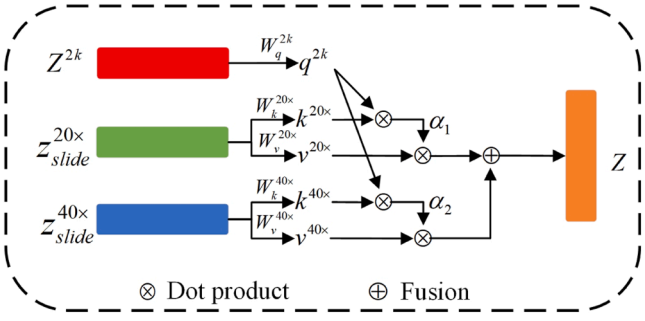


Fig. 6. Multiscale Adaptive Feature Fusion Module.

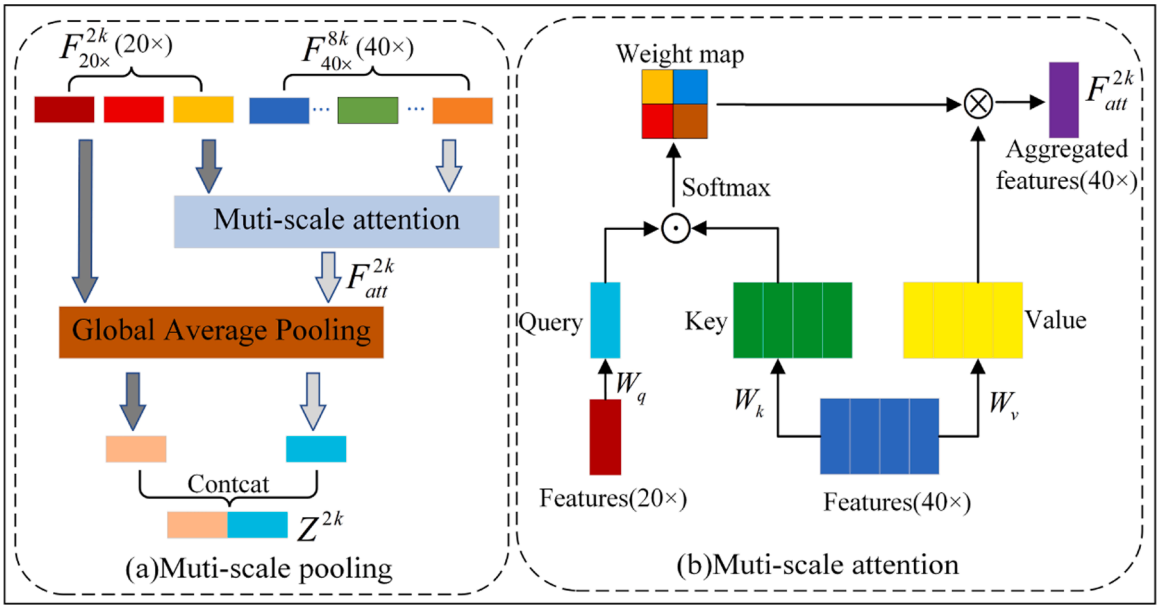


Fig. 5. Multi-scale pooling module. (a) Multi-scale pooling. (b) Multi-scale attention module.

more comprehensive and distinctive features of each branch to improve the performance and efficiency of the model.

Specifically, the attention weights of the two branches are obtained by cross-attention of the aggregated feature representations of Top2k-HSG and slide-level feature representations of each scale. Therefore, the corresponding attention weights can be given by:

$$\alpha_1 = softmax\left(\frac{q^{2k}\left(k^{20\times}\right)^T}{\sqrt{d_1}}\right) = softmax\left(\frac{\left(Z^{2k}W_q^{2k}\right)\cdot\left(z_{slide}^{20\times}\cdot W_k^{20\times}\right)^T}{\sqrt{d_1}}\right) \quad (25)$$

$$\alpha_2 = softmax\left(\frac{q^{2k}\left(k^{40\times}\right)^T}{\sqrt{d_2}}\right) = softmax\left(\frac{\left(Z^{2k}W_q^{2k}\right)\cdot\left(z_{slide}^{40\times}\cdot W_k^{40\times}\right)^T}{\sqrt{d_2}}\right) \quad (26)$$

Where α_1 and α_2 denote the attention weights of the two branches. Z^{2k} indicate the aggregated feature representation of Top2k-HSG. $z_{slide}^{20\times}$ and $z_{slide}^{40\times}$ are slide-level feature representations of WSI at 20 × and 40 × , respectively. d_1 and d_2 denote the feature dimension of $k^{20\times}$ and $k^{40\times}$, respectively. Here, $d_1 = d_2 = 256$. W_q^{2k}, $W_k^{40\times}$ and $W_k^{20\times}$ are also trainable feature transformation matrices of the module.

Then, the output feature is obtained by the weighted sum of the features in scale 20 × and 40 × .

$$Z = \alpha_1\left(v^{20\times}\right) + \alpha_2\left(v^{40\times}\right) = \alpha_1\left(z_{slide}^{20\times}W_v^{20\times}\right) + \alpha_2\left(z_{slide}^{40\times}W_v^{40\times}\right) \quad (27)$$

Where Z denotes the final prediction of the model after the fusion of two branch features, which could selectively retain the most significant features of the two branches. $v^{20\times}$ and $v^{40\times}$ are feature representations of scale 20 × and 40 × . $W_v^{20\times}$and $W_v^{40\times}$ are also trainable parameters.

2.4.7. Model optimization

Drop Edge. For gated hypergraph network branch, the connections between image patches in each slide are redundant because each hyper edge can connect more than two nodes in the constructed hypergraph. Therefore, drop edge [46] is applied to improve the generalization ability and prevent overfitting issues. Here, we drop 20 % of the hyper edges in the hypergraph, which is randomly selected from the constructed hypergraph. This operation can also reduce computation burden and prevent overfitting in the iteration.

Loss Functions. Designing a task-specific loss function is a critical issue for weakly supervised models, as it can drive the model to approximate the optimal solution in a refined manner.

(1) **Cross-entropy Loss.** For our classification task, we adopt conventional cross-entropy loss as a main part of the loss function, which can be written as:

$$L_{CE} = -\log(P_t) \quad (28)$$

Where P_t stands for the class prediction probability of the model, the value of which is ranging from 0 to 1.

(1) **Self-supervised Contrastive Learning Loss**. In clinical scenarios, the diagnostic results obtained by pathologists at low magnification should be generally consistent with those obtained at high magnification. Therefore, self-supervised contrastive learning constraint are introduced on the slide-level feature representations at two scales to ensure the consistency of predictions. Here, we adopt InfoNCE loss [47] as self-supervised contrastive learning loss function and expressed as:

$$L_S = -\log\sigma\left(f_D\left(z_{slide}^{20\times}, z_{slide}^{40\times}\right)\right) - \log\sigma\left(1 - f_D\left(z_{slide}^{20\times}, \tilde{z}_{slide}^{40\times}\right)\right) \quad (29)$$

Where $\tilde{z}_{slide}^{40\times}$ denote the negative samples obtained by corrupting spatial coordinates of instances in $z_{slide}^{40\times}$by row-wise and column-wise shuffling. That is, $\tilde{z}_{slide}^{40\times}$ is a slide-level representation of the WSI at scale 40 × after the instances are spatially shuffled. Therefore, the $z_{slide}^{20\times}$and $z_{slide}^{40\times}$are slide-level feature representations of a specific WSI at scale 20 × and 40 × , which forms a positive sample pair. Whereas $z_{slide}^{20\times}$ and $\tilde{z}_{slide}^{40\times}$ form a negative sample pair. Note that, the slide-level presentation $\tilde{z}_{slide}^{40\times}$ is not unique for a specific $z_{slide}^{20\times}$ in the training process. In the function, $f_D(\cdot,\cdot)$ acts as a discriminator, which takes the slide-level feature vectors of each scale as input and evaluates the agreement score between them through a simple dot-product operation. Thus, the self-supervised contrastive learning loss is to ensure the consistency of slide-level representation $z_{slide}^{20\times}$ and $z_{slide}^{40\times}$ while discriminate $z_{slide}^{20\times}$and $\tilde{z}_{slide}^{40\times}$ by constraint the spatial coordinates of the corresponding instances at both scales. Therefore, it also ensures the interpretability of the patches at two scales. The total loss function L can be written as:

$$L = \beta L_{CE} + (1-\beta)L_S \quad (30)$$

Where β is the weighting factor of the two loss functions.

3. Results

We constructed the experiments on two whole slide image datasets: CAMELYON16 and the Cancer Genome Atlas non-small cell lung cancer (TCGA-NSCLC).

3.1. Datasets description

CAMELYON16: CAMELYON16 is an H&E-stained whole slide image of lymph node metastasis that contains pixel-level annotation and is publicly available. The dataset contains a total of 399 whole slide images of sentinel lymph nodes, which are officially divided to training set and test set. The training set consists of 270 whole slide images, including 111 positive slides and 159 negative slides. To better ensure the reliability of the model, the 270 official training slides was randomly split into training set and validation set again according to the ratio of 8.5:1.5. The official testing set consists of 129 slides containing 49 positive slides and 80 negative slides. In the preprocessing stage, the background in each slide was discarded and the foreground was retained. The foreground is then cropped into image patches with the size of 256 × 256 at 20 × and 40 × , respectively. Meanwhile, the x-y coordinates of these patches are preserved in each scale. The average number of image patches at these two scales is approximately 9066 and 45,253, respectively.

TCGA-NSCLC: The dataset includes a total of 993 whole slide images of two lung carcinoma subtypes: lung squamous cell carcinoma (LUSC) and lung adenocarcinoma (LUAD). The number of LUADs was 507 WSI images from 444 confirmed cases, while the number of LUSCs was 486 WSI images from 452 confirmed cases. The TCGA-NSCLC dataset only provide slide-level labels for H&E-stained whole slide images. In our case, we randomly split it into training, validation and testing sets in the ratio of 65:15:25. Here, the preprocessing procedure is exactly the same as CAMELYON16. Thus, an average of 11,589 and 45,890 non-overlapping patches with sizes of 256 × 256 are obtained at 20 × and 40 × , respectively.

3.2. Implementation details

In the implementation, HSG-MGAF Net follows an end-to-end learning manner with only slide-level labels available for both datasets. For CAMELYON16 dataset, the learning rate and weight decay are set to 1×10^{-4} and 5×10^{-6}, respectively. While for TCGA-NSCLC dataset, the corresponding parameters are set as $0.8 \times 10^{-4}$and 3×10^{-6}, respectively. During the training process, if the loss of the

validation set does not decrease for 5 consecutive epochs, the learning rate will decay by the factor of 0.8. Meanwhile, the early stopping is also performed to prevent overfitting issue. The hyper parameter β is set to 0.8 as the model has to balance classification accuracy and consistency of the predictions in two branches. In both cases, our models use the Adam optimizer and are trained for 100 epochs on a Tesla V100 with a batch size of 1. The hyperparameter k in Top2k-HSG is set as 5 and 10 for CAMELYON16 and TCGA-NSCLC dataset, respectively. The choice of hyperparameter k is empirically based on the statistical features of each data set. In the experiment, we use the area under curve (AUC), accuracy (ACC) and F_1 score to evaluate the performance of the model.

3.3. Results

Table 1 shows the performance of our HSG-MGAF Net and SOTA weakly supervised approaches on both datasets. As shown in Table 1, the overall prediction ACC/AUC/ F_1 score of the HSG-MGAF Net can achieve 92.7%/0.951/0.892 on CAMELYON16, and 92.2%/0.957/0.919 on TCGA-NSCLC after 5 trials, respectively. The results indicated that the HSG-MGAF Net has outperformed other comparative models on both datasets. The performance improvement is higher on CAMELYON16 than that of TCGA-NSCLC. This is mainly because most positive slides in the CAMELYON16 only contain a small portion of the positive area. Therefore, it is more effective to utilize heterogeneous subgraphs of top-ranked instances at multiple scales to guide the model for predictions. While positive slides in the TCGA-NSCLC dataset generally have relatively more positive regions than CAMELYON16. Thus, the performance improvement on the TCGA-NSCLC dataset is relatively limited.

3.4. Discussion

To further illustrate the effectiveness of the HSG-MGAF Net, we divide the SOTA models into 5 main categories for comparison. (i) The existing approaches are mainly based on the instances *i.i.d* assumption supplemented by instance constraints, such as traditional MIL pooling models, ABMIL [44], MIL-RNN [23], CLAM [24] and DTFD-MIL [33]. However, the performances of these frameworks are limited as they ignore the inter-instance dependencies within the slide. (ii) DSMIL [35] simulates the context-aware features of slides by concatenating patch features at multiple scales, which can be regarded as a representative study of local context-based MIL. However, it only considers the correlation between the highest-ranked instance and others after feature embedding at different scales, which cannot fully describe the implicit contextual information between the image patches in the slide. (iii) Currently, the Transformer-based MIL becomes an effective branch for giga-pixel pathology slide prediction. It introduces a self-attention mechanism to emphasize the global relationship between image patches in the slide, which have yielded SOTA performance, such as TransMIL [34]. However, it also calculates attention scores between unrelated instances, which not only suffers from computational redundancy but also causes performance degradation. (iv) Patch-GCN [40], ABMIL-GCN [21] and Graph-Transformer [28] are attempt to establish the interaction between image patches and their surrounding neighbors to describe the slide, which is more applicable in real clinical scenarios. Nevertheless, the multi-scale heterogeneous relationships that exist in slide have still not been represented and utilized. (v) Recently, heterogeneous graph concept has been introduced in whole slide image processing. That is, image patches at each scale are treated as heterogeneous nodes, which are utilized to build a hierarchical heterogeneous graph to simulate the correlation of the heterogeneous nodes at the multiple scales, i.e. H^2-MIL [42]. However, it suffers from huge computational burden. This is because H^2-MIL constructs a heterogeneous graph with the "resolution" attribute to explicitly simulate the feature and spatial-scale relationships of all patches at multiple resolutions. That is, the method not only includes the message passing for all the nodes from homogeneous scale, but also those from the heterogeneous scale. In contrast, our model exploits only the most critical heterogeneous subgraphs instead of all heterogeneous information at both scales. The computational cost of our model is 8.2 GFlops and 5.7 GFlops for CAMELYON16 and TCGA-NSCLC dataset, respectively. While that of the H^2-MIL is 12.7 GFlops and 8.9 GFlops for these two datasets. Therefore, our model can greatly reduce the computational burden while effectively leveraging heterogeneous information compared with H^2-MIL.

While HSG-MGAF Net not only exploits the low-order and potential high-order correlations of image patches at each scale, but also leverages the Top2k-HSG at these scales to guide the model for adaptive prediction and ROI localization. Therefore, it outperforms all the other competing methods, even H^2-MIL. Besides, contrastive learning can be viewed as a spatially adaptive refinement constraint on each node that maintains high predictive consistency at both scales. This feature also improves the interpretability of the obtained heatmap.

Feature Visualization. To gain a deeper understanding of the HSG-MGAF Net, the feature representations of the whole slide images in the latent space are visualized via T-SNE, as shown in Fig. 7. Fig. 7(a)–(c) and (d)–(f) are slide-level features of two datasets at 20 × , 40 × and fusion scales, respectively. Each slide is denoted as a dot in the latent space, where red and blue dots represent the positive slides and negative slides, respectively. In Fig. 7, the results indicate that HSG-MGAF Net with two scales can clearly distinguish two categories in the latent space for both datasets. Compared with models based on a single-scale, they achieve smaller intra-class variance and larger inter-class variance on both datasets. This means that HSG-MGAF Net can effectively capture discriminative heterogeneous information of key patches based on

Table 1
Results on the CAMELYON16 and TCGA-NSCLC dataset (The values highlighted in bold denote the optimal solutions, the values underlined denote the suboptimal solutions).

	CAMELYON16			TCGA-NSCLC		
Method	Accuracy	AUC	F_1 score	Accuracy	AUC	F_1 score
MaxPooling	0.642±0.029	0.701±0.013	0.564±0.018	0.794±0.016	0.851±0.021	0.742±0.011
MeanPooling	0.598±0.031	0.539±0.003	0.315±0.085	0.751±0.022	0.807±0.017	0.725±0.048
ABMIL	0.823±0.011	0.858±0.013	0.710±0.026	0.834±0.017	0.848±0.006	0.821±0.014
MIL-RNN	0.834±0.014	0.857±0.027	0.781±0.013	0.864±0.014	0.865±0.008	0.847±0.007
DSMIL	0.833±0.023	0.878±0.019	0.812±0.022	0.875±0.031	0.897±0.024	0.873±0.017
CLAM-SB	0.817±0.033	0.843±0.028	0.796±0.037	0.871±0.003	0.938±0.021	0.863±0.024
CLAM-MB	0.830±0.009	0.873±0.012	0.730±0.008	0.878±0.043	0.949±0.019	0.874±0.028
TransMIL	0.801±0.027	0.828±0.036	0.773±0.028	0.886±0.019	0.946±0.013	0.868±0.026
DTFD-MIL	0.892±0.014	0.923±0.012	0.862±0.026	0.891±0.017	0.947±0.024	0.883±0.021
Patch-GCN	0.876±0.031	0.918±0.017	0.864±0.027	0.875±0.019	0.935±0.019	0.857±0.027
ABMIL-GCN	0.909±0.003	0.921±0.004	0.875±0.005	0.895±0.011	0.950±0.013	0.891±0.017
Graph-Transformer	0.893±0.026	0.923±0.021	0.876±0.019	0.919±0.011	0.951±0.013	0.901±0.008
H^2-MIL	0.912±0.017	0.924±0.009	0.876±0.017	0.917±0.016	0.953±0.018	0.902±0.014
HSG-MGAF Net(Ours)	**0.927±0.018**	**0.951±0.016**	**0.892±0.019**	**0.922±0.009**	**0.957±0.014**	**0.919±0.012**

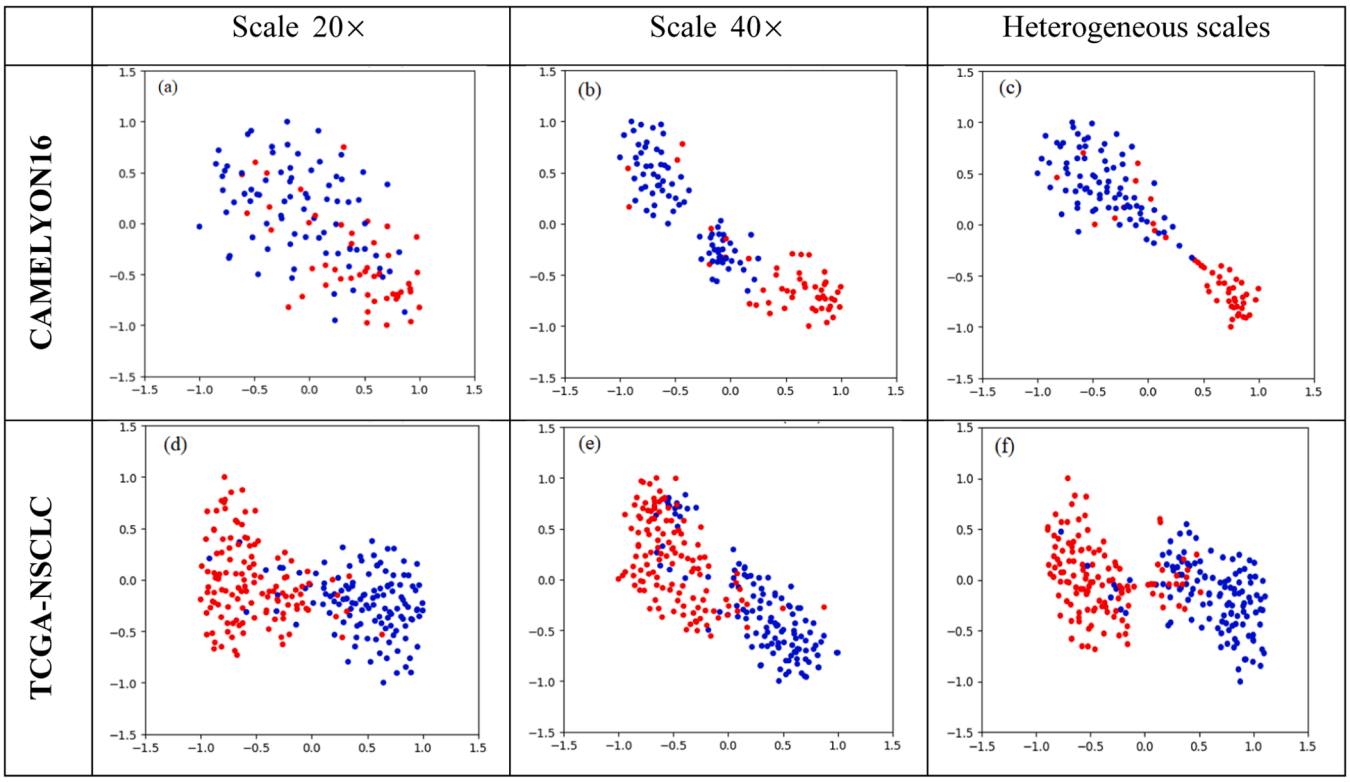


Fig. 7. Slide-level feature embedding for CAMELYON16 and TCGA-NSCLC using T-SNE. In (a)-(c), red dots represent the positive and blue dots represent the negative. In (d)-(f), red dots and blue dots represent LUAD and LUAC, respectively.

multi-scale hierarchical graph architecture for high-precision prediction.

Interpretability. The HSG-MGAF Net can not only establish the intrinsic low-order and potentially high-order nonlinear relationships between instances at two independent scales to obtain slide-level predictions, but also leverage the heterogeneous correlations of the highly ranked instances in two scales to guide the HSG-MGAF Net for adaptive ROI localization. Therefore, it can provide more interpretability from both macro morphological and microscopic features perspectives. Here, the attention weights of patches at each scale are obtained through attention-based MIL pooling modules of each branch. Specifically, this module achieves slide-level feature representation by adaptively aggregating the features of each patch. In this process, the weight of each patch can also be obtained according to Eq. (13). Subsequently, these weights are then converted into attention scores and mapped into the input space. Finally, the interpretable heatmaps on each scale can be obtained by index matching the patches with the attention scores. Here, we use red to highlight patches with high attention scores and blue to indicate patches with low attention scores. Fig. 8 shows the input slides with pixel-wise annotations, highlighted interpretable heatmaps at both scales, and the top-k instances selected from the ROIs, respectively. In Fig. 8 column 2 and 3, the results indicated that the obtained heat maps could accurately delineate the exact tumor boundaries, which were highly consistent with human pathology expertise. Meanwhile, as shown in Fig. 8(d), (h) and (l),the ROIs obtained from heatmaps at 20 × and 40 × have high prediction consistency, which demonstrated the effectiveness of the model. Furthermore, heatmaps at 20 × and 40 × can effectively complement each other when encountering uncertain patterns. Specifically, the patches at 20 × capture macroscopic infiltration between heterogeneous tissues, while the patches at 40 × provides microscopic evidence of the presence of cancer cells. This can also be illustrated in the subfigures in columns 2–4 of Fig. 8. This is due to the fact that the message passing mechanism between heterogeneous nodes in Top2k-HSG can refine the feature representation of every node in heterogeneous subgraph, thereby enhancing feature expression ability of each node. That is, the node feature is more discriminative and deterministic after heterogeneous message passing. This is similar to adding adaptive patch-wise pseudo labels to WSI at each scale. Meanwhile, the deterministic highlighted nodes at each scale will further promote the accuracy of node feature representation at homogeneous scales through the massage passing mechanism. Therefore, this iterative process can improve the interpretability of the model. The result in Fig. 8 (d), (h) and (l) also demonstrate that the heat map obtained under the high-magnification branch is a refinement and supplement of the result under the low-magnification branch.

However, HSG-MGAF Net still has limitations in selecting fixed number of instances to construct Top2k-HSG for all slides. In real scenarios, the spatial density of positive instances in each slide is not always the same. While the message passing through the heterogeneous subgraph with a fixed number of instances will lead to failure predictions in some cases. For instances, the Fig. 9 is a positive slide with a small portion of positive areas whereas the HSG-MGAF Net predicted as negative. This is because the hyper parameter k is inappropriate in this case. Selecting the top k instances will result in the introduction of negative instances. Therefore, the message passing of the nodes in Top2k-HSG will not improve the feature representation of these nodes. On the contrary, it will weaken the feature representation of the connected instances at 40 × through heterogeneous information passaging. Thus, the corresponding zoomed in areas in Fig. 9(b) and (c) are present false negative results (black arrow) compared to the pixel-wise annotation in Fig. 9(a). In the future, we will attempt to develop an adaptive heterogeneous feature selection algorithm for further improve the performance of the model.

3.5. Ablation study

To verify the effectiveness of proposed components in our HSG-MGAF Net, we conduct a series of ablation studies on both datasets,

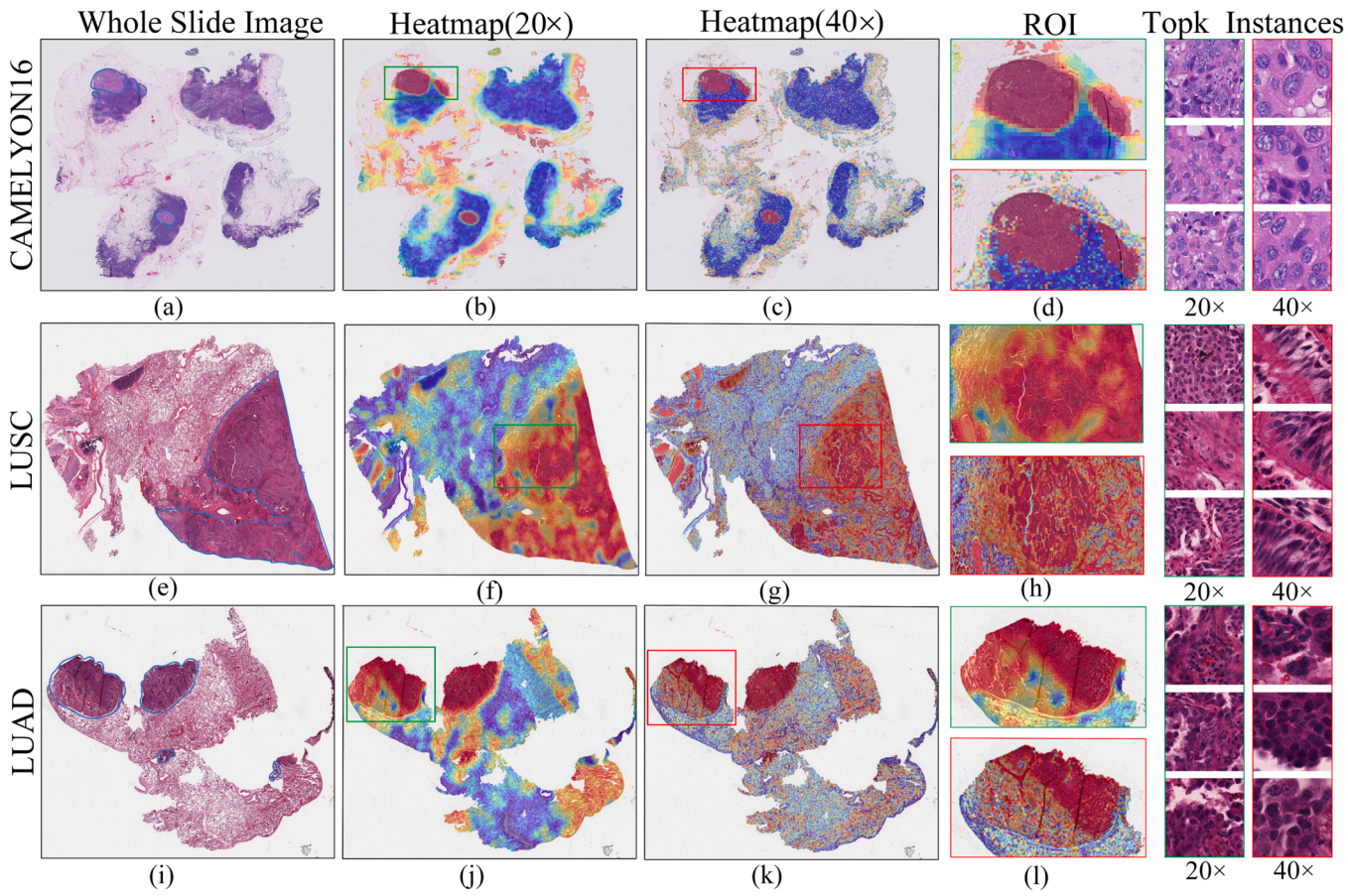


Fig. 8. Interpretable heatmaps at scale 20 × and 40 × . The subfigures in the first column are the input slides and pixel-level annotations. The subfigures in second and third columns present the corresponding attention heatmaps generated by HSG-MGAF Net at scale 20 × and 40 × , respectively. The subfigures in the fourth and fifth columns show the zoomed-in ROIs and highlighted top-k instances at scales 20 × and 40 × , respectively.

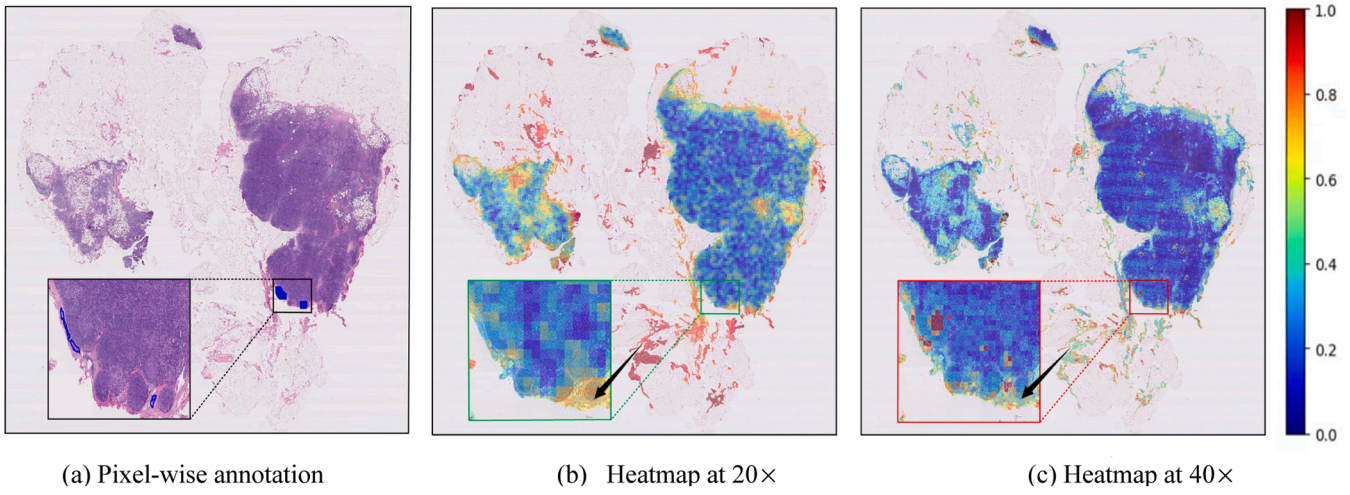


Fig. 9. Failure case in WSI prediction.

such as model without (w/o) Top2k-HSG, Topk-HSG, gated fusion layer (GFL) and contrastive loss (L_S). These results are shown in Table 2. From Table 2, the result indicated that HSG-MGAF Net outperforms the baseline model by approximately 4.3 and 5.5 % for CAMELYON16 and TCGA-NSCLC dataset, respectively. Furthermore, the results also showed that the performance of the model is degrade without top-K subgraph, and further deteriorates without top-2k heterogeneous subgraph. This is because the top-k heterogeneous subgraph of branch 1 is obtained by gradient backpropagation based on the graph structure, while another top-k heterogeneous subgraph of branch 2 the is obtained based on gradient backpropagation based on the hypergraph framework. Note that, the top-k heterogeneous subgraph from branch 1 and branch 2 have some complementary properties. Therefore, these features are both significant for WSI images. The heterogeneous subgraph generated by using only one of the branches will cause the loss of critical information in the other branch. Therefore, it can be concluded that

Table 2
Ablation study for HSG-MGAF Net (The values highlighted in bold denote the optimal solutions).

	CAMELYON16			TCGA-NSCLC		
Model	Accuracy	AUC	F_1 score	Accuracy	AUC	F_1 score
Baseline	0.884±0.013	0.919±0.015	0.839±0.024	0.867±0.018	0.932±0.007	0.862±0.031
w/o GFL	0.923±0.023	0.942±0.017	0.883±0.014	0.911±0.009	0.943±0.013	0.872±0.021
w/o L_S	0.915±0.011	0.930±0.008	0.879±0.019	0.915±0.014	0.934±0.028	0.902±0.016
w/o Top2k-HSG	0.907±0.026	0.931±0.018	0.870±0.017	0.899±0.024	0.938±0.013	0.884±0.015
w/o Topk-HSG	0.926±0.020	0.946±0.019	0.891±0.022	0.914±0.029	0.953±0.016	0.889±0.020
Ours	**0.927±0.018**	**0.951±0.016**	**0.892±0.019**	**0.922±0.009**	**0.957±0.014**	**0.919±0.012**

Top2k-HSG module fully exploits the significant heterogeneity information of highly ranked instances at both scales to guide the model in each branch to capture more discriminative and fine-grained information for further prediction. Meanwhile, the result in Table 2 showed that contrastive learning is also a significant factor in our HSG-MGAF Net. Since self-supervised contrastive learning aims to maximize the consistency of predictions at two scales for each slide, thereby reducing the occurrence of false negative and false positive instances. This feature can also greatly improve the interpretability of heatmaps. While for GFL, its function is to fuse features in deep layer and shallow layer to enhance the representation of meaningful information and suppress noise. Therefore, the performance of GFL is relatively limited compared with Top2k-HSG module and self-supervised contrastive learning loss.

4. Conclusions

In this paper, we propose an HSG-MGAF Net, which build the heterogeneous subgraph topology of critical image patches at two scales for adaptive slide label prediction and ROI localization using only slide-level labels. It not only fully exploits the low-order and potential high-order correlations of image patches though graph and hypergraph architecture in two independent scales, but also leverages the existed heterogeneous information of these scales to guide the backbone network for high-precision analysis and prediction. Meanwhile, Self-Supervised contrastive learning is introduced across the two scales to maximize the consistency of predictions on both scales. The result demonstrated that the Top2k-HSG can achieve significant performance improvement with relatively less computational cost, which lays the foundation for heterogeneous graph applications in computation pathology.

Statements of ethical approval

This work was based on the open-source public dataset CAMELYON16 (https://camelyon17.grand-challenge.org/Data) and The Cancer Genome Atlas diagnostic whole-slide data (NSCLC) (TCGA, https://www.cancer.gov/tcga). The patients involved in the database have given ethical approval.

CRediT authorship contribution statement

Meiyan Liang: Writing – review & editing, Visualization, Methodology, Funding acquisition, Conceptualization. **Xing Jiang:** Writing – original draft, Software, Methodology, Formal analysis. **Jie Cao:** Supervision, Methodology. **Shupeng Zhang:** Data curation. **Haishun Liu:** Methodology, Investigation. **Bo Li:** Visualization. **Lin Wang:** Visualization, Investigation. **Cunlin Zhang:** Writing – review & editing, Conceptualization. **Xiaojun Jia:** Supervision.

Declaration of competing interest

The authors declared that they have no conflicts of interest in this work.

Acknowledgments

This work is supported in part by National Natural Science Foundation of China under Grant 11804209, Research Project Supported by Shanxi Scholarship Council of China No. 2023-010, Natural Science Foundation of Shanxi Province under Grant 202303021211014, 20210302123411 and 202203021222015, China Postdoctoral Science Foundation 2023M742577.

References

[1] M.N. Gurcan, L.E. Boucheron, A. Can, A. Madabhushi, N.M. Rajpoot, B. Yener, Histopathological image analysis: a review, IEEE Rev. Biomed. Eng. 2 (2009) 147–171. https://ieeexplore.ieee.org/abstract/document/5299287.

[2] E. Abels, L. Pantanowitz, F. Aeffner, M.D. Zarella, J. van der Laak, M.M. Bui, C. Kozlowski, Computational pathology definitions, best practices, and recommendations for regulatory guidance: a white paper from the digital pathology association, J. Pathol. 249 (3) (2019) 286–294. https://pathsocjournals.onlinelibrary.wiley.com/doi/full/10.1002/path.5331.

[3] S. Deng, X. Zhang, W. Yan, E.I.C. Chang, Y. Fan, M. Lai, Y. Xu, Deep learning in digital pathology image analysis: a survey, Front. Med. 14 (2020) 470–487. https://link.springer.com/article/10.1007/s11684-020-0782-9.

[4] L. He, L.R. Long, S. Antani, G.R. Thoma, Histology image analysis for carcinoma detection and grading, Comput. Methods Progr. Biomed. 107 (3) (2012) 538–556. https://www.sciencedirect.com/science/article/pii/S0169260711003245.

[5] 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. https://arxiv.org/abs/2012.03583.

[6] L. Duran-Lopez, J.P. Dominguez-Morales, A.F. Conde-Martin, S. Vicente-Diaz, A. Linares-Barranco, PROMETEO: a CNN-based computer-aided diagnosis system for WSI prostate cancer detection, IEEE Access 8 (2020) 128613–128628. https://ieeexplore.ieee.org/abstract/document/9139241.

[7] A. Echle, N.T. Rindtorff, T.J. Brinker, T. Luedde, A.T. Pearson, J.N. Kather, Deep learning in cancer pathology: a new generation of clinical biomarkers, Br. J. Cancer 124 (4) (2021) 686–696. https://www.nature.com/articles/s41416-020-01122-x.

[8] M.G. Hanna, A. Parwani, S.J. Sirintrapun, Whole slide imaging: technology and applications, Adv. Anat. Pathol. 27 (4) (2020) 251–259. https://www.ingentaconnect.com/content/wk/adapa/2020/00000027/00000004/art00005.

[9] S. Maksoud, K. Zhao, P. Hobson, A. Jennings, B.C. Lovell, Sos: selective objective switch for rapid immunofluorescence whole slide image classification, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 3862–3871. https://openaccess.thecvf.com/content_CVPR_2020/html/Maksoud_SOS_Selective_Objective_Switch_for_Rapid_Immunofluorescence_Whole_Slide_Image_CVPR_2020_paper.html.

[10] H.R. Tizhoosh, L. Pantanowitz, Artificial intelligence and digital pathology: challenges and opportunities, J. Pathol. Inform. 9 (1) (2018) 38. https://www.sciencedirect.com/science/article/pii/S2153353922003510.

[11] H. Cai, W. Yi, Y. Li, W. Liao, J. Song, A regional multiple instance learning network for whole slide image segmentation, in: Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine (BIBM), IEEE, 2022, pp. 922–928. https://ieeexplore.ieee.org/abstract/document/9995017.

[12] X. Wang, H. Chen, C. Gan, H. Lin, Q. Dou, E. Tsougenis, P.A. Heng, Weakly supervised deep learning for whole slide lung cancer image analysis, IEEE Trans. Cybern. 50 (9) (2019) 3950–3962. https://ieeexplore.ieee.org/abstract/document/8822590.

[13] J. Chen, H.M. Cheung, L. Milot, A.L. Martel, AMINN: autoencoder-based multiple instance neural network improves outcome prediction in multifocal liver metastases, in: Proceedings of the Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, Springer International Publishing, 2021, pp. 752–761. September 27–October 1, 2021, Proceedings, Part V 24, https://link.springer.com/chapter/10.1007/978-3-030-87240-3_72.

[14] P. Bandi, O. Geessink, Q. Manson, M. Van Dijk, M. Balkenhol, M. Hermsen, G. Litjens, From detection of individual metastases to classification of lymph node status at the patient level: the camelyon17 challenge, IEEE Trans. Med. Imaging 38 (2) (2018) 550–560. https://ieeexplore.ieee.org/abstract/document/8447230.

[15] K. Das, S. Conjeti, J. Chatterjee, D. Sheet, Detection of breast cancer from whole slide histopathological images using deep multiple instance CNN, IEEE Access 8 (2020) 213502–213511. https://ieeexplore.ieee.org/abstract/document/9269335.
[16] L. Liu, C. Li, Comparative study of deep learning models on the images of biopsy specimens for diagnosis of lung cancer treatment, J. Radiat. Res. Appl. Sci. 16 (2) (2023) 100555. https://www.sciencedirect.com/science/article/pii/S168785072300033X.
[17] T.S. Sheikh, J.Y. Kim, M. Cho, Refined attention module for WSI cancer diagnosis, in: Proceedings of the IEEE 5th International Conference on Knowledge Innovation and Invention (ICKII), IEEE, 2022, pp. 30–34. https://ieeexplore.ieee.org/abstract/document/9983555.
[18] J. Vorndran, C. Neuner, R. Coras, L. Hoffmann, S. Geffers, J. Honke, S. Jabari, A deep learning-based histopathology classifier for focal cortical dysplasia, Neural Comput. Appl. 35 (17) (2023) 12775–12792. https://link.springer.com/article/10.1007/s00521-023-08364-9.
[19] A. De, R. Mhatre, M. Tiwari, A.S. Chowdhury, Brain tumor classification from radiology and histopathology using deep features and graph convolutional network, in: Proceedings of the 26th International Conference on Pattern Recognition (ICPR), IEEE, 2022, pp. 4420–4426. https://ieeexplore.ieee.org/abstract/document/9956229.
[20] R. Li, J. Yao, X. Zhu, Y. Li, J. Huang, Graph CNN for survival analysis on whole slide pathological images, in: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Cham, Springer International Publishing, 2018, pp. 174–182. https://link.springer.com/chapter/10.1007/978-3-030-00934-2_20.
[21] M. Liang, Q. Chen, B. Li, L. Wang, Y. Wang, Y. Zhang, C. Zhang, Interpretable classification of pathology whole-slide images using attention based context-aware graph convolutional neural network, Comput. Methods Progr. Biomed. 229 (2023) 107268. https://www.sciencedirect.com/science/article/pii/S0169260722006496.
[22] J. Xiang, X. Wang, X. Wang, J. Zhang, S. Yang, W. Yang, Y. Liu, Automatic diagnosis and grading of Prostate Cancer with weakly supervised learning on whole slide images, Comput. Biol. Med. 152 (2023) 106340. https://www.sciencedirect.com/science/article/pii/S0010482522010484.
[23] G. Campanella, M.G. Hanna, L. Geneslaw, A. Miraflor, V. Werneck Krauss Silva, K. J. Busam, T.J. Fuchs, Clinical-grade computational pathology using weakly supervised deep learning on whole slide images, Nat. Med. 25 (8) (2019) 1301–1309. https://www.nature.com/articles/s41591-019-0508-1.
[24] M.Y. Lu, D.F. Williamson, T.Y. Chen, R.J. Chen, M. Barbieri, F. Mahmood, Data-efficient and weakly supervised computational pathology on whole-slide images, Nat. Biomed. Eng. 5 (6) (2021) 555–570. https://www.nature.com/articles/s41551-020-00682-w.
[25] Su, Z., Rezapour, M., Sajjad, U., Gurcan, M.N., & Niazi, M.K.K. (2023). Attention2Minority: a salient instance inference-based multiple instance learning for classifying small lesions in whole slide images. arXiv preprint arXiv: 2301.07700. https://arxiv.org/abs/2301.07700.
[26] Z. Huang, H. Chai, R. Wang, H. Wang, Y. Yang, H. Wu, Integration of patch features through self-supervised learning and transformer for survival analysis on whole slide images, in: Proceedings of the Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, Springer International Publishing, 2021, pp. 561–570. September 27–October 1, 2021, Proceedings, Part VIII 24, https://link.springer.com/chapter/10.1007/978-3-030-87237-3_54.
[27] W. Li, V.D. Nguyen, H. Liao, M. Wilder, K. Cheng, J. Luo, Patch transformer for multi-tagging whole slide histopathology images, in: Proceedings of the Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, Springer International Publishing, 2019, pp. 532–540. October 13–17, 2019, Proceedings, Part I 22, https://link.springer.com/chapter/10.1007/978-3-030-32239-7_59.
[28] Y. Zheng, R.H. Gindra, E.J. Green, E.J. Burks, M. Betke, J.E. Beane, V. B. Kolachalama, A graph-transformer for whole slide image classification, IEEE Trans. Med. Imaging 41 (11) (2022) 3003–3015. https://ieeexplore.ieee.org/abstract/document/9779215.
[29] N. Hashimoto, D. Fukushima, R. Koga, Y. Takagi, K. Ko, K. Kohno, I. Takeuchi, 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, 2020, pp. 3852–3861. https://openaccess.thecvf.com/content_CVPR_2020/html/Hashimoto_Multi-scale_Domain-adversarial_Multiple-instance_CNN_for_Cancer_Subtype_Classification_with_Unannotated_CVPR_2020_paper.html.
[30] K. Thandiackal, B. Chen, P. Pati, G. Jaume, D.F. Williamson, M. Gabrani, O. Goksel, Differentiable zooming for multiple instance learning on whole-slide images, in: Proceedings of the European Conference on Computer Vision, Cham, Springer Nature Switzerland, 2022, pp. 699–715. https://link.springer.com/chapter/10.1007/978-3-031-19803-8_41.
[31] H. Tokunaga, Y. Teramoto, A. Yoshizawa, R. Bise, Adaptive weighting multi-field-of-view CNN for semantic segmentation in pathology, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 12597–12606. https://openaccess.thecvf.com/content_CVPR_2019/html/Tokunaga_Adaptive_Weighting_Multi-Field-Of-View_CNN_for_Semantic_Segmentation_in_Pathology_CVPR_2019_paper.html.
[32] X. Wang, H. Chen, C. Gan, H. Lin, Q. Dou, Q. Huang, P.A. Heng, Weakly supervised learning for whole slide lung cancer image classification, in: Proceedings of the Medical Imaging with Deep Learning, 2022. https://openreview.net/forum?id=SJwod1hjz.
[33] H. Zhang, Y. Meng, Y. Zhao, Y. Qiao, X. Yang, S.E. Coupland, Y. Zheng, DTFD-MIL: double-tier feature distillation multiple instance learning for histopathology whole slide image classification, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 18802–18812. https://openaccess.thecvf.com/content/CVPR2022/html/Zhang_DTFD-MIL_Double-Tier_Feature_Distillation_Multiple_Instance_Learning_for_Histopathology_Whole_CVPR_2022_paper.html.
[34] Z. Shao, H. Bian, Y. Chen, Y. Wang, J. Zhang, X. Ji, Transmil: transformer based correlated multiple instance learning for whole slide image classification, Adv. Neural Inf. Process Syst. 34 (2021) 2136–2147, in: https://proceedings.neurips.cc/paper_files/paper/2021/hash/10c272d06794d3e5785d5e7c5356e9ff-Abstract.html.
[35] B. Li, Y. Li, K.W. Eliceiri, Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 14318–14328. https://openaccess.thecvf.com/content/CVPR2021/html/Li_Dual-Stream_Multiple_Instance_Learning_Network_for_Whole_Slide_Image_Classification_CVPR_2021_paper.html.
[36] G. Jaume, P. Pati, B. Bozorgtabar, A. Foncubierta, A.M. Anniciello, F. Feroce, O. Goksel, Quantifying explainers of graph neural networks in computational pathology, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 8106–8116. https://openaccess.thecvf.com/ctent/CVPR2021/html/Jaume_Quantifying_Explainers_of_Graph_Neural_Networks_in_Computational_Pathology_CVPR_2021_paper.html.
[37] W. Lu, S. Graham, M. Bilal, N. Rajpoot, F. Minhas, Capturing cellular topology in multi-gigapixel pathology images, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2020, pp. 260–261. https://openaccess.thecvf.com/content_CVPRW_2020/html/w16/Lu_Capturing_Cellular_Topology_in_Multi-Gigapixel_Pathology_Images_CVPRW_2020_paper.html.
[38] A. Nair, H. Arvidsson, V. Gatica, E. J, N. Tudzarovski, K. Meinke, R.V Sugars, A graph neural network framework for mapping histological topology in oral mucosal tissue, BMC Bioinform. 23 (1) (2022) 506. https://link.springer.com/article/10.1186/s12859-022-05063-5.
[39] M. Sureka, A. Patil, D. Anand, A. Sethi, Visualization for histopathology images using graph convolutional neural networks, in: Proceedings of the IEEE 20th International Conference on Bioinformatics and Bioengineering (BIBE), IEEE, 2020, pp. 331–335. https://ieeexplore.ieee.org/abstract/document/9288001.
[40] R.J. Chen, M.Y. Lu, M. Shaban, C. Chen, T.Y. Chen, D.F. Williamson, F. Mahmood, Whole slide images are 2d point clouds: context-aware survival prediction using patch-based graph convolutional networks, in: Proceedings of the Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, Springer International Publishing, 2021, pp. 339–349. September 27–October 1, 2021, Proceedings, Part VIII 24, https://link.springer.com/chapter/10.1007/978-3-030-87237-3_33.
[41] P. Liu, L. Ji, F. Ye, B. Fu, GraphLSurv: a scalable survival prediction network with adaptive and sparse structure learning for histopathological whole-slide images, Comput. Methods Progr. Biomed. 231 (2023) 107433. https://www.sciencedirect.com/science/article/pii/S0169260723001001.
[42] W. Hou, L. Yu, C. Lin, H. Huang, R. Yu, J. Qin, L. Wang, Hˆ 2-MIL: exploring hierarchical representation with heterogeneous multiple instance learning for whole slide image analysis, in: Proceedings of the AAAI Conference on Artificial Intelligence 36, 2022, pp. 933–941. https://ojs.aaai.org/index.php/AAAI/article/view/19976.
[43] S. Bai, F. Zhang, P.H. Torr, Hypergraph convolution and hypergraph attention, Pattern Recognit. 110 (2021) 107637. https://www.sciencedirect.com/science/article/pii/S0031320320304404.
[44] M. Ilse, J. Tomczak, M. Welling, Attention-based deep multiple instance learning, in: Proceedings of the International conference on machine learning, PMLR, 2018, pp. 2127–2136, in: https://proceedings.mlr.press/v80/ilse18a.html?ref=https://githubhelp.com.
[45] X. Wang, H. Ji, C. Shi, B. Wang, Y. Ye, P. Cui, P.S. Yu, Heterogeneous graph attention network, in: Proceedings of the World Wide Web Conference, 2019, pp. 2022–2032. https://dl.acm.org/doi/abs/10.1145/3308558.3313562.
[46] Rong, Y., Huang, W., Xu, T., & Huang, J. (2019). Dropedge: towards deep graph convolutional networks on node classification. arXiv preprint arXiv:1907.10903. https://arxiv.org/abs/1907.10903.
[47] X. Xia, H. Yin, J. Yu, Q. Wang, L. Cui, X. Zhang, Self-supervised hypergraph convolutional networks for session-based recommendation, in: Proceedings of the AAAI Conference on Artificial Intelligence 35, 2021, pp. 4503–4511. https://ojs.aaai.org/index.php/AAAI/article/view/16578.