H^2-MIL: Exploring Hierarchical Representation with Heterogeneous Multiple Instance Learning for Whole Slide Image Analysis

Wentai Hou[1,2,*], **Lequan Yu**[3,*], **Chengxuan Lin**[2,4,*], **Helong Huang**[4], **Rongshan Yu**[4], **Jing Qin**[5], **Liansheng Wang**[4,†]

[1] Information and Communication Engineering Department at School of Informatics, Xiamen University, Xiamen, China
[2] National Institute for Data Science in Health and Medicine, Xiamen University, Xiamen, China
[3] Department of Statistics and Actuarial Science, The University of Hong Kong, Hong Kong SAR, China
[4] Department of Computer Science at School of Informatics, Xiamen University, Xiamen, China
[5] Center for Smart Health, School of Nursing, The Hong Kong Polytechnic University, Hong Kong SAR, China
houwt@stu.xmu.edu.cn, lqyu@hku.hk, lincx@stu.xmu.edu.cn, hlhuang@stu.xmu.edu.cn, rsyu@xmu.edu.cn, harry.qin@polyu.edu.hk, lswang@xmu.edu.cn

Abstract

Current representation learning methods for whole slide image (WSI) with pyramidal resolutions are inherently homogeneous and flat, which cannot fully exploit the multi-scale and heterogeneous diagnostic information of different structures for comprehensive analysis. This paper presents a novel graph neural network-based multiple instance learning framework (*i.e.*, H^2-MIL) to learn hierarchical representation from a heterogeneous graph with different resolutions for WSI analysis. A heterogeneous graph with the "resolution" attribute is constructed to explicitly model the feature and spatial-scaling relationship of multi-resolution patches. We then design a novel resolution-aware attention convolution (RAConv) block to learn compact yet discriminative representation from the graph, which tackles the heterogeneity of node neighbors with different resolutions and yields more reliable message passing. More importantly, to explore the task-related structured information of WSI pyramid, we elaborately design a novel iterative hierarchical pooling (IH-Pool) module to progressively aggregate the heterogeneous graph based on scaling relationships of different nodes. We evaluated our method on two public WSI datasets from the TCGA project, *i.e.*, esophageal cancer and kidney cancer. Experimental results show that our method clearly outperforms the state-of-the-art methods on both tumor typing and staging tasks.

Introduction

Pathological slide examination is considered as the "gold standard" for diagnosis and treatment planning of many diseases (Yao et al. 2020; Cai et al. 2021; Ortega et al. 2018). To facilitate preservation and retrieval, pathological slides are usually scanned into whole slide images (WSI) with pyramidal resolutions for quantitative analysis. As shown in Figure 1, at different resolutions, pathologists can clearly observe the tissue features ranging from cellular-scale to millimeter-scale. However, it is a time-consuming and tedious task for pathologists to perform manual inspection on a WSI due to (1) the gigapixels and the pyramidal resolutions of the WSI and (2) the complex colors and patterns of different structures. Therefore, it is highly demanded to develop automatic and accurate inspection tools for WSI analysis to reduce the workload of pathologists and improve the accuracy and efficiency of the examination (Cui and Zhang 2021; Farris et al. 2021).

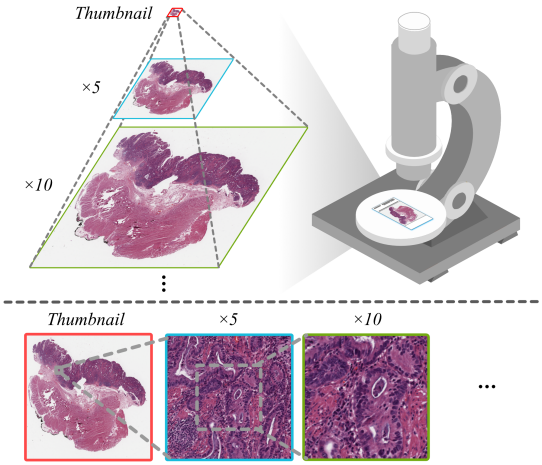

Figure 1. Top: A WSI can be scaled at different resolutions, forming a pyramid structure. Bottom: Fixed size patches obtained from WSI pyramid present varying diagnostic information ranging from global level to cellular level.

Due to the huge size and high annotation cost of WSI (*e.g.*, the usual size is 40000×40000), multiple instance learning (MIL) (Maron and Lozano-Pérez 1998) is a promising framework to learn effective representations for WSI analysis. Specifically, we consider each WSI as a bag and the numerous cropped patches in WSI (*e.g.*, by sliding window) as instances. With deep neural networks (DNNs), the features of patches (instances) are extracted and aggregated to produce WSI-level prediction (Ilse, Tomczak, and Welling 2018; Tu et al. 2019). However, the previous MIL works usually applied to a single resolution of WSIs (Tellez et al. 2021; Courtiol et al. 2018; Campanella et al. 2019),

*Equal contribution.
†Corresponding author.

ignoring the multiscale feature information of WSIs. Recently, inspired by the diagnosis procedure of pathologists, some researchers have extended MIL to learn representations from WSI pyramid (Li, Li, and Eliceiri 2021; Chen et al. 2021), and achieved better performance than previous single-resolution-based methods.

However, existing methods have not tapped the full potential of WSI pyramid to produce better representations due to the following limitations. First, the patches with different resolutions present quite different diagnostic information ranging from global-scale, cellular-scale (*e.g.*, nucleus and micro-environment) to tissue-scale (*e.g.*, vessels and glands). The heterogeneity of different resolution patches should be sufficiently considered during the learning process. While in existing methods, the extracted features of different resolution patches are often simply concatenated or linked, which may lead to suboptimal learning. Second, the existing methods utilize global pooling (*e.g.*, max or average pooling) to aggregate the local representations of WSI pyramid, which are inherently flat and incapable of capturing hierarchical structure information from the WSI. These limitations prohibit existing algorithms from extracting key information (*e.g.*, differentiation degree and invasion depth of tumor) of the WSI pyramid for analysis.

In this paper, we propose a novel H^2-MIL framework to learn hierarchical representation from heterogeneous pyramidal WSIs for more comprehensive slide-level analysis. The proposed framework consists of three key components: (1) a novel heterogeneous graph with an extra "resolution" attribute, which is constructed from the WSI pyramid and can be served as an effective data structure for modeling WSI pyramid and retaining the heterogeneity of multi-resolution patches, (2) a new resolution-aware attention convolution (RAConv) block, which is proposed for more reliable message passing by considering both resolution-level attention and node-level attention during the learning process, and (3) a novel iterative hierarchical pooling (IHPool) module, which is elaborately designed to explore the task-related latent structures of WSI pyramid in order to progressively aggregate the heterogeneous graph, leading to better analysis performance and richer interpretability. We extensively evaluated our method on the esophageal cancer (ESCA) and kidney cancer (KICA) cohorts from TCGA project, and our method clearly outperforms state-of-the-art methods (SOTAs) on WSI typing and staging tasks. Our main contributions can be summarized as follows.

- We pioneer the usage of heterogeneous graph for WSI pyramid analysis. A new heterogeneous graph is constructed to explicitly model the spatial-scaling relationships and conveniently retain the heterogeneity of multi-resolution patches in WSI.
- To facilitate the discriminative representation learning from the graph, we design a new RAConv block to tackle the heterogeneity of graph nodes with different resolutions and a novel IHPool module to progressively aggregate the graph, yielding more reliable message passing and richer interpretability.
- Extensive experiments with promising results on two public TCGA datasets validate the effectiveness of the proposed method for WSI analysis. The codes are available at https://github.com/lin-lcx/H2-MIL.

Related Work

Multiple Instance Learning for WSI. MIL itself is a widely studied topic. Readers can refer Carbonneau et al. (2018) for a comprehensive survey. We will briefly review some typical works. According to the representation objects, existing WSI methods can be divided into *single-resolution* and *multi-resolution* oriented methods. For single-resolution oriented methods, all instances are extracted from a certain resolution of WSIs. For example, Tellez *et al.* (2021) recombined the feature vectors of patches into compressed WSI according to spatial relationship, and trained a Convolutional Neural Network (CNN) on this compressed WSI to predict WSI-level label. Courtiol *et al.* (2018) proposed a Deep Neural Network (DNN) based MinMax model to aggregate the local descriptors of patches, to obtain the WSI slide-level representation. Campanella *et al.* (2019) adopted a Recurrent Neural Network (RNN) to integrate the feature vectors of patches into the prediction of WSI. However, these methods ignored the use of WSI pyramid for representation learning.

In recent years, multi-resolution oriented methods aroused the interest of researchers. Hashimoto *et al.* (2020) applied instance-wise attention to aggregate the features of multi-resolution patches, which is extracted by scale-specific extractor networks. Li *et al.* (2021) concatenated the feature embeddings of multi-resolution patches and trained a dual-stream DNN aggregator for WSI prediction. Chen *et al.* (2021) first selected several patches based on an attention map of thumbnail and connected their corresponding multi-resolution patches as a tree structure. After that, a relevance enhanced GNN model was proposed to investigate this tree structure and learn representation for WSI. However, these methods do not fully consider the heterogeneity of multi-resolution patches and are lack of hierarchical analysis of these patches, without fully exploiting the richer pyramid information for WSI analysis.

Attention Mechanism on Graph Network. The message passing of traditional GCN (Kipf and Welling 2016) excessively depends on the structure of the graph (*i.e.*, edges and edge weights). Since the attention mechanism allows the model to dynamically focus on certain informative parts of the input to perform the task (Vaswani et al. 2017), it was widely adopted and achieved state-of-the-art performance for different tasks (Wang et al. 2019a,b; Guo et al. 2021; Hou et al. 2021). For example, Petar *et al.* (2017) first proposed a graph attention network (GAT) to directly learn the attention weight of neighborhoods with non-linear layers. However, in nature, the heterogeneous graph (Zhang et al. 2019) is a more common data format than homogeneous graph. Thus, Yang *et al.* (2021) proposed a Heterogeneous GAT (HGAT) to deal with the attention score calculation for heterogeneous graph through a nested GAT network, which improves the performance of short text classification task. Due to the huge

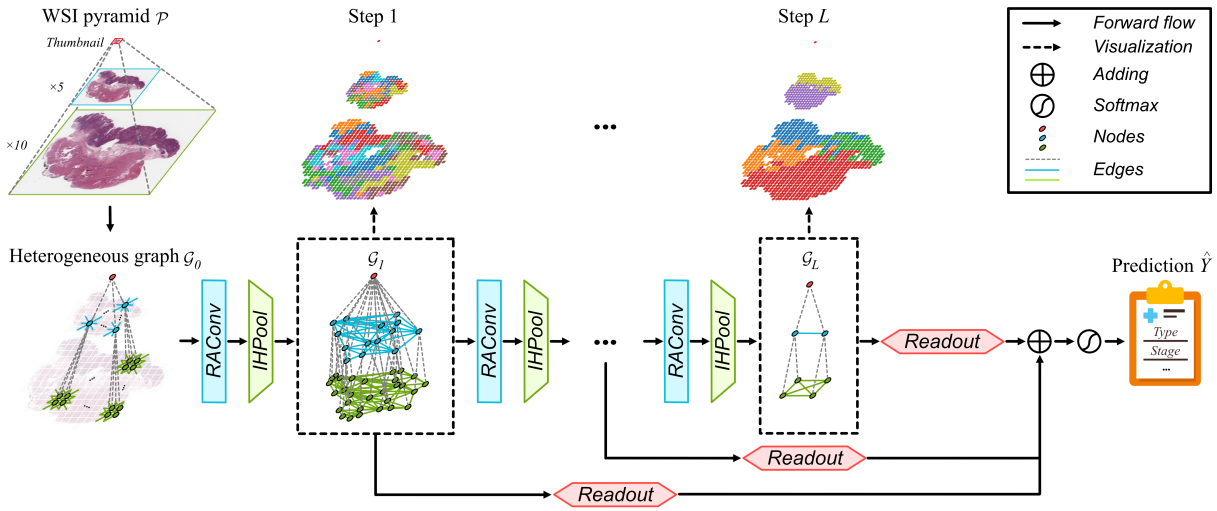

Figure 2. Overview of the proposed H^2-MIL framework. A WSI pyramid and multi-resolution patches are considered as a bag and instances, respectively. Our proposed heterogeneous GCN with stacked RAConv blocks and IHPool modules is a typical MIL operator to progressively aggregate instances information for bag prediction.

size of WSI, the attention mechanism is important to the representation learning for WSI pyramid. And it is also urgent to design a heterogeneous graph convolution suitable for the representation learning of WSI pyramid.

Hierarchical Graph Pooling. The traditional global pooling methods (Kipf and Welling 2016) directly pool the whole graph into a node, which may ignore some structured information in the large-scale graph for graph analysis. For example, the functional groups of molecules are essential to the molecular properties prediction. In this regard, Ying *et al.* (2018) proposed a DiffPool to cluster all nodes in a graph with a differential module. Gao *et al.* (2019) proposed a TopK pooling scheme for sparse graph classification, which projects all node features to one dimension and performs Max operation to filter nodes. Meanwhile, Lee (2019) proposed a SAGP network, which uses self-attention mechanism to learn an allocation result for each node. The hierarchical pooling mechanism is also important for GNN-based WSI pyramidal representation learning, as some key structured information (*e.g.*, invasion depth of tumor) have important reference value for many WSI prediction tasks, such as tumor staging (Rice, Patil, and Blackstone 2017) and survival prediction (Wang et al. 2018). However, to the best of our knowledge, few studies are focused on hierarchical pooling method for heterogeneous graph representation learning.

Methodology

To fully exploit the pyramid feature information of WSIs, we propose a novel H^2-MIL framework for whole slide image analysis. Figure 2 illustrates the pipeline of the proposed framework. Given a WSI pyramid $\mathcal{P}$, our framework predicts the slide-level label $\hat{Y}$ by fully considering the heterogeneity between the patches with different resolutions and exploring the task-related structured information of the WSI. As shown in Figure 2, we first construct a heterogeneous graph $\mathcal{G}_0$ by taking the feature embeddings of multi-resolution patches as nodes and their spatial-scaling relationships as edges. Then, the generated heterogeneous graph will be fed into the proposed H^2-MIL network, equipped with RAConv blocks and IHPool modules, to extract compact yet discriminative representation, as well as mine hierarchical structure semantics for WSI analysis. Finally, a WSI-level classifier is employed to obtain the slide-level prediction based on the learned graph representation. In the following subsections, we will detail the heterogeneous graph construction, RAConv calculation, IHPool design, and the learning strategy of the whole framework.

Heterogeneous Graph Construction

In clinic, pathologists comprehensively observe the tissue-level information (*e.g.*, vessels and glands) from low-resolution WSI and cellular-level information (*e.g.*, nucleus and micro-environment) from high-resolution WSI for diagnosis. To simulate this reading scenario, it is necessary to model the WSI pyramid completely and flexibly. Here we constructed a heterogeneous graph with multi-resolution attribute to achieve this purpose, which is able to explicitly represent the spatial-scaling relationships and conveniently retain the heterogeneity of multi-resolution patches.

As shown in Figure 2, given a WSI pyramid $\mathcal{P}$ scanned by $R = \{\text{“}Thumbnail\text{”}, \times 5, \times 10, ...\}$ different resolutions, we first use a sliding window strategy to crop $\mathcal{P}$ into numerous multi-resolution patches. Note that we also use a simple thresholding method (*i.e.*, variance of RGB value is less than a threshold) to filter those background regions and only crop patches from the foreground regions. Then, a heterogeneous graph $\mathcal{G}_0 = \{\mathcal{V}, \mathcal{E}, \mathcal{R}\}$ is constructed based on the feature embeddings and spatial-scaling relationship of the cropped multi-resolution patches. Here $\mathcal{V}$ is the set of feature embeddings of multi-resolution patches, which is usually extracted by a CNN encoder, such as KimiaNet (Riasatian et al. 2021),

ImageNet pretrained ResNets (He et al. 2016) and *etc.*Here we use the KimiaNet to extract the feature. The set of edges $\mathcal{E}$ represents the relations between patches, including spatial and scaling relations shown in Figure 2. Specifically, the spatial relations describe the bordering relationship between patches with the same resolution (colored solid lines) and the scaling relationship describes the relationship between patches with different resolutions at the same location (gray dashed lines). Moreover, $\mathcal{R}$ is an extra attribute set which contains the "resolution" attribute of each node to facilitate the following calculation.

As can be observed from Figure 2, the constructed heterogeneous graph has the following characteristics. First, each heterogeneous node has the same feature dimension and unique resolution attribute. Second, the number of nodes with different resolution attributes is unbalanced. Third, the heterogeneous graphs are dense, as the dependencies among nodes are established in an 8-adjacent manner. Overall, the constructed heterogeneous graph can completely represent the feature and spatial-scaling relations of WSI pyramid, and the heterogeneity of multi-resolution patches can be conveniently extracted for further analysis.

Resolution-aware Attention Convolution

Different from previous graph-based methods, each node in the heterogeneous graph is associated with a resolution attribute, which introduces richer information to be exploited in the learning process. Unfortunately, the existing graph attention convolutions are not suitable for directly processing the heterogeneous graph of the WSI pyramid due to several reasons. First, the traditional GAT (Veličković et al. 2017) suffers from reduced performance as it ignores the heterogeneity of nodes. Second, the unbalanced number of heterogeneous nodes in different resolutions is not considered in the HGAT (Yang et al. 2021), leading to the neglect of low resolution nodes with global semantic information.

To address these issues, we propose a novel RAConv, which considers the heterogeneity of neighbors and alleviates the bias caused by the imbalance number of heterogeneous neighbors when calculating the attention scores. Specifically, we work on a heterogeneous graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}, \mathcal{R}\}$ generated from a WSI pyramid with $R = \{"Thumbnail", \times 5, \times 10, ...\}$ different resolutions. Supposing $H^{(l)} \in \mathbb{R}^{|\mathcal{V}| \times d}$ is the hidden representation of $\mathcal{G}_l$ in the l^{th} layer of H^2-MIL network, where $|\mathcal{V}|$ is a number of nodes and d is the dimension of each node, the layer-wise propagation rule is as follows

$$H^{(l+1)} = \sigma\left(\tilde{A} \cdot H^{(l)} \cdot W^{(l)}\right). \tag{1}$$

Here $\tilde{A}$ denotes the symmetric normalized adjacency matrix, $W^{(l)}$ denotes a layer-specific learnable matrix, and $\sigma(\cdot)$ denotes an activation function, such as ReLU. For the WSI classification task, the neighboring nodes with different resolutions may have different influences for message passing. In this regard, we propose a new dual-stream attention mechanism to further re-calibrate the node-level attention by taking the resolution-level attention into account.

For a source node v, we denote its all neighboring nodes with resolution r as $\mathcal{N}_r$, where $r \in R$ and R is the total number of resolutions. To compute resolution-level attention of the source node v, we first calculate the prototype of resolution r as h_r, which is the mean feature embedding of $\mathcal{N}_r$. Based on the source node embedding h_v and the resolution prototype h_r, the r-th resolution-level attention scores for v is calculated as

$$\alpha_r = \frac{\exp(\beta(U_r^T[h_v||h_r]))}{\sum_{r' \in R} \exp(\beta(U_r^T[h_v||h_{r'}]))}, \tag{2}$$

where U_r is a learnable attention layer for resolution r, $||$ denotes the concatenation operation, and $\beta(\cdot)$ denotes an activation function, such as LeakyReLU. Besides the resolution-level attention, we also employ the node-level attention mechanism to strengthen the key neighbors and suppress the noise neighbors for each resolution. Specifically, based on the source node embedding h_v and the neighboring node embedding $h_{v'}$, $v' \in \mathcal{N}_r$, the node-level attention scores can be calculated as

$$\alpha_{v'} = \frac{\exp(\beta(V^T[h_v||h_{v'}]))}{\sum_{v'' \in \mathcal{N}_r} \exp(\beta(V^T[h_v||h_{v''}]))}. \tag{3}$$

Here V is a learnable attention layer for neighboring nodes of v. Finally, the attention score of v to v' is calculated by re-calibrating the node-level attention using resolution-level attention:

$$\alpha_{vv'} = \alpha_r \cdot \alpha_{v'}. \tag{4}$$

Therefore, the layer-wise propagation rule of Eq. (1) can be replaced as

$$H^{(l+1)} = \sigma\left(\mathcal{A} \cdot H^{(l)} \cdot W^{(l)}\right), \tag{5}$$

where $\mathcal{A}$ represents the attention matrix and the v^{th} row and v'^{th} column element $\alpha_{vv'}$ is defined as Eq. (4).

Iterative Hierarchical Pooling

As mentioned before, the heterogeneous graph representation of the WSI pyramid is dense. It is thereby necessary to introduce a pooling layer into the graph to improve the receptive field and reduce redundant calculation in the learning process. Moreover, for some WSI classification tasks, such as tumor staging, scoring and *etc.*, structured feature (*e.g.*, aggregation morphology and invasion depth) has great reference value. Therefore, progressively exploring these latent structured information may be conducive to learn discriminative representation for the WSI pyramid, which is ignored by most of the existing MIL methods. To this end, as shown in Figure 2, we proposed a novel Iterative Hierarchical Pooling (IHPool) module to adaptively aggregate nodes with similar semantic features and spatial distribution while maintaining the scaling relationship between heterogeneous nodes, leading to richer multi-resolution structured information and interpretability for decision-making.

The design principle of proposed IHPool are as follows: (1) to maintain the pyramid structure and global information, the thumbnail nodes are always retained; (2) to prevent

Algorithm 1: The IHPool algorithm.

Input: Heterogeneous graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}, \mathcal{R}\}$ with node feature X and adjacent matrix A; Pooling ratio k; Resolutions $R = \{\text{“}Thumbnail\text{”}, \times 5, \times 10, ...\}$; Learnable projection layer $\mathbf{P}$.

Output: Pooled heterogeneous graph $\mathcal{G}'$ with node feature X' and adjacency matrix A'.

Initialize an empty node assignment list S;
Initialize the number of clusters N;
Initialize a counter c;
for r in R **do**
  if r is “$Thumbnail$” **then**
    The thumbnail node is individually assigned;
    Append the assignment result to S;
    $N \leftarrow 1$.
  else
    $c \leftarrow 0$;
    for n in N **do**
      Determine the node set $\mathcal{V}_r^n$ to be pooled based on scaling relations;
      Calculate the fitness set $\phi_r^n \leftarrow tanh(\frac{\mathcal{V}_r^n \cdot \mathbf{P}}{||\mathbf{P}||})$;
      Sample $\lceil k \cdot |\mathcal{V}_r^n| \rceil$ clusters based on ϕ_r^n;
      Assign $\mathcal{V}_r^n$ to clusters based on the spatial distance and fitness difference;
      Append the assignment result of $\mathcal{V}_r^n$ to S;
      $c \leftarrow c + \lceil k \cdot |\mathcal{V}_r^n| \rceil$.
    $N \leftarrow c$.
Matrix S;
$X' \leftarrow S^T X$; // Aggregate node features.
$A' \leftarrow S^T A S$. // Maintain graph connectivity.

contradictory results, the nodes to be pooled in each iteration are depended on the pooling results of corresponding low-resolution nodes; (3) pooling centers are dynamically selected according to the pooling ratio and learned fitness set φ; and (4) node assignment are determined by combining spatial distance and fitness difference. The pseudocode of IHPool is shown in Algorithm 1. To the best of our knowledge, the proposed IHPool is the first attempt to explore the latent structured information of WSI pyramid.

Network Architecture and Training Strategy

The RAConv and IHPool are the basic components of our proposed H^2-MIL. After going through a $L-$layer H^2-MIL, the coarse-grained information of the constructed heterogeneous graph is extracted dynamically, which can be considered as different level representations of tissues (*e.g.*, $\mathcal{G}_1$ and $\mathcal{G}_L$ in Figure 2). As shown in Figure 2, we further employ a residual connection-like structure to aggregate these coarse-grained information for decision-making. Formally, the final prediction is

$$\hat{Y} = Softmax(\sum_{l=1}^{L} Readout(\mathcal{G}_l)), \tag{6}$$

where $Readout$ is a global mean or max pooling layer for generating representation for each sub-graph.

For the network training, the cross-entropy loss is adopted for WSI classification tasks and the objective loss is defined as

$$\mathcal{L} = -\frac{1}{M} \sum_{i=1}^{M} \sum_{j=1}^{C} Y_{ij} log(\hat{Y}_{ij}), \tag{7}$$

where M is the number of samples, C is the number of classes, and Y is the one-hot label matrix. The gradient descent algorithm is adopted for network optimizing.

Experiments

Datasets and Experiment Setting. In this study, we evaluate the performance of our proposed H^2-MIL on two public clinical WSI benchmark datasets from The Cancer Genome Atlas (TCGA) project, *i.e.*, ESCA and KICA. For each dataset, the WSI classification tasks include tumor typing and staging. Note that patients with TNM label of I/II are considered as early stage and patients with TNM label of III/IV are considered as late stage. We only include patients whose diagnostic WSI, type and stage records are all available. All WSIs are standardized in to 3-level pyramids, where the magnifications are $Thumbnail$, $\times 5$ and $\times 10$, respectively.

- **ESCA** is the esophageal cancer cohort that contains 135 cases. For the typing task, all cases are divided into adenocarcinoma (86) and squamous cell carcinoma (49). For the staging task, all cases are divided into early stage (80) and late stage (55).
- **KICA** is the kidney cancer cohort contains 275 cases. For the typing task, all cases are divided into chromophobe (192) and renal papillary cell carcinoma (83). For the staging task, all cases are divided into early stage (205) and late stage (70).

The accuracy (ACC) and the area under the curve (AUC) of receiver operating characteristic (ROC) are used as metrics for both tumor typing and staging tasks. The mean results and standard deviation of 5 repeated 5-fold cross-validation (5-fold CV) are reported. Note that during the cross-validation procedure, 25% of the training data are also randomly split as the validation data for choosing the checkpoint.

Implementation Details. The proposed framework was implemented with Pytorch Geometric framework (Fey and Lenssen 2019) and all experiments were conducted on a workstation with four RTX 3090 (24 GB) GPUs. For fair comparison with other MIL methods, we used pretrained KimiaNet (Riasatian et al. 2021) as the feature extraction network for multi-resolution patches, as KimiaNet achieves better balance between representation and efficiency. The size of multi-resolution patches was fixed as 512×512 and converted into 1024-dimensional features. The Adam optimizer was adopted, and the network was trained for 50 epochs. The learning rate was tuned from {5e-4, 1e-4, 5e-5}. The output dimension of readout operation was tuned from {128, 256, 512}.The dropout ratio of linear layers was tuned from {0, 0.1, 0.2, 0.3, 0.4}. The pooling ratio of IHPool was tuned from {0.1, 0.2, 0.3, 0.4}.

Experiment	Method	Typing			Staging		
		ACC	AUC	F1	ACC	AUC	F1
SOTAs	MIL-CNN	68.74 ± 3.90	75.50 ± 4.82	67.89 ± 2.59	56.59 ± 2.46	57.73 ± 7.02	53.20 ± 2.82
	MinMax	86.44 ± 2.69	91.78 ± 3.09	85.94 ± 3.71	58.52 ± 3.14	60.88 ± 4.37	57.15 ± 3.78
	MIL-RNN	81.33 ± 3.16	87.50 ± 2.65	80.20 ± 1.99	64.74 ± 3.43	63.25 ± 3.05	63.05 ± 2.69
	MS-DA-MIL	88.89 ± 3.70	92.27 ± 4.67	88.22 ± 3.88	69.13 ± 2.13	72.53 ± 3.32	64.14 ± 5.36
	DS-MIL	87.26 ± 2.01	92.47 ± 1.65	86.26 ± 1.99	66.67 ± 1.78	69.23 ± 2.66	65.38 ± 2.63
	PTree-Net	88.29 ± 1.27	91.03 ± 0.97	86.32 ± 1.12	65.78 ± 1.44	64.52 ± 3.18	63.42 ± 3.49
Ablation study	GCN + IHPool	87.41 ± 2.34	92.58 ± 1.71	85.80 ± 2.85	68.00 ± 2.74	71.06 ± 2.38	67.97 ± 2.89
	GAT + IHPool	89.93 ± 1.60	91.68 ± 2.09	89.10 ± 1.60	69.18 ± 1.44	71.16 ± 2.20	68.23 ± 1.43
	HGAT + IHPool	90.96 ± 1.19	93.15 ± 1.48	89.71 ± 0.83	69.48 ± 2.31	72.14 ± 0.38	69.86 ± 2.96
	RAConv+GAP	90.07 ± 1.66	91.57 ± 1.51	88.58 ± 2.18	67.25 ± 2.93	69.59 ± 3.87	65.11 ± 5.03
	RAConv+TopK	89.63 ± 1.41	90.84 ± 2.48	87.97 ± 1.71	68.00 ± 1.60	69.43 ± 1.23	67.26 ± 1.08
	RAConv+SAGP	87.70 ± 1.97	90.22 ± 2.79	87.09 ± 1.71	68.89 ± 1.23	71.25 ± 2.24	65.21 ± 2.74
Ours	H^2-MIL	$\mathbf{91.56 \pm 1.60}$	$\mathbf{96.40 \pm 0.59}$	$\mathbf{91.01 \pm 1.83}$	$\mathbf{72.89 \pm 2.32}$	$\mathbf{76.36 \pm 1.69}$	$\mathbf{71.92 \pm 2.56}$

Table 1. Classification results on the ESCA cohort. We use the same patch feature extractor for all methods.

Experiment	Method	Typing			Staging		
		ACC	AUC	F1	ACC	AUC	F1
SOTAs	MIL-CNN	68.44 ± 4.69	67.66 ± 5.76	60.40 ± 4.69	58.98 ± 3.78	59.68 ± 1.05	49.72 ± 1.68
	MinMax	92.76 ± 0.77	96.69 ± 0.36	91.37 ± 0.69	64.29 ± 4.83	59.22 ± 3.34	54.27 ± 3.63
	MIL-RNN	89.75 ± 2.27	94.41 ± 1.32	87.37 ± 2.29	62.55 ± 3.64	56.80 ± 3.43	54.22 ± 4.18
	MS-DA-MIL	93.52 ± 1.31	94.77 ± 2.96	87.68 ± 2.78	65.45 ± 1.82	64.69 ± 4.18	62.06 ± 4.25
	DS-MIL	91.19 ± 1.91	95.87 ± 1.13	90.10 ± 1.88	67.41 ± 4.45	60.25 ± 3.81	56.08 ± 4.66
	PTree-Net	91.20 ± 2.19	94.94 ± 1.46	88.98 ± 3.40	66.18 ± 4.66	56.95 ± 2.83	53.00 ± 3.13
Ablation study	GCN + IHPool	91.78 ± 0.71	95.04 ± 1.17	89.97 ± 1.03	65.89 ± 4.03	61.70 ± 3.22	56.68 ± 2.72
	GAT + IHPool	93.45 ± 0.97	96.53 ± 1.10	92.33 ± 1.55	69.01 ± 3.40	62.44 ± 2.86	58.52 ± 1.66
	HGAT + IHPool	93.16 ± 1.11	96.57 ± 0.67	91.99 ± 1.25	69.75 ± 4.03	60.42 ± 2.55	54.23 ± 1.44
	RAConv + GAP	92.58 ± 1.35	94.51 ± 1.08	91.00 ± 1.20	64.80 ± 4.06	62.51 ± 5.72	58.73 ± 3.20
	RAConv + TopK	89.45 ± 1.21	93.89 ± 0.88	88.32 ± 1.19	66.04 ± 2.96	61.69 ± 5.01	57.63 ± 3.24
	RAConv + SAGP	90.83 ± 0.84	93.09 ± 1.67	88.04 ± 1.67	65.16 ± 1.58	60.43 ± 3.35	58.37 ± 4.48
Ours	H^2-MIL	$\mathbf{95.05 \pm 0.49}$	$\mathbf{98.05 \pm 0.57}$	$\mathbf{93.30 \pm 1.34}$	$\mathbf{70.62 \pm 2.60}$	$\mathbf{69.16 \pm 1.52}$	$\mathbf{63.51 \pm 5.01}$

Table 2. Classification results on the KICA cohort. We use the same patch feature extractor for all methods.

Comparison with State-of-the-art Methods. We compare the proposed method with three single-resolution oriented MIL methods: (1) MIL-CNN (2021), (2) MinMax (2018), and (3) MIL-RNN (2019), as well as three multi-resolution oriented MIL methods: (4) MS-DA-MIL (2020), (5) DS-MIL (2021), and (6) PTree-Net (2021). The comparative results on ESCA and KICA datasets are shown in the top half of Table 1 and Table 2, respectively.

Generally, due to the advantage of the WSI pyramid, the overall performance of multi-resolution MIL methods is obviously better than that of single-resolution MIL methods. As our method not only considers the heterogeneity of multi-resolution patches but also excavates the potential structured information to facilitate tumor analysis, our method clearly outperforms the existing SOTAs by a large margin on both datasets for both two tasks. Moreover, we observe that our H^2-MIL achieves more obvious improvement for tumor staging, as the structured information explored by IHPool would reflect the aggregation morphology and infiltration depth of tumor, which provides important value for tumor staging.

Ablation Study. We conduct an ablation study to demonstrate the effectiveness of each proposed component. The results are shown in the bottom half of Table 1 and Table 2. We first compare the proposed RAConv with traditional GCN (2016), GAT (2017), and HGAT (2021). As can be observed that the proposed RAConv is superior to the existing graph convolution methods, as RAConv not only considers the heterogeneity of nodes with different resolutions, but also prevents the attention deviation caused by the node imbalance with different resolutions. We also compare the proposed IHPool with widely used pooling methods, such as global average pool (GAP), TopK (2019) and SAGP (2019). Overall, the proposed IHPool module largely outperforms the widely used pooling methods, as the iterative design principle of IHPool effectively maintains the pyramid structure of the heterogeneous graph and effectively avoids the contradiction between the pooling results of different resolutions.

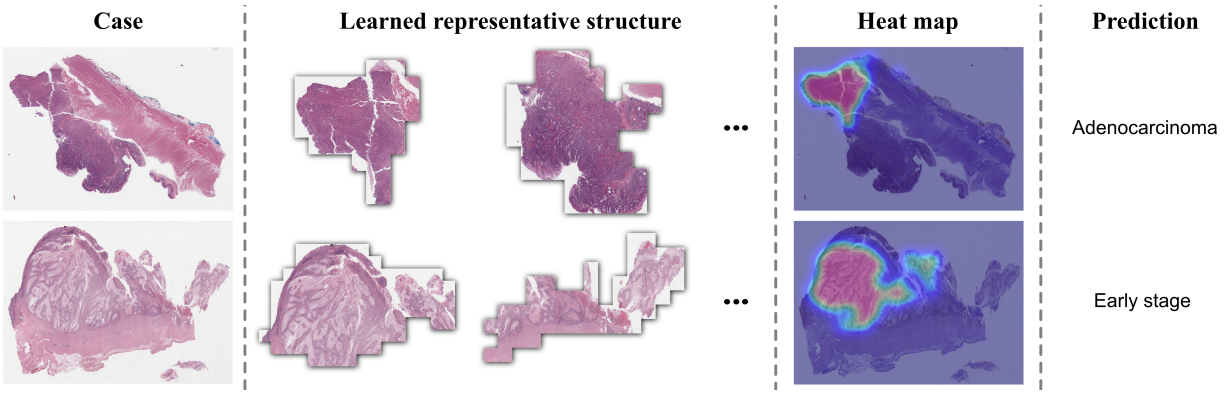

Figure 3. Visualization of representative structures learned by the network. For each case, the representative structures are extracted by the last IHPool layer and ordered by the fitness score from left to right. The heatmap is obtained by normalizing and weighting the fitness of patch and their corresponding structure during the forward propagation.

Scheme	ESCA		KICA	
	Typing	Staging	Typing	Staging
Thu. + Mid.	87.17	70.98	93.31	64.59
Thu. + Bot.	94.49	73.20	96.13	64.57
Thu. + Mid. + Bot.	**96.40**	**76.36**	**98.05**	**69.16**

Table 3. Comparison of different multi-resolution schemes in H^2-MIL. Thu., Mid., and Bot. denotes the thumbnail, middle-level, and bottom-level of WSI pyramid, respectively. The mean AUC is reported.

Setting	ESCA		KICA	
	Typing	Staging	Typing	Staging
$L = 1$	94.64	76.04	97.50	**69.16**
$L = 2$	**96.40**	**76.36**	**98.05**	68.18
$L = 3$	95.46	74.86	97.44	65.42
$L = 4$	93.30	74.09	97.52	64.24

Table 4. Performance of H^2-MIL with respect to different settings of L. The mean AUC is reported.

Investigation of Multi-resolution Scheme. We also investigate the impact of different multiscale schemes (*i.e.*, using different resolutions) on the performance of the proposed H^2-MIL and the results are shown in Table 3. The "Thu. + Mid. + Bot." scheme shows the best performance than other schemes. This is because the extracted thumbnail, mid-level and bottom-level WSIs contain complementary diagnostic information, ranging from global-level, tissue-level, to cell-level. These experimental results also show that our RAConv can effectively capture and make use of the heterogeneity between multi-resolution patches and further dig the advantages of WSI pyramid.

Investigation of Hierarchical Pooling. The number of pooling L is an important hyperparameter affecting the performance of the hierarchical analysis. Intuitively, if L is too large, the features learned by each node will tend to be homogeneous and the network would be overfitting. Thus, it is not recommended setting the H^2-MIL too deep. We investigate the performance of proposed H^2-MIL with different settings of L and the results are shown in Table 4. It is observed that for ESCA typing, ESCA staging and KICA typing tasks, the best performance of H^2-MIL is achieved by setting L to 2, while for KICA staging task, the best performance of H^2-MIL is achieved by setting L to 1.

Visualization of Learned Representations. In clinic, tumor typing and staging tasks require not only high performance but also a strong rationale for judgment. Our H^2-MIL network could provide abundant interpretability by visualizing the learned task-related structures. As shown in Figure 2, the proposed network can effectively deconstruct WSI at multiple levels and the extracted structure could well describe the aggregation morphology and infiltration depth of tumors, which would be helpful to improve the performance of many WSI analysis tasks. The other cases shown in Figure 3 also confirm to this observation.

Conclusion

In this paper, we proposed a novel H^2-MIL network to learn the hierarchical representation from the heterogeneous graph of WSI pyramid for WSI analysis. Specifically, a heterogeneous graph is constructed to explicitly represent the spatial-scaling relationships and heterogeneity of multi-resolution patches. During the learning process, a RAConv is proposed to realize a more reliable node updating by considering the heterogeneity of the neighbors with different resolutions, and an IHPool module is designed to dynamically explore the task-related structured information of WSI, leading to performance improvement and richer interpretability. Extensive experiments validate the effectiveness of the proposed method. In the future, we will develop more computation-efficient strategy to accelerate the computation of the framework and evaluate our framework on other tasks.

Acknowledgments

This work was supported by Ministry of Science and Technology of the People's Republic of China (2021ZD0201900) (2021ZD0201903).

References

Cai, Z.; Song, H.; Fingerhut, A.; Sun, J.; Ma, J.; Zhang, L.; Li, S.; Yu, C.; Zheng, M.; and Zang, L. 2021. A greater lymph node yield is required during pathological examination in microsatellite instability-high gastric cancer. *BMC cancer*, 21(1): 1–9.

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

Carbonneau, M.-A.; Cheplygina, V.; Granger, E.; and Gagnon, G. 2018. Multiple instance learning: A survey of problem characteristics and applications. *Pattern Recognition*, 77: 329–353.

Chen, Z.; Zhang, J.; Che, S.; Huang, J.; Han, X.; and Yuan, Y. 2021. Diagnose Like A Pathologist: Weakly-Supervised Pathologist-Tree Network for Slide-Level Immunohistochemical Scoring. In *35th AAAI Conference on Artificial Intelligence (AAAI-21)*, 47–54. AAAI Press.

Courtiol, P.; Tramel, E. W.; Sanselme, M.; and Wainrib, G. 2018. Classification and disease localization in histopathology using only global labels: A weakly-supervised approach. *arXiv preprint arXiv:1802.02212*.

Cui, M.; and Zhang, D. Y. 2021. Artificial intelligence and computational pathology. *Laboratory Investigation*, 101(4): 412–422.

Farris, A. B.; Vizcarra, J.; Amgad, M.; Cooper, L. A.; Gutman, D.; and Hogan, J. 2021. Artificial intelligence and algorithmic computational pathology: an introduction with renal allograft examples. *Histopathology*, 78(6): 791–804.

Fey, M.; and Lenssen, J. E. 2019. Fast graph representation learning with PyTorch Geometric. *arXiv preprint arXiv:1903.02428*.

Gao, H.; and Ji, S. 2019. Graph u-nets. In *international conference on machine learning*, 2083–2092. PMLR.

Guo, D.; Shao, Y.; Cui, Y.; Wang, Z.; Zhang, L.; and Shen, C. 2021. Graph attention tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9543–9552.

Hashimoto, N.; Fukushima, D.; Koga, R.; Takagi, Y.; Ko, K.; Kohno, K.; Nakaguro, M.; Nakamura, S.; Hontani, H.; and Takeuchi, I. 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*, 3852–3861.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hou, W.; Wang, L.; Cai, S.; Lin, Z.; Yu, R.; and Qin, J. 2021. Early neoplasia identification in Barrett's esophagus via attentive hierarchical aggregation and self-distillation. *Medical image analysis*, 72: 102092.

Ilse, M.; Tomczak, J.; and Welling, M. 2018. Attention-based deep multiple instance learning. In *International conference on machine learning*, 2127–2136. PMLR.

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

Lee, J.; Lee, I.; and Kang, J. 2019. Self-attention graph pooling. In *International Conference on Machine Learning*, 3734–3743. PMLR.

Li, B.; Li, Y.; and 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*, 14318–14328.

Maron, O.; and Lozano-Pérez, T. 1998. Attention is all you need. In *Advances in neural information processing systems*, 570–576.

Ortega, S.; Fabelo, H.; Camacho, R.; De la Luz Plaza, M.; Callicó, G. M.; and Sarmiento, R. 2018. Detecting brain tumor in pathological slides using hyperspectral imaging. *Biomedical optics express*, 9(2): 818–831.

Riasatian, A.; Babaie, M.; Maleki, D.; Kalra, S.; Valipour, M.; Hemati, S.; Zaveri, M.; Safarpoor, A.; Shafiei, S.; Afshari, M.; et al. 2021. Fine-Tuning and training of densenet for histopathology image representation using TCGA diagnostic slides. *Medical Image Analysis*, 70: 102032.

Rice, T. W.; Patil, D. T.; and Blackstone, E. H. 2017. AJCC/UICC staging of cancers of the esophagus and esophagogastric junction: application to clinical practice. *Annals of cardiothoracic surgery*, 6(2): 119.

Tellez, D.; Litjens, G.; van der Laak, J.; and Ciompi, F. 2021. Neural Image Compression for Gigapixel Histopathology Image Analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(2): 567–578.

Tu, M.; Huang, J.; He, X.; and Zhou, B. 2019. Multiple instance learning with graph neural networks. *arXiv preprint arXiv:1906.04881*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in neural information processing systems*, 5998–6008.

Veličković, P.; Cucurull, G.; Casanova, A.; Romero, A.; Lio, P.; and Bengio, Y. 2017. Graph attention networks. *arXiv preprint arXiv:1710.10903*.

Wang, L.; Huang, Y.; Hou, Y.; Zhang, S.; and Shan, J. 2019a. Graph attention convolution for point cloud semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10296–10305.

Wang, S.; Chen, A.; Yang, L.; Cai, L.; Xie, Y.; Fujimoto, J.; Gazdar, A.; and Xiao, G. 2018. Comprehensive analysis of lung cancer pathology images to discover tumor shape and boundary features that predict survival outcome. *Scientific reports*, 8(1): 1–9.

Wang, X.; He, X.; Cao, Y.; Liu, M.; and Chua, T.-S. 2019b. Kgat: Knowledge graph attention network for recommendation. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 950–958.

Yang, T.; Hu, L.; Shi, C.; Ji, H.; Li, X.; and Nie, L. 2021. HGAT: Heterogeneous Graph Attention Networks for Semi-supervised Short Text Classification. *ACM Transactions on Information Systems (TOIS)*, 39(3): 1–29.

Yao, X.-H.; He, Z.-C.; Li, T.-Y.; Zhang, H.-R.; Wang, Y.; Mou, H.; Guo, Q.; Yu, S.-C.; Ding, Y.; Liu, X.; et al. 2020. Pathological evidence for residual SARS-CoV-2 in pulmonary tissues of a ready-for-discharge patient. *Cell research*, 30(6): 541–543.

Ying, R.; You, J.; Morris, C.; Ren, X.; Hamilton, W. L.; and Leskovec, J. 2018. Hierarchical graph representation learning with differentiable pooling. *arXiv preprint arXiv:1806.08804*.

Zhang, C.; Song, D.; Huang, C.; Swami, A.; and Chawla, N. V. 2019. Heterogeneous graph neural network. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 793–803.