Histopathology Whole Slide Image Analysis with Heterogeneous Graph Representation Learning WED-PM-315

Tsai Hor Chan[1] Fernando Julio Cendra[1]
Lan Ma[2] Guosheng Yin[1, 3] Lequan Yu[1]

[1]Department of Statistics and Actuarial Science
The University of Hong Kong

[2]TCL Corporate Research, Hong Kong

[3]Department of Mathematics
Imperial College London

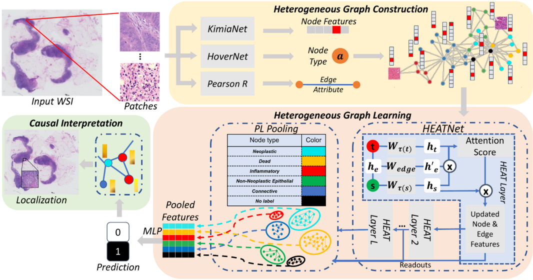
June 2023

Table of Contents

1. Summary
2. Introduction
3. Problem Definition
    - Definitions
    - The Problem
    - Contributions
4. Methodology
    - Instance Selections
    - Graph Construction
    - Model Design
    - Causal Interpretations
5. Experiments
    - Dataset
    - Comparable Methods
    - Results
    - Ablations

Summary

- We propose a novel framework for WSI analysis, which leverages a heterogeneous graph to learn the inter-relationships among different types of nodes and edges.
- We propose a graph aggregation algorithm which incorporate node and edge attributes in a heterogeneous graph, together with a pseudo-label pooling algorithm.
- We adopt a localization method based on Granger causality which shown improved performance.



Node type	Color
Neoplastic	🟦
Dead	🟨
Inflammatory	🟥
Non-Neoplastic Epithelial	🟩
Connective	🟦
No label	⬛

Table of Contents

1. Summary
2. Introduction
3. Problem Definition
   - Definitions
   - The Problem
   - Contributions
4. Methodology
   - Instance Selections
   - Graph Construction
   - Model Design
   - Causal Interpretations
5. Experiments
   - Dataset
   - Comparable Methods
   - Results
   - Ablations

Digital Histopathology

- In digital pathology, whole-slide scanners are used to digitize glass slides containing tissue specimens into whole-slide images (WSI) at high resolution (up to 160nm per pixel).
- It's time-consuming and tedious for pathologists to manually inspect a WSI due to the huge size (e.g., the usual size is 60,000 × 60,000) and complex patterns.
- Machine learning solutions are introduced to reduce the workload on pathologists.
- Recently the emergence of graph learning provides powerful solutions to WSI analysis.

Table of Contents

1 Summary

2 Introduction

3 Problem Definition

- Definitions
- The Problem
- Contributions

4 Methodology

- Instance Selections
- Graph Construction
- Model Design
- Causal Interpretations

5 Experiments

- Dataset
- Comparable Methods
- Results
- Ablations

Preliminary Definition

- **Heterogeneous Graph**: A heterogeneous graph is defined by a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{A}, \mathcal{R})$, where $\mathcal{V}, \mathcal{E}, \mathcal{A}$ represent the set of entities (vertices or nodes), relations (edges), and entity types, respectively. And $\mathcal{R}$ represents the space of edge attributes. For $v \in \mathcal{V}$, v is mapped to an entity type by a function $\tau(v) \in \mathcal{A}$. An edge $e = (s, r, t) \in \mathcal{E}$ links the source node s and the target node t, and r is mapped to an edge attribute by a function $\phi(e) = r \in \mathcal{R}$. Every node v has a d-dimensional node feature $x \in \mathcal{X}$, where $\mathcal{X}$ is the embedding space of node features.
- **Granger Causality** Granger [1969], Lin et al. [2021]: Let $\mathcal{I}$ be all the available information and $\mathcal{I}_{-X}$ be the information excluding variable X. If we can make a better prediction of Y using $\mathcal{I}$ than using $\mathcal{I}_{-X}$, we conclude that X Granger-causes Y.

Problem — WSI Classification

WSI Classification: Given a WSI X and a heterogeneous graph $\mathcal{G}$ constructed from X, we wish to predict the label y with a GNN model $\mathcal{M}$. We also aim to assign an importance score $f(v)$ to each node $v \in \mathcal{V}$ in $\mathcal{G}$ as the causal contribution of each patch to the prediction for localization.

Contributions

- We propose a novel framework for WSI analysis, which leverages a heterogeneous graph to learn the inter-relationships among different types of nodes and edges.
- The heterogeneous graph introduces a "nucleus-type" attribute to each node, which can serve as an effective data structure for modeling the structural interactions among the nuclei in the WSI.
- To tackle the aggregation process in the heterogeneous graph, we propose a novel heterogeneous-graph edge attribute transformer (HEAT) architecture which can take advantage of the edge and node heterogeneity. Thus, the diverse structural relations among different biological entities in the WSI can be incorporated to guide the GNN for more accurate prediction.

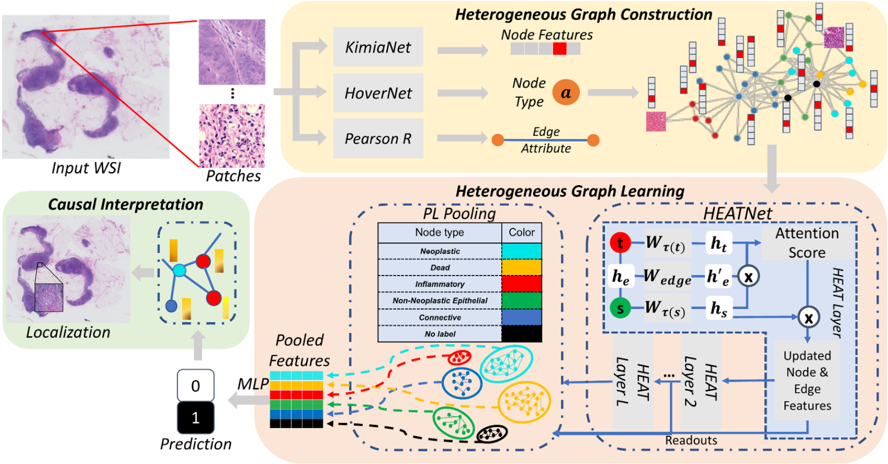
Contributions

- Further, to obtain the graph-level representations for slide-level prediction, we propose a semantic-consistent pooling mechanism — pseudo-label (PL) pooling, which pools node features to graph level based on clusters with a fixed definition (i.e., nucleus type). The proposed PL pooling can regularize the graph pooling process by distilling the context knowledge (i.e., pathological knowledge) from a pretrained model to alleviate the over-parameterization issue [Balaji et al., 2020].
- Additionally, we propose a Granger causality [Granger, 1969] based localization method to identify the potential regions of interest with clinical relevance to provide more insights to pathologists and promote the clinical usability of our approach.

Table of Contents

1. Summary
2. Introduction
3. Problem Definition
   - Definitions
   - The Problem
   - Contributions
4. Methodology
   - Instance Selections
   - Graph Construction
   - Model Design
   - Causal Interpretations
5. Experiments
   - Dataset
   - Comparable Methods
   - Results
   - Ablations

The Framework

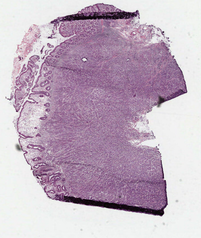


Node type	Color
Neoplastic	🟦
Dead	🟨
Inflammatory	🟥
Non-Neoplastic Epithelial	🟩
Connective	🟦
No label	⬛

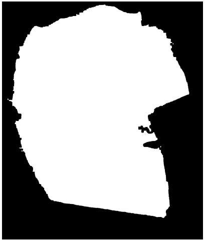
Figure: The workflow of our proposed framework

Instance Selections

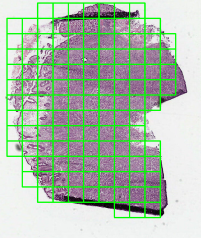
Three procedures were used to generate the heterogeneous graphs i.e., Otsu's thresholding method [Otsu, 1979] to automatically segment the nuclei from input histopathology images.



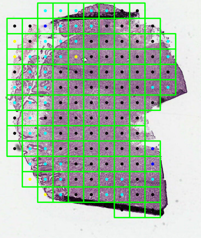
Input WSI data



Segmented mask



Extracted patches



Predicted node types

Instance Selections

Patch extraction

- Given the segmented masks from Otsu's thresholding [Otsu, 1979] with a magnification factor of 20, it generates a patch-level image with a patch size of 256x256.
- Output: uniform patch-level image with its corresponding patch coordinates.

Node type prediction

- Node prediction: use HoverNet's nuclei classifier [Graham et al., 2019] (pretrained using PanNuke dataset).
- Node type assignment per patch: use majority vote operation to find the most frequent node type in a patch.

Instance Selections

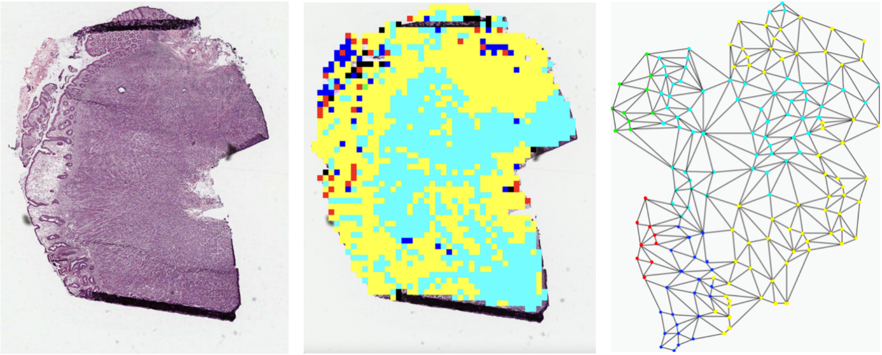


Figure: A WSI with selected patches and predicted node types

Model WSI with Heterogeneous Graph

- Node construction: each patch is a node with node types predicted by Hovernet.
- Edge construction: use Pearson R to determine the correlation between the feature vectors of the nodes.
- We then have a heterogeneous graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{A}, \mathcal{R})$ to model the WSI image in a graphical manner.
- We design a novel architecture to propagate information on $\mathcal{G}$ and predict image-level label $\hat{y}$.

Meta-Relations in a WSI

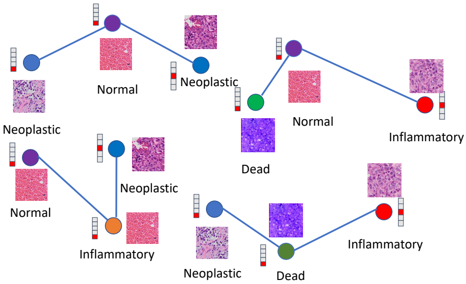


Figure: Examples of meta-relations in a heterogeneous graph constructed from a WSI.

The HEAT Algorithm

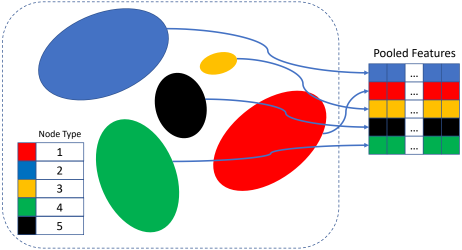
We propose a heterogeneous edge attribute transformer (HEAT) layer to incorporate the continuous edge features in a heterogeneous graph.

Algorithm The HEAT algorithm.

```
Input:
Heterogeneous graph $\mathcal{G}_{l-1}$ with node features $\{H_i^{(l-1)}, \forall i \in \mathcal{V}\}$ and
edge attribute $\{h_e^{(l-1)}, \forall e \in \mathcal{E}\}$;
Node-type specific projection layers $\{\boldsymbol{W}_a^i, \forall a \in \mathcal{A}\}$
Edge attribute transformation layer $W_{\text{edge}}$.
Output: The updated graph $\mathcal{G}_l$ with node features $\{H_i^{(l)}, \forall i \in \mathcal{V}\}$, and
the edge features $\{h_e^{(l)}, \forall e \in \mathcal{E}\}$
Initialize projection layers for each node type
for $e = (s, t) \in \mathcal{E}$ do
    $\boldsymbol{h}_{\text{key}}^i = \boldsymbol{W}_{\tau(s)}^i H_s^{(l-1)}$                ▷ Project the source node
    $\boldsymbol{h}_{\text{value}}^i = \boldsymbol{W}_{\tau(s)}^i H_s^{(l-1)}$              ▷ Compute value vector
    $\boldsymbol{h}_{\text{query}}^i = \boldsymbol{W}_{\tau(t)}^i H_t^{(l-1)}$              ▷ Project the target node
    $h'_e \leftarrow W_{\text{edge}} \cdot h_e^{(l-1)}$                   ▷ Project the edge attribute
    $\text{ATT}(e, i) = \left(\boldsymbol{h}_{\text{key}}^i h'_e \boldsymbol{h}_{\text{query}}^i\right) / \sqrt{d}$
    $\text{Attention}(e) = \underset{\forall s \in N(t)}{\text{softmax}}(\|_{i \in [1,h]} \text{ATT}(e, i))$
    $h_e^{(l)} \leftarrow h'_e$                                  ▷ Compute latent edge features
end for
for $t \in \mathcal{V}$ do
    $H_t^{(l)} = \oplus_{\forall s \in N(t)}(\|_{i \in [1,h]} \boldsymbol{h}_{\text{value}}^i \cdot \text{Attention}(e))$
end for
return $\mathcal{G}_l$
```

Pooling by Pseudo Labels

We introduce a novel pooling method — PL Pool, to aggregate information with respect to the pseudo-labels (i.e., node types) predicted from a pretrained teacher network (e.g., HoverNet Graham et al. [2019]).



Node Type	
🟥	1
🟦	2
🟨	3
🟩	4
⬛	5

Figure: Mechanism of Pseudo-label Pool

Model Training

The predicted label from the network is

$$\hat{y} = \text{softmax}(\sum_{l=1}^{L} \text{readout}(HEAT(\mathcal{G}_l))),$$

where readout is an arbitrary pooling method (e.g., average pooling). We adopt the cross-entropy loss to train the network and the objective is defined as the loss function

$$\mathcal{L} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{K} y_{ij} \log(\hat{y}_{ij}),$$

where N is the number of samples, K is the number of the classes, and $y \in \mathbb{R}^{N \times K}$ are the one-hot labels.

Causal Interpretations

We make use of the Granger causality to outline causal regions in the WSI with the causal graph explainer. The causal contribution of each node v is given by [Lin et al., 2021]

$$\Delta_v = \mathcal{L}(y, \tilde{y}_{\mathcal{G}}) - \mathcal{L}(y, \tilde{y}_{\mathcal{G}\backslash\{v\}}),$$

where y is the true label and $\tilde{y}_{\mathcal{G}} = \mathcal{M}(\mathcal{G})$ and $\tilde{y}_{\mathcal{G}\backslash\{v\}} = \mathcal{M}(\mathcal{G}\backslash\{v\})$ are the predicted labels from the GNN $\mathcal{M}$ with input graphs $\mathcal{G}$ and $\mathcal{G}\backslash\{v\}$, respectively. $\mathcal{L}(y, \hat{y})$ is the cross-entropy loss between the ground-truth label y and the predicted label $\hat{y}$.

Table of Contents

1. Summary
2. Introduction
3. Problem Definition
    - Definitions
    - The Problem
    - Contributions
4. Methodology
    - Instance Selections
    - Graph Construction
    - Model Design
    - Causal Interpretations
5. Experiments
    - Dataset
    - Comparable Methods
    - Results
    - Ablations

Datasets

Table: The distribution of classes in TCGA-COAD, TCGA-BRCA, and Camelyon16 datasets.

Classification sets	Tumor data		Normal data	
TCGA-COAD	1325		99	
TCGA-BRCA	1365		347	
Camelyon 16	160		239	
Staging sets	Stage I	Stage II	Stage III	Stage IV
TCGA-COAD	267	561	397	209
TCGA-BRCA	276	967	368	37
Typing sets	Type I		Type II	
TCGA-BRCA	190		30	
TCGA-ESCA	89		65	

Comparable Methods

- ABMIL [Ilse et al., 2018]: an MIL framework aggregating bag-level instance information by attention mechanism.
- DSMIL [Li et al., 2021]: a dual-stream multiple instance learning method using max pooling and attention to aggregate the signals from the individual patches.
- GTNMIL [Zheng et al., 2022]: a graph-based MIL method based on graph transformer network [Yun et al., 2019].
- Patch GCN [Chen et al., 2021]: a hierarchical graph-based model on survival data with patient-level and WSI-level aggregations. We adapt this method as a GCN model with Global attention pooling [Li et al., 2015].
- H^2-MIL [Hou et al., 2022]: a tree-graph-based multiple instance learning that utilizes different magnification levels to represent hierarchical features.

Results — TCGA–COAD and TCGA–BRCA

		Cancer Staging (Four Stages)			**Cancer Classification**		
	Model	**AUC**	**Accuracy**	**Macro-F1**	**AUC**	**Accuracy**	**Macro-F1**
TCGA–COAD	ABMIL Ilse et al. [2018]	53.8 (3.7)	19.2 (7.8)	35.8 (4.4)	97.7 (2.3)	98.3 (0.9)	95.8 (2.2)
	DSMIL Li et al. [2021]	59.3 (1.4)	35.7 (5.7)	37.9 (2.8)	99.7 (0.2)	98.6 (0.5)	96.9 (0.9)
	ReMix Yang et al. [2022]	58.3 (1.5)	33.9 (7.8)	24.8 (7.5)	94.3 (3.4)	96.0 (4.6)	92.8 (5.9)
	PatchGCN Chen et al. [2021]	62.5 (4.9)	38.2 (3.1)	38.5 (5.7)	91.1 (5.3)	97.1 (2.0)	98.8 (1.0)
	GTNMIL Zheng et al. [2022]	54.2 (2.6)	29.3 (1.4)	24.3 (3.9)	97.3 (2.6)	98.1 (1.3)	95.9 (2.4)
	H^2-MIL Hou et al. [2022]	58.6 (2.7)	38.5 (5.4)	33.0 (5.0)	99.7 (0.4)	99.2 (0.5)	97.4 (1.7)
	HEAT (Ours)	**63.4 (2.5)**	**40.0 (2.1)**	**41.3 (2.7)**	**99.9 (0.2)**	**99.9 (0.3)**	**99.2 (0.4)**
TCGA–BRCA	ABMIL Ilse et al. [2018]	54.7 (4.6)	19.0 (10.0)	23.9 (3.2)	97.3 (1.7)	98.3 (1.1)	97.3 (1.6)
	DSMIL Li et al. [2021]	51.4 (4.7)	18.3 (14.9)	23.2 (2.3)	98.7 (0.5)	95.6 (1.4)	93.3 (2.0)
	ReMix Yang et al. [2022]	58.8 (2.2)	35.6 (16.2)	27.6 (5.8)	96.1 (0.7)	95.8 (2.6)	93.0 (3.4)
	PatchGCN Chen et al. [2021]	50.3 (0.2)	41.6 (0.5)	25.1 (0.3)	96.2 (1.7)	98.2 (0.8)	98.4 (0.8)
	GTNMIL Zheng et al. [2022]	53.0 (3.7)	41.3 (4.4)	25.1 (2.3)	94.7 (1.0)	94.5 (0.2)	93.7 (1.7)
	H^2-MIL Hou et al. [2022]	52.1 (7.2)	53.7 (2.6)	21.2 (2.5)	97.9 (2.7)	98.0 (1.5)	97.6 (2.2)
	HEAT (ours)	**61.9 (3.8)**	**55.8 (6.4)**	**27.7 (16.3)**	**98.8 (0.7)**	**98.3 (0.5)**	**99.5 (0.7)**

Table: Cancer staging and classification results [%] of various methods on TCGA–COAD and TCGA–BRCA datasets.

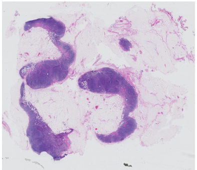
Results — TCGA–ESCA

Model	AUC	Accuracy	Macro-F1
ABMIL Ilse et al. [2018]	79.5 (7.5)	80.3 (8.4)	81.3 (7.4)
DSMIL Li et al. [2021]	92.5 (1.7)	87.3 (2.0)	86.3 (2.0)
ReMix Yang et al. [2022]	92.5 (7.2)	90.0 (8.1)	90.3 (7.7)
PatchGCN Chen et al. [2021]	88.6 (3.5)	92.1 (2.3)	92.3 (2.4)
GTNMIL Yun et al. [2019]	89.7 (4.7)	81.2 (4.8)	89.2 (4.9)
H^2-MIL Hou et al. [2022]	92.1 (3.9)	88.2 (5.8)	88.0 (5.8)
HEAT (ours)	**92.8 (2.5)**	**92.7 (2.2)**	**93.3 (1.9)**

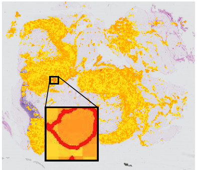
Table: Cancer typing results [%] of our method compared to various methods on the TCGA–ESCA dataset.

Results — Qualitative Evaluation

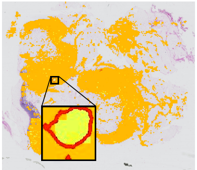
We modified the GNN Explainers to apply masks on nodes (instead of features in the original paper) to calculate the contributions of each node. We use GNN Explainer as the baseline method to interpret the important regions to cancer prediction.



(a) Original WSI



(b) GNN Explainer



(c) Causal Explainer

Ablation — GNN Architectures

GNN Architecture	**AUC**	**Accuracy**	**Macro-F1**
GCN Welling and Kipf [2016]	90.8	90.9	90.0
GAT Veličković et al. [2017]	85.8	86.4	88.9
GIN Xu et al. [2018]	91.6	90.9	83.3
HetRGCN Schlichtkrull et al. [2018]	82.5	83.3	88.9
HGT Hu et al. [2020]	87.8	87.5	83.3
HEAT (ours)	**92.8**	**92.7**	**93.2**

Table: Cancer typing results [%] of our method compared to various GNN architectures on the TCGA–ESCA dataset.

Ablations — Pooling Methods

- We perform binary classification on the COAD dataset compared to pooling methods.

Table: Cancer classification results on TCGA–COAD of our pooling method to various comparable pooling methods using GCN and KimiaNet feature encoder.

Method	**Accuracy**	**Macro-F1**	**AUC**
Sum pooling	99.3	99.2	95.5
Max pooling	98.6	99.2	95.1
Mean pooling	95.8	100.0	97.7
Global attention pooling Li et al. [2015]	97.9	99.2	94.7
IH-Pool Hou et al. [2022]	97.2	88.1	99.3
ASAP Ranjan et al. [2020]	98.6	95.1	99.2
PL-Pool (Ours)	**99.3**	**100.0**	**99.6**

References I

Yogesh Balaji, Mohammadmahdi Sajedi, Neha Mukund Kalibhat, Mucong Ding, Dominik Stöger, Mahdi Soltanolkotabi, and Soheil Feizi. Understanding over-parameterization in generative adversarial networks. In *International Conference on Learning Representations*, 2020.

Richard J Chen, Ming Y Lu, Muhammad Shaban, Chengkuan Chen, Tiffany Y Chen, Drew FK Williamson, and Faisal Mahmood. Whole slide images are 2d point clouds: Context-aware survival prediction using patch-based graph convolutional networks. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 339–349. Springer, 2021.

Simon Graham, Quoc Dang Vu, Shan E Ahmed Raza, Ayesha Azam, Yee Wah Tsang, Jin Tae Kwak, and Nasir Rajpoot. Hover-net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images. *Medical Image Analysis*, 58:101563, 2019.

References II

Clive WJ Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: journal of the Econometric Society*, pages 424–438, 1969.

Wentai Hou, Lequan Yu, Chengxuan Lin, Helong Huang, Rongshan Yu, Jing Qin, and Liansheng Wang. H2-mil: Exploring hierarchical representation with heterogeneous multiple instance learning for whole slide image analysis. 2022.

Ziniu Hu, Yuxiao Dong, Kuansan Wang, and Yizhou Sun. Heterogeneous graph transformer. In *Proceedings of The Web Conference 2020*, pages 2704–2710, 2020.

Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *International conference on machine learning*, pages 2127–2136. PMLR, 2018.

References III

Bin Li, Yin Li, and Kevin 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*, pages 14318–14328, 2021.

Yujia Li, Daniel Tarlow, Marc Brockschmidt, and Richard Zemel. Gated graph sequence neural networks. *Proceedings of ICLR'16*, 2015.

Wanyu Lin, Hao Lan, and Baochun Li. Generative causal explanations for graph neural networks. In *International Conference on Machine Learning*, pages 6666–6679. PMLR, 2021.

Nobuyuki Otsu. A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, 9(1):62–66, 1979. doi: 10.1109/TSMC.1979.4310076.

References IV

Ekagra Ranjan, Soumya Sanyal, and Partha Talukdar. Asap: Adaptive structure aware pooling for learning hierarchical graph representations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 5470–5477, 2020.

Michael Schlichtkrull, Thomas N Kipf, Peter Bloem, Rianne van den Berg, Ivan Titov, and Max Welling. Modeling relational data with graph convolutional networks. In *European semantic web conference*, pages 593–607. Springer, 2018.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *International Conference on Learning Representations*, 2017.

Max Welling and Thomas N Kipf. Semi-supervised classification with graph convolutional networks. In *J. International Conference on Learning Representations (ICLR 2017)*, 2016.

References V

Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? In *International Conference on Learning Representations*, 2018.

Jiawei Yang, Hanbo Chen, Yu Zhao, Fan Yang, Yao Zhang, Lei He, and Jianhua Yao. Remix: A general and efficient framework for multiple instance learning based whole slide image classification. *Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part II*, 2022.

Seongjun Yun, Minbyul Jeong, Raehyun Kim, Jaewoo Kang, and Hyunwoo J Kim. Graph transformer networks. *Advances in neural information processing systems*, 32, 2019.

References VI

Yi Zheng, Rushin H Gindra, Emily J Green, Eric J Burks, Margrit Betke, Jennifer E Beane, and Vijaya B Kolachalama. A graph-transformer for whole slide image classification. *IEEE transactions on medical imaging*, 41(11):3003–3015, 2022.