Integrative Graph-Transformer Framework for Histopathology Whole Slide Image Representation and Classification

Zhan Shi[1], Jingwei Zhang[1], Jun Kong[2], and Fusheng Wang[1]

[1]Stony Brook University, USA
[2]Georgia State University, USA
{shzhan, jingwezhang, fushwang}@cs.stonybrook.edu
jkong@gsu.edu

Abstract. In digital pathology, the multiple instance learning (MIL) strategy is widely used in the weakly supervised histopathology whole slide image (WSI) classification task where giga-pixel WSIs are only labeled at the slide level. However, existing attention-based MIL approaches often overlook contextual information and intrinsic spatial relationships between neighboring tissue tiles, while graph-based MIL frameworks have limited power to recognize the long-range dependencies. In this paper, we introduce the integrative graph-transformer framework that simultaneously captures the context-aware relational features and global WSI representations through a novel Graph Transformer Integration (GTI) block. Specifically, each GTI block consists of a Graph Convolutional Network (GCN) layer modeling neighboring relations at the local instance level and an efficient global attention model capturing comprehensive global information from extensive feature embeddings. Extensive experiments on three publicly available WSI datasets: TCGA-NSCLC, TCGA-RCC and BRIGHT, demonstrate the superiority of our approach over current state-of-the-art MIL methods, achieving an improvement of 1.0% to 2.6% in accuracy and 0.7%-1.6% in AUROC.

Keywords: Whole slide image classification · Multiple instance learning· Graph Transformer

1 Introduction

With the significant advance in high-throughput whole slide tissue scanning technology, digital pathology leverages high-quality whole slide images (WSIs) and is an actively developing component in pathology study [24]. As WSIs are often giga-pixels by scale and lack of pixel-level annotations, an efficient and effective way to analyze such high-resolution WSIs becomes critical to facilitate cancer diagnosis and prognosis. Due to the remarkable performance, deep-learning based multiple instance learning (MIL) is often employed in such weakly-supervised scenarios where only slide-level labels are available [12,14,15,30]. By the MIL scheme, each image patch or instance is first encoded as a feature embedding using a pretrained feature extractor [15]. These embeddings are next passed to an aggregator module that compiles embeddings into a comprehensive bag-level representation before the classification [27].

Multiple digital pathology studies in the MIL framework adopt attention mechanisms and achieve promising results with the global WSI representations [13,21]. However, these methods assume that all instances are independent and thus ignores the critical correlations across different tissue regions. The self-attention mechanism from vision transformers (ViT) [10] has been used to address this problem, where pairwise similarity scores across all instances are computed [26,5,28]. However, such pairwise calculation exhibits quadratic complexity, often too demanding to support a large number of input instances for the WSI classification. The Nyström-attention [29] has been applied to alleviate this problem in TransMIL [26]. It utilizes a subset of landmarks to approximate the self-attention process. Similarly, FlashAttention [8] achieves the full self-attention ability and uses the IO-aware mechanism to enhance the attention efficiency.

Besides the correlation across instances, the tissue spatial relationship is crucial for the WSI analysis. However, it is often overlooked in existing MIL studies. The use of position encoding in Transformers for fixed-length sequences [10] can preserve positional information, but it cannot be directly used in the WSI analysis due to the variable lengths of input instance embeddings. To address this, TransMIL [26] employ Convolution Neural Network(CNN) to characterize the spatial information. However, it reorganizes the tissue patches and thus does not accurately reflect the genuine spatial relationships among patches. Consequently, the inherent potential for spatial arrangement within WSI has not been thoroughly explored.

By contrast, the graph structure is widely known for its intrinsic merit for spatial relationship representations and graph-based MIL methods have increasingly gained attention for the histopathology WSI analysis. The Graph Convolution Network (GCN) utilizes a foundational local message-passing mechanism to capture spatial interactions and integrate neighboring instances and provides a cutting-edge graph-based paradigm for the digital pathology study. However, such GCN-based frameworks may suffer from over-smoothing [16] due to the repeated aggregation of local information, and over-squashing [2] as a result of the increased model depth. Moreover, graph-based MIL frameworks exhibit limitations in recognizing long-range dependency.

Recent research has demonstrated that integrating self-attention mechanisms into graph-based approach can effectively mitigate the limitations of message-passing mechanism, such as over-squashing and over-smoothing, thereby enhancing the model's capability for representation [25]. Furthermore, the application of graph transformers has extended to multi-modal, multi-task, and multi-scale analysis of WSIs [23,31,9]. The GTP [32] has been developed for WSI classification which employs a clustering-based mincutpool [3] to bridge GCN and transformer layers. However, the GCN layers in GTP are still prone to over-squashing, and the inevitable information loss from the pooling layer constrains the transformer's capabilities.

To alleviate these limitations, we develop a novel Integrative Graph-Transformer (IGT) framework for WSI representation and classification. The core architecture of the IGT framework consists of a sequence of graph transformer integration blocks, where each block integrates a GCN layer for encoding spatial relationships among adjacent instances and a global attention module capturing global WSI representations. Our framework is able to simultaneously models spatial relationships at the local instance level and long-range pairwise correlations across all instances. We demonstrate the efficacy of our method on three public WSI datasets, TCGA-NSCLC, TCGA-RCC and BRIGHT. With extensive testing on these datasets, our IGT framework presents a superior performance to the state-of-the-art methods, achieving a 1.0% to 2.6% improvement in accuracy and a 0.7% and 1.6% increase in AUROC.

2 Method

Illustrated in Fig. 1, the proposed IGT framework consists of three key components: graph construction, the backbone, and the downstream process. During the graph construction, feature vectors are extracted, and the corresponding adjacency matrix is created. The backbone module processes this undirected WSI graph representation, serving as an efficient encoder. Finally, the refined features from the last GTI block are provided to the downstream model for classification.

2.1 Graph Construction

For each WSI graph G construction, we first partition a WSI into non-overlapping 256×256 tissue region patches/instances. Note that the number of extracted instances N varies for different WSIs. A ResNet50 [11] model pre-trained on ImageNet is used to encode each instance into a 1024-dimensional feature vector $\{\mathbf{h}_i \in \mathbb{R}^{1024}, i = 1, 2...N\}$. Each feature is regarded as a node in the WSI graph and we assemble instance feature vectors as the node feature matrix $\{\mathbf{H} \in \mathbb{R}^{N\times 1024}\}$ for each WSI. To depict the node connectivity in the WSI graph, we preserve the patch spatial coordinates in the WSI and find adjacent nodes by the K-Nearest Neighbor algorithm (i.e. *k-NN*, k=8) [7]. Thus, we build the WSI graph $G = (\mathbf{H}, \mathbf{A})$ in the Euclidean space, where $\{\mathbf{A} = [A_{ij}], \mathbf{A} \in \mathbb{R}^{N\times N}\}$ is the adjacency matrix. Its entry $A_{ij} = 1$ when there exists a connection between node i and j by the *k-NN* algorithm on node feature representations $(\mathbf{h}_i, \mathbf{h}_j)$. Otherwise, $A_{ij} = 0$. This graph models the local neighborhood information across the entire WSI.

2.2 Graph-Transformer Integration Block

The spatial relationships across tissue instances are crucial for the WSI representation and classification [1]. Therefore, we design the GTI block to concurrently aggregate the local instance relationships and capture long-range pairwise correlations across the entire tissue domain.

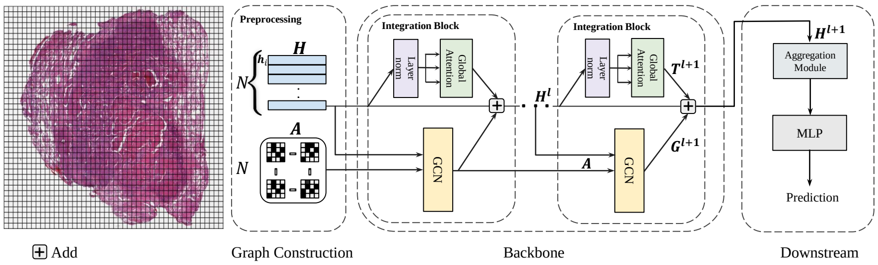

Fig. 1. Overview of the proposed method: The process begins with the graph construction module where a graph representation $G = (\mathbf{H}, \mathbf{A})$ is generated for the subsequent backbone network. Here, the $\mathbf{H}$ is the feature matrix and $\mathbf{A}$ denotes the associated adjacent matrix. Within each l-th integration block of the backbone, a global attention layer processes $\mathbf{H}^l$ to produce the feature matrix $\mathbf{T}^{l+1}$, and a GCN layer processes both $\mathbf{H}^l$ and $\mathbf{A}$ to update the graph representation $\mathbf{G}^{l+1}$. Finally, the integrated feature $\mathbf{H}^{l+1}$ from the last block is utilized for prediction in the downstream module.

As depicted in the integration block (Fig. 1), the l-th GTI block operates on the GCN and $GlobalAttention$ layers in parallel and integrates their outputs through a simple summation as follows:

$$\mathbf{G}^{l+1} = GCN\left(\mathbf{H}^l,\ \mathbf{A}\right) \tag{1}$$

$$\mathbf{T}^{l+1} = GlobalAttn\left(\mathbf{H}^l\right) \tag{2}$$

$$\mathbf{H}^{l+1} = GTI\left(\mathbf{H}^l, \mathbf{A}\right) = \mathbf{G}^{l+1}\ +\ \mathbf{T}^{l+1} \tag{3}$$

where $\mathbf{G}^{l+1} \in \mathbb{R}^{N \times d}$ is the generated graph representation, $\mathbf{T}^{l+1} \in \mathbb{R}^{N \times d}$ is the global attention feature matrix and d is the dimension of the feature embedding.

(1) *GCN:* The message passing functions of the general GCN operator, acting on the local neighborhood of node u at l-th layer, can be represented as follows:

$$\mathbf{m}_u^l = AGG\left(\left\{\mathbf{m}_{uv}^l = \rho\left(\mathbf{h}_u^l, \mathbf{h}_v^l, \mathbf{h}_{e_{uv}}^l\right),\ v \in \mathcal{N}(u)\right\}\right) \tag{4}$$

$$\mathbf{g}_u^{l+1} = \phi\left(\mathbf{h}_u^l,\ \mathbf{m}_u^l\right) \tag{5}$$

where ρ, AGG, and ϕ are differentiable functions. The message construction function ρ constructs a message for node u by integrating the node u feature $\mathbf{h}_u^l$, features of its neighbors $\mathbf{h}_v^l$, and the edge features $\mathbf{h}_{e_{uv}}^l$. The AGG is a permutation invariant function that aggregates all messages directed towards node u. In essence, the AGG function executes MIL-manner operations within a graph's local neighborhood. The resulting feature $\mathbf{g}_u^{l+1}$ of node u is then updated by merging the original node feature $\mathbf{h}_u^l$ and the aggregated message $\mathbf{m}_u^l$ via the update function ϕ. As the choice of these GCN related functions is flexible, we adopt the generalized graph convolution $GENConv$ from the DeeperGCN [18].

The corresponding message passing functions are defined as follows:

$$\mathbf{m}^l_{uv} = \ ReLU\left(\mathbf{h}^l_v + \mathbf{1}\left(\mathbf{h}^l_{e_{uv}}\right)\cdot\ \mathbf{h}^l_{e_{uv}}\right) \ + \ \epsilon \tag{6}$$

$$\mathbf{m}^l_u \ = \sum_{v\in\mathcal{N}(u)} \frac{exp\left(\beta\mathbf{m}^l_{uv}\right)}{\sum\limits_{v\in\mathcal{N}(u)} exp\left(\beta\mathbf{m}^l_{uv}\right)}\cdot\mathbf{m}^l_{uv} \tag{7}$$

$$\mathbf{g}^{l+1}_u = \ \phi\left(\mathbf{h}^l_u,\ \mathbf{m}^l_u\right) = \ MLP\left(\mathbf{h}^l_u\ +\mathbf{m}^l_u\right) \tag{8}$$

The message is constructed by a ReLU activation function with the neighboring node feature $\mathbf{h}^l_v$ and the associated edge feature between node u and v where $\mathbf{1}(\cdot)$ is an indicator function. A small positive constant ($\epsilon = 10^{-7}$) is added to the ReLU activation function output to ensure positive feature values for the numerical stability. The resulting messages from neighboring nodes are summed with weights by the SoftMax function where hyper-parameter β denotes the inverse temperature. This aggregation method concentrates on the local instance interactions. Finally, the update function is structured as a two-layer MLP. These configurations ensure an effective feature transformation and message propagation.

(2) *Global Attention:* While GNNs can be used to describe the entire WSI graph, they can be constrained for long-range dependency characterization due to the limited receptive field. Although an increase in a GNN depth could be a potential remedy, it can result in indistinguishable node representations, an issue known as over-smoothing or over-squashing. To alleviate these problems, we implement a global attention modules in parallel to the GCN (Fig. 1). This design enhances the ability to identify discriminating node representations from the entire WSI graph. Specifically, the global attention layer employs the self-attention mechanism, with its formulation given below:

$$GlobalAttn(\mathbf{Q},\mathbf{K},\mathbf{V}) \ = \ softmax\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_q}}\right)\mathbf{V} \tag{9}$$

where feature representations $\mathbf{Q},\mathbf{K},\mathbf{V}$ are calculated by projecting instance feature matrix $\mathbf{H}$ using distinct three weight matrix $\mathbf{W}_i \ \in \ \mathbb{R}^{d\times d_i}$. While the self-attention mechanism in the original transformer is effective and well-suited for this scenario, its $O(N^2)$ computational complexity limits its ability to process a large number of input instances efficiently. To address this limitation, we leverage FlashAttention (FA) [8] to fully harness the potential of the multi-head self-attention mechanism without information loss or an expensive computational cost. Integrating global feature embeddings with those from the GCN branch, we produce effective and expressive WSI representations that are able to capture the global contextual information and the local neighbor interactions.

After the feature processing via GTI blocks, a straightforward attention-based MIL pooling [13] strategy is used for feature aggregation in the downstream phase in Fig. 1. The resulting bag-level representation $\mathbf{h_{bag}} \ \in \ \mathbb{R}^{1\times d}$ is computed by the weighted average of the instance representations by the attention scores α as follows:

$$\mathbf{h_{bag}} \ = \ \alpha^T \mathbf{H}^L \tag{10}$$

In the final phase, the bag-level feature $\mathbf{h_{bag}}$ is provided to the MLP layer to achieve the final bag-level classification.

3 Experiments

3.1 Datasets

To demonstrate the efficacy of our novel IGT framework, we conduct experiments and compare our method with SOTA methods on three widely used public datasets: TCGA-NSCLC (The Caner Genome Atlas Non-Small Cell Lung Cancer), TCGA-RCC (Renal Cell Carcinoma) and BRIGHT [4]. We use the official data split if it is available, otherwise, we split the train, validation, and test sets by an ratio of 6.5:1.5:2.0. All WSIs in these datasets are cropped at 20× magnification.

TCGA-NSCLC is a lung cancer dataset and includes two distinct cancer subtypes: Lung Adenocarcinoma (LUAD) and Lung Squamous Cell Carcinoma (LUSC). It has 1,043 diagnostic digital WSIs with 531 and 512 WSIs of LUAD and LUSC, respectively. We follow the same random split for DSMIL study [17].

TCGA-RCC is a kidney cancer dataset and consists of 940 WSIs. Specifically, there are 121 WSIs of 109 Kidney Chromophobe Renal Cell Carcinoma (TCGA-KICH) cases, 519 WSIs of 513 Kidney Renal Clear Cell Carcinoma (TCGA-KIRC) cases, and 300 WSIs of 276 Kidney Renal Papillary Cell Carcinoma (TCGA-KIRP) cases.

BRIGHT is a breast cancer dataset and contains 503 diagnostic slides across six breast tumor subtypes: Pathological Benign (PB), Usual Ductal Hyperplasia (UDH), Flat Epithelia Atypia (FEA), Atypical Ductal Hyperplasia (ADH), Ductal Carcinoma in Situ (DCIS), and Invasive Carcinoma (IC). We use the official data split, where 423 WSIs are for training and 80 WSIs for testing.

3.2 Implementation Details

In the graph construction phase, background patches with a saturation level of less than 15 are discarded. The processed 1024-dimensional feature vector $h_i \in \mathbb{R}^{1024}$ is downscaled to 256 and assembled for the node feature matrix $H \in R^{N \times 256}$ [22,26], before being taken as input. For model training, the cross-entropy loss function is utilized, and the batch size is set to 1. We adopt the Rectified Adam optimizer [20] for optimization with a weight decay of 1e-5. We train the IGT framework for 40 epochs on both TCGA-NSCLC and TCGA-RCC datasets, and for 30 epochs on BRIGHT dataset. The learning rate starts at 1e-3, decaying to 1e-4 at epoch 20 for TCGA-NSCLC, and at epoch 15 for TCGA-RCC and BRIGHT. We employ two GTI blocks for TCGA-NSCLC and TCGA-RCC, and three GTI blocks for BRIGHT. All models are implemented by PyTorch 2.0, and executed on an NVIDIA GeForce RTX 3090Ti GPU.

Table 1. Comparison of accuracy and AUROC on three public datasets. The reported metrics are presented as percentages and averaged for three times. Our IGT framework consistently outperforms existing state-of-art MIL methods.

Method	TCGA-NSCLC		TCGA-RCC		BRIGHT-6class	
	ACC(%)	AUC(%)	ACC(%)	AUC(%)	ACC(%)	AUC(%)
Mean-pooling	77.6	86.2	82.3	94.2	26.1	64.1
Max-pooling	79.0	85.8	84.0	96.1	29.3	66.0
ABMIL[13]	84.1	91.3	86.8	97.1	30.8	67.0
DSMIL[17]	86.0	93.9	87.7	97.7	36.4	72.5
CLAM-SB[22]	85.5	90.9	88.5	98.0	33.1	69.1
CLAM-MB[22]	87.9	92.9	89.9	97.9	38.1	71.7
TransMIL[26]	89.3	94.2	90.2	97.7	39.6	71.8
GCN-ABMIL[19]	87.3	94.4	89.2	97.6	33.4	68.1
Patch-GCN[6]	88.8	95.0	89.7	98.1	38.2	71.2
GTP[32]	90.5	95.8	91.4	97.7	40.8	72.9
IGT (Ours)	**91.6**	**96.7**	**92.4**	**98.4**	**43.4**	**74.5**

3.3 Results

Performance comparison with the SOTA methods: We compare the proposed IGT with ten baselines: Seven of them are none graph based methods, including max/mean-pooing, ABMIL [13], DSMIL [17], CLAM-SB [22], CLAM-MB [22] and TransMIL [26]. Three of them are graph-based MIL methods, including GCN-ABMIL [19], PatchGCN [6] and GTP [32]. Note both TransMIL and GTP use Transformers. We chose overall accuracy (ACC) and area under receiver operating characteristic curve (AUROC) as the evaluation metrics.

As illustrated in Table 1, our IGT framework surpasses current SOTA methods. To be specific, compared with the best performing graph-based method, GTP, our method achieves a 1.1% improvement in accuracy and a 0.9% increase in AUROC for the binary classification on the TCGA-NSCLC dataset. In multi-class classification, our method shows a 1.0% improvement in accuracy and a 0.7% increase in AUROC for the TCGA-RCC dataset, and a 2.6% improvement in accuracy with a 2.5% increase in AUROC on the BRIGHT dataset. Similarly, when compared with the leading non-graph-based method, TransMIL, our method shows a substantial 2.2%-3.8% improvement in accuracy and a 0.7%-2.7% enhancement in AUROC. In conclusion, our graph-transformer-based method significantly outperforms current both graph-based and transformer-based approaches, indicating the advantages of integrating local neighborhood information with global context for enhanced performance.

Ablation Studies To demonstrate the efficacy of the developed GTI block and investigate the necessity of model components, we conduct a ablation study to quantify the separate benefit of the individual global-attention and GCN module using ABMIL and DSMIL as the aggregation modules. As shown in Table 2, compared with GTI block without self-attention, our GTI achieves a 2.4% to 5.7% improvement in accuracy. It proves that the self-attention mechanism in our GTI captures pairwise correlation across all instances and thus improves the performance. In comparing our GTI block with GTI block without the GCN branch, our GTI block achieves a 3.7% to 5.6% increase in accuracy. It shows the necessity of spatial information for WSI analysis.

An interesting finding is that the method equipped exclusively with the global attention module exhibit inferior performance compared to those only utilizing the GCN. This discrepancy can be attributed to the lack of spatial information when directly applying the self-attention mechanism for WSI analysis.

Table 2. An ablation study conducted to evaluate the importance of each component within the GTI block, utilizing ABMIL and DSMIL as the base aggregation models.

Aggregator	Backbone	TCGA-NSCLC		TCGA-RCC		BRIGHT-6class	
		ACC(%)	AUC(%)	ACC(%)	AUC(%)	ACC(%)	AUC(%)
ABMIL	-	84.1	91.3	86.8	96.1	30.8	67.0
	GTI w/o Attn	89.2	95.2	88.8	98.1	38.7	72.5
	GTI w/o GCN	86.0	93.1	87.8	98.0	38.1	71.2
	GTI	**91.6**	**96.7**	**92.4**	**98.4**	**43.4**	**74.5**
DSMIL	-	86.0	93.9	87.7	97.7	36.4	72.5
	GTI w/o Attn	87.9	94.3	89.8	98.1	39.0	73.7
	GTI w/o GCN	87.4	95.2	88.7	98.0	37.4	**73.8**
	GTI	**91.1**	**95.5**	**91.7**	**98.5**	**42.9**	73.4

4 Conclusion

In this paper, we introduce a new integrative graph-transformer framework, IGT, that simultaneously captures the context-aware relational features from local tissue regions and global WSI representations across instance embeddings for histopathology WSI classification. We integrate the graph convolutional network with a global attention module to construct the Graph-Transformer Integration block. Specifically, the graph convolutional network explores the local neighbor interactions and the multi-head self-attention model captures the long-range dependencies from all instances. The efficacy of the developed framework is manifested with three public WSI datasets. When compared with multiple state-of-the-art methods, our method consistently presents a superior performance, suggesting its promising potential to support computational histopathology analyses.

References

1. Ahmedt-Aristizabal, D., Armin, M.A., Denman, S., Fookes, C., Petersson, L.: A survey on graph-based deep learning for computational histopathology. Computerized Medical Imaging and Graphics **95**, 102027 (2022)
2. Alon, U., Yahav, E.: On the bottleneck of graph neural networks and its practical implications. ICLR (2021)
3. Bianchi, F.M., Grattarola, D., Alippi, C.: Spectral clustering with graph neural networks for graph pooling. In: International conference on machine learning. pp. 874–883. PMLR (2020)
4. Brancati, N., Anniciello, A.M., Pati, P., Riccio, D., Scognamiglio, G., Jaume, G., De Pietro, G., Di Bonito, M., Foncubierta, A., Botti, G., et al.: Bracs: A dataset for breast carcinoma subtyping in h&e histology images. Database **2022**, baac093 (2022)
5. Chen, R.J., Chen, C., Li, Y., Chen, T.Y., Trister, A.D., Krishnan, R.G., Mahmood, F.: Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In: CVPR (2022)
6. Chen, R.J., Lu, M.Y., Shaban, M., Chen, C., Chen, T.Y., Williamson, D.F., Mahmood, F.: Whole slide images are 2d point clouds: Context-aware survival prediction using patch-based graph convolutional networks. In: MICCAI (2021)
7. Cover, T., Hart, P.: Nearest neighbor pattern classification. IEEE transactions on information theory **13**(1), 21–27 (1967)
8. Dao, T., Fu, D., Ermon, S., Rudra, A., Ré, C.: Flashattention: Fast and memory-efficient exact attention with io-awareness. NIPS (2022)
9. Ding, S., Li, J., Wang, J., Ying, S., Shi, J.: Multi-scale efficient graph-transformer for whole slide image classification. arXiv preprint arXiv:2305.15773 (2023)
10. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
11. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016)
12. Hou, L., Samaras, D., Kurc, T.M., Gao, Y., Davis, J.E., Saltz, J.H.: Patch-based convolutional neural network for whole slide tissue image classification. In: CVPR (2016)
13. Ilse, M., Tomczak, J., Welling, M.: Attention-based deep multiple instance learning. In: International conference on machine learning. pp. 2127–2136. PMLR (2018)
14. Khened, M., Kori, A., Rajkumar, H., Krishnamurthi, G., Srinivasan, B.: A generalized deep learning framework for whole-slide image segmentation and analysis. Scientific reports **11**(1), 11579 (2021)
15. Kim, D.W., Lee, S., Kwon, S., Nam, W., Cha, I.H., Kim, H.J.: Deep learning-based survival prediction of oral cancer patients. Scientific reports **9**(1), 6994 (2019)
16. Kreuzer, D., Beaini, D., Hamilton, W., Létourneau, V., Tossou, P.: Rethinking graph transformers with spectral attention. Advances in Neural Information Processing Systems **34**, 21618–21629 (2021)
17. Li, B., Li, Y., Eliceiri, K.W.: Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In: CVPR (2021)
18. Li, G., Xiong, C., Thabet, A., Ghanem, B.: Deepergcn: All you need to train deeper gcns. arXiv preprint arXiv:2006.07739 (2020)

19. Liang, M., Chen, Q., Li, B., Wang, L., Wang, Y., Zhang, Y., Wang, R., Jiang, X., Zhang, C.: Interpretable classification of pathology whole-slide images using attention based context-aware graph convolutional neural network. Computer Methods and Programs in Biomedicine **229**, 107268 (2023)
20. Liu, L., Jiang, H., He, P., Chen, W., Liu, X., Gao, J., Han, J.: On the variance of the adaptive learning rate and beyond. In: ICLR (2019)
21. Lu, M.Y., Williamson, D.F., Chen, T.Y., Chen, R.J., Barbieri, M., Mahmood, F.: Data-efficient and weakly supervised computational pathology on whole-slide images. Nature biomedical engineering **5**(6), 555–570 (2021)
22. Lu, M.Y., Williamson, D.F., Chen, T.Y., Chen, R.J., Barbieri, M., Mahmood, F.: Data-efficient and weakly supervised computational pathology on whole-slide images. Nature biomedical engineering **5**(6), 555–570 (2021)
23. Nakhli, R., Moghadam, P.A., Mi, H., Farahani, H., Baras, A., Gilks, B., Bashashati, A.: Sparse multi-modal graph transformer with shared-context processing for representation learning of giga-pixel images. In: CVPR (2023)
24. Niazi, M.K.K., Parwani, A.V., Gurcan, M.N.: Digital pathology and artificial intelligence. The lancet oncology **20**(5), e253–e261 (2019)
25. Rampášek, L., Galkin, M., Dwivedi, V.P., Luu, A.T., Wolf, G., Beaini, D.: Recipe for a general, powerful, scalable graph transformer. NIPS (2022)
26. Shao, Z., Bian, H., Chen, Y., Wang, Y., Zhang, J., Ji, X., et al.: Transmil: Transformer based correlated multiple instance learning for whole slide image classification. NIPS (2021)
27. Wang, X., Yan, Y., Tang, P., Bai, X., Liu, W.: Revisiting multiple instance neural networks. Pattern Recognition **74**, 15–24 (2018)
28. Xiong, C., Chen, H., Sung, J., King, I.: Diagnose like a pathologist: Transformer-enabled hierarchical attention-guided multiple instance learning for whole slide image classification. arXiv preprint arXiv:2301.08125 (2023)
29. Xiong, Y., Zeng, Z., Chakraborty, R., Tan, M., Fung, G., Li, Y., Singh, V.: Nyströmformer: A nyström-based algorithm for approximating self-attention. In: AAAI (2021)
30. Yao, J., Zhu, X., Jonnagaddala, J., Hawkins, N., Huang, J.: Whole slide images based cancer survival prediction using attention guided deep multiple instance learning networks. Medical Image Analysis **65**, 101789 (2020)
31. Zhao, W., Wang, S., Yeung, M., Niu, T., Yu, L.: Mulgt: Multi-task graph-transformer with task-aware knowledge injection and domain knowledge-driven pooling for whole slide image analysis. AAAI (2023)
32. Zheng, Y., Gindra, R.H., Green, E.J., Burks, E.J., Betke, M., Beane, J.E., Kolachalama, V.B.: A graph-transformer for whole slide image classification. IEEE transactions on medical imaging **41**(11), 3003–3015 (2022)