HIGT: Hierarchical Interaction Graph-Transformer for Whole Slide Image Analysis

Ziyu Guo[1], Weiqin Zhao[1], Shujun Wang[2], and Lequan Yu[1(✉)]

[1] The University of Hong Kong, Hong Kong SAR, China
{gzypro,wqzhao98}@connect.hku.hk, lqyu@hku.hk
[2] The Hong Kong Polytechnic University, Hong Kong SAR, China
shu-jun.wang@polyu.edu.hk

Abstract. In computation pathology, the pyramid structure of gigapixel Whole Slide Images (WSIs) has recently been studied for capturing various information from individual cell interactions to tissue microenvironments. This hierarchical structure is believed to be beneficial for cancer diagnosis and prognosis tasks. However, most previous hierarchical WSI analysis works (1) only characterize local or global correlations within the WSI pyramids and (2) use only unidirectional interaction between different resolutions, leading to an incomplete picture of WSI pyramids. To this end, this paper presents a novel Hierarchical Interaction Graph-Transformer (*i.e.*, HIGT) for WSI analysis. With Graph Neural Network and Transformer as the building commons, HIGT can learn both short-range local information and long-range global representation of the WSI pyramids. Considering that the information from different resolutions is complementary and can benefit each other during the learning process, we further design a novel Bidirectional Interaction block to establish communication between different levels within the WSI pyramids. Finally, we aggregate both coarse-grained and fine-grained features learned from different levels together for slide-level prediction. We evaluate our methods on two public WSI datasets from TCGA projects, *i.e.*, kidney carcinoma (KICA) and esophageal carcinoma (ESCA). Experimental results show that our HIGT outperforms both hierarchical and non-hierarchical state-of-the-art methods on both tumor subtyping and staging tasks.

Keywords: WSI analysis · Hierarchical representation · Interaction · Graph neural network · Vision transformer

Z. Guo and W. Zhao: Contributed equally to this work.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-43987-2_73.

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14225, pp. 755–764, 2023.
https://doi.org/10.1007/978-3-031-43987-2_73

1 Introduction

Histopathology is considered the gold standard for diagnosing and treating many cancers [19]. The tissue slices are usually scanned into Whole Slide Images (WSIs) and serve as important references for pathologists. Unlike natural images, WSIs typically contain billions of pixels and also have a pyramid structure, as shown in Fig. 1. Such gigapixel resolution and expensive pixel-wise annotation efforts pose unique challenges to constructing effective and accurate models for WSI analysis. To overcome these challenges, Multiple Instance Learning (MIL) has become a popular paradigm for WSI analysis. Typically, MIL-based WSI analysis methods have three steps: (1) crop the huge WSI into numerous image patches; (2) extract instance features from the cropped patches; and (3) aggregate instance features together to obtain slide-level prediction results. Many advanced MIL models emerged in the past few years. For instance, ABMIL [9] and DeepAttnMIL [18] incorporated attention mechanisms into the aggregation step and achieved promising results. Recently, Graph-Transformer architecture [17] has been proposed to learn short-range local features through GNN and long-range global features through Transformer simultaneously. Such Graph-Transformer architecture has also been introduced into WSI analysis [15,20] to mine the thorough global and local correlations between different image patches. However, current Graph-Transformer-based WSI analysis models only consider the representation learning under one specific magnification, thus ignoring the rich multi-resolution information from the WSI pyramids.

Different resolution levels in the WSI pyramids contain different and complementary information [3]. The images at a high-resolution level contain cellular-level information, such as the nucleus and chromatin morphology features [10]. At a low-resolution level, tissue-related information like the extent of tumor-immune localization can be found [1], while the whole WSI describes the entire tissue microenvironment, such as intra-tumoral heterogeneity and tumor invasion [3]. Therefore, analyzing from only a single resolution would lead to an incomplete picture of WSIs. Some very recent works proposed to characterize and analyze WSIs in a pyramidal structure. H2-MIL [7] formulated WSI as a hierarchical heterogeneous graph and HIPT [3] proposed an inheritable ViT framework to model WSI at different resolutions. Whereas these methods only characterize local or global correlations within the WSI pyramids and use only unidirectional interaction between different resolutions, leading to insufficient capability to model the rich multi-resolution information of the WSI pyramids.

In this paper, we present a novel Hierarchical Interaction Graph-Transformer framework (*i.e.*, HIGT) to simultaneously capture both local and global information from WSI pyramids with a novel Bidirectional Interaction module. Specifically, we abstract the multi-resolution WSI pyramid as a heterogeneous hierarchical graph and devise a Hierarchical Interaction Graph-Transformer architecture to learn both short-range and long-range correlations among different image patches within different resolutions. Considering that the information from different resolutions is complementary and can benefit each other, we specially design a Bidirectional Interaction block in our Hierarchical Interaction ViT mod-

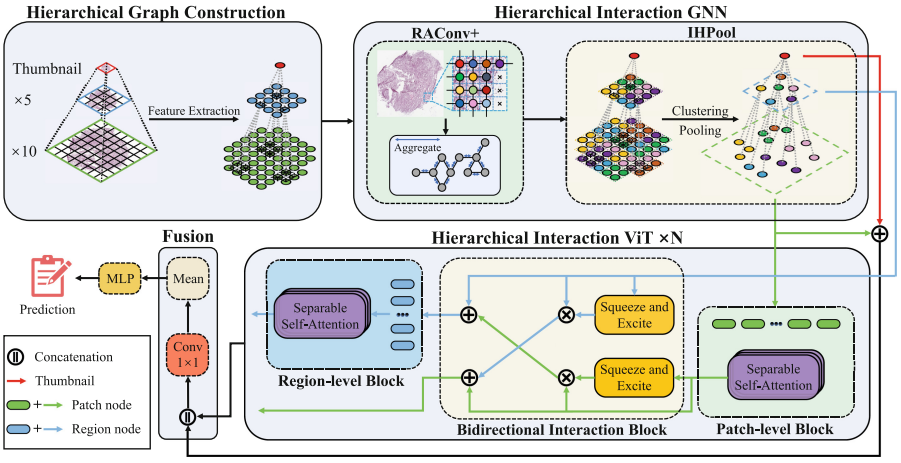

Fig. 1. Overview of the proposed HIGT framework. A WSI pyramid will be constructed as a hierarchical graph. Our proposed Hierarchical Interaction GNN and Hierarchical Interaction ViT block can capture the local and global features, and the Bidirectional Interaction module in the latter allows the nodes from different levels to interact. And finally, the Fusion block aggregates the coarse-grained and fine-grained features to generate the slide-level prediction.

ule to establish communication between different resolution levels. Moreover, a Fusion block is proposed to aggregate features learned from the different levels for slide-level prediction. To reduce the tremendous computation and memory cost, we further adopt the efficient pooling operation after the hierarchical GNN part to reduce the number of tokens and introduce the Separable Self-Attention Mechanism in Hierarchical Interaction ViT modules to reduce the computation burden. The extensive experiments with promising results on two public WSI datasets from TCGA projects, *i.e.*, kidney carcinoma (KICA) and esophageal carcinoma (ESCA), validate the effectiveness and efficiency of our framework on both tumor subtyping and staging tasks. The codes are available at https://github.com/HKU-MedAI/HIGT.

2 Methodology

Figure 1 depicts the pipeline of HIGT framework for better exploring the multi-scale information in hierarchical WSI pyramids. First, we abstract each WSI as a hierarchical graph, where the feature embeddings extracted from multi-resolution patches serve as nodes and the edge denotes the spatial and scaling relationships of patches within and across different resolution levels. Then, we feed the constructed graph into several hierarchical graph convolution blocks to learn the short-range relationship among graph nodes, following pooling operations to aggregate local context and reduce the number of nodes. We further devise a Separable Self-Attention-based Hierarchical Interaction Transformer

architecture equipped with a novel Bidirectional Interaction block to learn the long-range relationship among graph nodes. Finally, we design a fusion block to aggregate the features learned from the different levels of WSI pyramids for final slide-level prediction.

2.1 Graph Construction

As shown in Fig. 1, a WSI is cropped into numerous non-overlapping 512 × 512 image patches under different magnifications (*i.e.*, ×5, ×10) by using a sliding window strategy, where the OTSU algorithm [4] is used to filter out the background patches. Afterwards, we employ a pre-trained KimiaNet [16] to extract the feature embedding of each image patch. The feature embeddings of the slide-level $\boldsymbol{T}$ (Thumbnail), region-level $\boldsymbol{R}$ (×5), and the patch-level $\boldsymbol{P}$ (×10) can be represented as,

$$\begin{aligned} &\boldsymbol{T} = \{\boldsymbol{t}\}, \\ &\boldsymbol{R} = \{\boldsymbol{r}_1, \boldsymbol{r}_2, \cdots, \boldsymbol{r}_N\}, \\ &\boldsymbol{P} = \{\boldsymbol{P}_1, \boldsymbol{P}_2, \cdots, \boldsymbol{P}_N\}, \boldsymbol{P_i} = \{\boldsymbol{p}_{i,1}, \boldsymbol{p}_{i,2}, \cdots, \boldsymbol{p}_{i,M}\}, \end{aligned} \tag{1}$$

where $\boldsymbol{t}, \boldsymbol{r}_i, \boldsymbol{p}_{i,j} \in \mathbb{R}^{1\times C}$ correspond to the feature embeddings of each patch in thumbnail, region, and patch levels, respectively. N is the total number of the region nodes and M is the number of patch nodes belonging to a certain region node, and C denotes the dimension of feature embedding (1,024 in our experiments). Based on the extracted feature embeddings, we construct a hierarchical graph to characterize the WSI, following previous H^2-MIL work [7]. Specifically, the cropped patches serve as the nodes of the graph and we employ the extracted feature embedding as the node embeddings. There are two kinds of edges in the graph: spatial edges to denote the 8-adjacent spatial relationships among different patches in the same levels, and scaling edges to denote the relationship between patches across different levels at the same location.

2.2 Hierarchical Graph Neural Network

To learn the short-range relationship among different patches within the WSI pyramid, we propose a new hierarchical graph message propagation operation, called RAConv+. Specifically, for any source node j in the hierarchical graph, we define the set of it all neighboring nodes at resolution k as $\mathcal{N}_k$ and $k \in K$. Here K means all resolutions. And the h_k is the mean embedding of the node j's neighboring nodes in resolution k. And $h_{j\prime}$ is the embedding of the neighboring nodes of node j in resolution k and $h_{j\prime} \in \mathcal{N}_k$. The formula for calculating the attention score of node j in resolution-level and node-level:

$$\alpha_k = \frac{\exp\left(\boldsymbol{a}^\top \cdot \text{LeakyReLU}\left([\boldsymbol{U}\boldsymbol{h}_j \| \boldsymbol{U}\boldsymbol{h}_k]\right)\right)}{\sum_{k'\in\mathcal{K}} \exp\left(\boldsymbol{a}^\top \cdot \text{LeakyReLU}\left([\boldsymbol{U}\boldsymbol{h}_j \| \boldsymbol{U}\boldsymbol{h}_{k'}]\right)\right)},$$

$$\alpha_{j\prime} = \frac{\exp\left(\boldsymbol{b}^\top \cdot \text{LeakyReLU}\left([\boldsymbol{V}\boldsymbol{h}_j \| \boldsymbol{V}\boldsymbol{h}_{j\prime\prime}]\right)\right)}{\sum_{h_{j\prime\prime}\in\mathcal{N}_k} \exp\left(\boldsymbol{b}^\top \cdot \text{LeakyReLU}\left([\boldsymbol{V}\boldsymbol{h}_j \| \boldsymbol{V}\boldsymbol{h}_{j\prime\prime}]\right)\right)},$$

$$\alpha_{j,j\prime} = \alpha_k + \alpha_{j\prime}, \tag{2}$$

where $\alpha_{j,j\prime}$ is the attention score of the node j to node $j\prime$ and h_j is the source node j embedding. And U, V, a and b are four learnable layers. The main difference from H2-MIL [6] is that we pose the non-linear $LeakyReLU$ between a and U, b and V, to generate a more distinct attention score matrix which increases the feature differences between different types of nodes [2]. Therefore, the layer-wise graph message propagation can be represented as:

$$H^{(l+1)} = \sigma\left(\mathcal{A} \cdot H^{(l)} \cdot W^{(l)}\right), \tag{3}$$

where $\mathcal{A}$ represents the attention score matrix, and the attention score for the j-th row and j$\prime$-th column of the matrix is given by Eq. (2). At the end of the hierarchical GNN part, we use the IHPool [6] progressively aggregate the hierarchical graph.

2.3 Hierarchical Interaction ViT

We further propose a Hierarchical Interaction ViT (HIViT) to learn long-range correlation within the WSI pyramids, which includes three key components: Patch-level (PL) blocks, Bidirectional Interaction (BI) blocks, and Region-level (RL) blocks.

Patch-Level Block. Given the patch-level feature set $\boldsymbol{P} = \bigcup_{i=1}^{N} \boldsymbol{P}_i$, the PL block learns long-term relationships within the patch level:

$$\hat{\boldsymbol{P}}^{l+1} = PL(\boldsymbol{P}^l) \tag{4}$$

where $l = 1, 2, ..., L$ is the index of the HIViT block. $PL(\cdot)$ includes a Separable Self Attention (SSA) [13], 1×1 Convolution, and Layer Normalization in sequence. Note that here we introduced SSA into the PL block to reduce the computation complexity of attention calculation from quadratic to linear while maintaining the performance [13].

Bidirectional Interaction Block. We propose a Bidirectional Interaction (BI) block to establish communication between different levels within the WSI pyramids. The BI block performs bidirectional interaction, and the interaction progress from region nodes to patch nodes is:

$$\begin{aligned} &\boldsymbol{r}_{\boldsymbol{i}}^{\boldsymbol{l'}} \in \boldsymbol{R}^{\boldsymbol{l'}}, \quad \boldsymbol{R}^{\boldsymbol{l'}} = SE(\boldsymbol{R}^{\boldsymbol{l}}) \cdot \boldsymbol{R}^{\boldsymbol{l}}, \\ &\boldsymbol{P}_i^{l+1} = \{\boldsymbol{p}_{i,1}^{l+1}, \boldsymbol{p}_{i,2}^{l+1}, \cdots, \boldsymbol{p}_{i,k}^{l+1}\}, \quad \boldsymbol{p}_{i,k}^{l+1} = \hat{\boldsymbol{p}}_{i,k}^{l+1} + \boldsymbol{r}_i^{l'}, \end{aligned} \tag{5}$$

where the $SE(\cdot)$ means the Sequeeze-and-Excite layer [8] and the $\boldsymbol{r}_{\boldsymbol{i}}^{\boldsymbol{l'}}$ means the i-th region node in $\boldsymbol{R}^{\boldsymbol{l'}}$, and $\hat{\boldsymbol{p}}_{i,k}^{l+1}$ is the k-th patch node linked to the i-th region

node after the interaction. Besides, another direction of the interaction is,

$$\bar{\boldsymbol{P}} = \{\bar{\boldsymbol{P}}_1^{l+1}, \bar{\boldsymbol{P}}_2^{l+1}, \cdots, \bar{\boldsymbol{P}}_n^{l+1}\}, \quad \bar{\boldsymbol{P}}_i^{l+1} = MEAN(\hat{\boldsymbol{P}}_i^{l+1})$$
$$\hat{\boldsymbol{R}}^{l+1} = SE(\bar{\boldsymbol{P}}^{l+1}) \cdot \bar{\boldsymbol{P}}^{l+1} + \boldsymbol{R}^{l}, \tag{6}$$

where the $MEAN(\cdot)$ is the operation to get the mean value of patch nodes set $\hat{\boldsymbol{P}}_i^{l+1}$ associated with the i-th region node and $\bar{\boldsymbol{P}}_1^{l+1} \in \mathcal{R}^{1\times C}$ and the C is the feature channel of nodes, and $\hat{\boldsymbol{R}}^{l+1}$ is the region nodes set after interaction.

Region-Level Block. The final part of this module is to learn the long-range correlations of the interacted region-level nodes:

$$\boldsymbol{R}^{l+1} = RL(\hat{\boldsymbol{R}}^{l+1}) \tag{7}$$

where $l = 1, 2, ..., L$ is the index of the HIViT module, $\boldsymbol{R} = \{\boldsymbol{r}_1, \boldsymbol{r}_2, \cdots, \boldsymbol{r}_N\}$, and $RL(\cdot)$ has a similar structure to $PL(\cdot)$.

2.4 Slide-Level Prediction

In the final stage of our framework, we design a Fusion block to combine the coarse-grained and fine-grained features learned from the WSI pyramids. Specifically, we use an element-wise summation operation to fuse the coarse-grained thumbnail feature and patch-level features from the Hierarchical Interaction GNN part, and then further fuse the fine-grained patch-level features from the HIViT part with a concatenation operation. Finally, a 1×1 convolution and mean operation followed by a linear projection are employed to produce the slide-level prediction.

3 Experiments

Datasets and Evaluation Metrics. We assess the efficacy of the proposed HIGT framework by testing it on two publicly available datasets (KICA and ESCA) from The Cancer Genome Atlas (TCGA) repository. The datasets are described below in more detail:

- **KICA dataset.** The KICA dataset consists of 371 cases of kidney carcinoma, of which 279 are classified as early-stage and 92 as late-stage. For the tumor typing task, 259 cases are diagnosed as kidney renal papillary cell carcinoma, while 112 cases are diagnosed as kidney chromophobe.
- **ESCA dataset.** The ESCA dataset comprises 161 cases of esophageal carcinoma, with 96 cases classified as early-stage and 65 as late-stage. For the tumor typing task, there are 67 squamous cell carcinoma cases and 94 adenocarcinoma cases.

Experimental Setup. The proposed framework was implemented by PyTorch [14] and PyTorch Geometric [5]. All experiments were conducted on a workstation with eight NVIDIA GeForce RTX 3090 (24 GB) GPUs. The shape of all nodes' features extracted by KimiaNet is set to 1×1024. All methods are trained with a batch size of 8 for 50 epochs. The learning rate was set as 0.0005, with Adam optimizer. The accuracy (ACC) and area under the curve (AUC) are used as the evaluation metric. All approaches were evaluated with five-fold cross-validations (5-fold CVs) from five different initializations.

Table 1. Comparison with other methods on ESCA. Top results are shown in bold.

Method	Staging		Typing	
	AUC	ACC	AUC	ACC
ABMIL [9]	64.53 ± 4.80	64.39 ± 5.05	94.11 ± 2.69	93.07 ± 2.68
CLAM-SB [12]	67.45 ± 5.40	67.29 ± 5.18	93.79 ± 5.52	93.47 ± 5.77
DeepAttnMIL [18]	67.96 ± 5.52	67.53 ± 4.96	95.68 ± 1.94	94.43 ± 3.04
DS-MIL [11]	66.92 ± 5.28	66.83 ± 5.57	95.96 ± 3.07	94.77 ± 4.10
LA-MIL [15]	63.93 ± 6.19	63.45 ± 6.19	95.23 ± 3.75	94.69 ± 3.94
H2-MIL [7]	63.20 ± 8.36	62.72 ± 8.32	91.88 ± 4.17	91.31 ± 4.18
HIPT [3]	68.59 ± 5.62	68.45 ± 6.39	94.62 ± 2.34	93.01 ± 3.28
Ours	$\mathbf{71.11 \pm 6.04}$	$\mathbf{70.53 \pm 5.41}$	$\mathbf{96.81 \pm 2.49}$	$\mathbf{96.16 \pm 2.85}$

Table 2. Comparison with other methods on KICA. Top results are shown in bold.

Method	Staging		Typing	
	AUC	ACC	AUC	ACC
ABMIL [9]	77.40 ± 3.87	75.94 ± 5.06	97.76 ± 1.74	98.86 ± 0.69
CLAM-SB [12]	77.16 ± 3.64	76.61 ± 4.31	96.76 ± 3.42	97.13 ± 2.99
DeepAttnMIL [18]	76.77 ± 1.94	75.94 ± 2.41	97.44 ± 1.04	96.30 ± 2.63
DS-MIL [11]	77.33 ± 4.11	76.57 ± 5.14	98.03 ± 1.13	97.31 ± 1.85
LA-MIL [15]	69.37 ± 5.27	68.73 ± 5.09	98.34 ± 0.98	97.71 ± 1.76
H2-MIL [7]	65.59 ± 6.65	64.48 ± 6.20	98.06 ± 1.43	96.99 ± 3.01
HIPT [3]	75.93 ± 2.01	75.34 ± 2.31	98.71 ± 0.49	97.32 ± 2.24
Ours	$\mathbf{78.80 \pm 2.10}$	$\mathbf{76.80 \pm 2.30}$	$\mathbf{98.90 \pm 0.60}$	$\mathbf{97.90 \pm 1.40}$

Comparison with State-of-the-Art Methods. We first compared our proposed HIGT framework with two groups of state-of-the-art WSI analysis methods: (1) non-hierarchical methods including: ABMIL [9], CLAM-SB [12], DeepAttnMIL [18], DS-MIL [11], LA-MIL [15], and (2) hierarchical methods including: H2-MIL [7], HIPT [3]. For LA-MIL [15] method, it was introduced with

a single-scale Graph-Transformer architecture. For H2-MIL [7] and HIPT [3], they were introduced with a hierarchical Graph Neural Network and hierarchical Transformer architecture, respectively. The results for ESCA and KICA datasets are summarized in Table 1 and Table 2, respectively. Overall, our model achieves a content result both in AUC and ACC of classifying the WSI, and especially in predicting the more complex task (i.e. Staging) compared with the SOTA approaches. Even for the non-hierarchical Graph-Transformer baseline LA-MIL and hierarchical transformer model HIPT, our model approaches at least around 3% and 2% improvement on AUC and ACC in the classification of the Staging of the KICA dataset. Therefore we believe that our model benefits a lot from its used modules and mechanisms.

Ablation Analysis. We further conduct an ablation study to demonstrate the effectiveness of the proposed components. The results are shown in Table 3. In its first row, we replace the RAConv+ with the original version of this operation. And in the second row, we replace the Separable Self Attention with a canonical transformer block. The third row changes the bidirectional interaction mechanism into just one direction from region-level to patch-level. And the last row, we remove the fusion block from our model. Finally, the ablation analysis results show that all of these modules we used actually improved the prediction effect of the model to a certain extent.

Table 3. Ablation analysis on KICA dataset.

Method	Staging		Typing	
	AUC	ACC	AUC	ACC
H2-MIL + HIViT	77.35 ± 3.41	$\mathbf{77.16 \pm 3.29}$	98.56 ± 1.01	95.00 ± 1.75
Ours w/o SSA	73.45 ± 8.48	71.47 ± 3.21	97.94 ± 2.51	97.42 ± 2.65
Ours w/o BI	72.42 ± 2.09	71.34 ± 7.23	98.04 ± 8.30	96.54 ± 2.80
Ours w/o Fusion	77.87 ± 2.09	76.80 ± 2.95	98.46 ± 0.88	97.35 ± 1.81
Ours	$\mathbf{78.80 \pm 2.10}$	76.80 ± 2.30	$\mathbf{98.90 \pm 0.60}$	$\mathbf{97.90 \pm 1.40}$

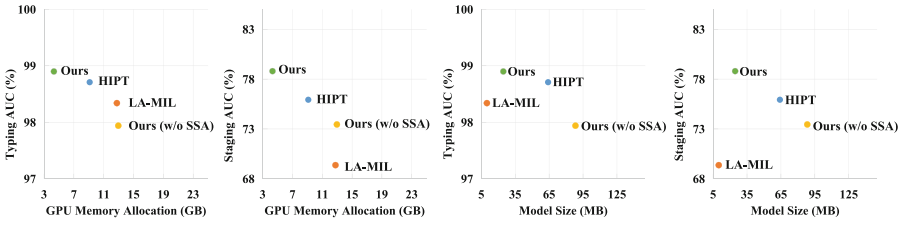

Fig. 2. Computational analysis of our framework and some selected SOTA methods. From left to right are scatter plots of Typing AUC v.s. GPU Memory Allocation, Staging AUC v.s. GPU Memory Allocation, Typing AUC v.s. Model Size, Staging AUC v.s. Model Size.

Computation Cost Analysis. We analyze the computation cost during the experiments to compare the efficiency between our methods and existing state-of-the-art approaches. Besides we visualized the model size (MB) and the training memory allocation of GPU (GB) v.s. performance in KICA's typing and staging task plots in Fig. 2. All results demonstrate that our model is able to maintain the promising prediction result while reducing the computational cost and model size effectively.

4 Conclusion

In this paper, we propose HIGT, a framework that simultaneously and effectively captures local and global information from the hierarchical WSI. Firstly, the constructed hierarchical data structure of the multi-resolution WSI is able to offer multi-scale information to the later model. Moreover, the redesigned H2-MIL and HIViT capture the short-range and long-range correlations among varying magnifications of WSI separately. And the bidirectional interaction mechanism and fusion block can facilitate communication between different levels in the Transformer part. We use IHPool and apply the Separable Self Attention to deal with the inherently high computational cost of the Graph-Transformer model. Extensive experimentation on two public WSI datasets demonstrates the effectiveness and efficiency of our designed framework, yielding promising results. In the future, we will evaluate on other complex tasks such as survival prediction and investigate other techniques to improve the efficiency of our framework.

Acknowledgement. The work described in this paper was partially supported by grants from the National Natural Science Fund (62201483), the Research Grants Council of the Hong Kong Special Administrative Region, China (T45-401/22-N), and The Hong Kong Polytechnic University (P0045999).

References

1. AbdulJabbar, K., et al.: Geospatial immune variability illuminates differential evolution of lung adenocarcinoma. Nature Med. **26**(7), 1054–1062 (2020)
2. Brody, S., Alon, U., Yahav, E.: How attentive are graph attention networks? arXiv preprint arXiv:2105.14491 (2021)
3. Chen, R.J., et al.: Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 16144–16155 (June 2022)
4. Chen, R.J., et al.: Whole slide images are 2D Point Clouds: context-aware survival prediction using patch-based graph convolutional networks. In: de Bruijne, M., et al. (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2021: 24th International Conference, Strasbourg, France, September 27 – October 1, 2021, Proceedings, Part VIII, pp. 339–349. Springer International Publishing, Cham (2021). https://doi.org/10.1007/978-3-030-87237-3_33
5. Fey, M., Lenssen, J.E.: Fast graph representation learning with pytorch geometric. arXiv preprint arXiv:1903.02428 (2019)

6. Hou, W., Wang, L., Cai, S., Lin, Z., Yu, R., Qin, J.: Early neoplasia identification in Barrett's esophagus via attentive hierarchical aggregation and self-distillation. Medical Image Anal. **72**, 102092 (2021). https://doi.org/10.1016/j.media.2021.102092
7. Hou, W., et al.: H^2-mil: Exploring hierarchical representation with heterogeneous multiple instance learning for whole slide image analysis. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, pp. 933–941 (2022)
8. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7132–7141 (2018)
9. Ilse, M., Tomczak, J., Welling, M.: Attention-based deep multiple instance learning. In: International Conference on Machine Learning, pp. 2127–2136. PMLR (2018)
10. Kumar, N., Verma, R., Sharma, S., Bhargava, S., Vahadane, A., Sethi, A.: A dataset and a technique for generalized nuclear segmentation for computational pathology. IEEE Trans. Med. Imaging **36**(7), 1550–1560 (2017)
11. Li, B., Li, Y., Eliceiri, K.W.: Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14318–14328 (2021)
12. 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 Biomed. Eng. **5**(6), 555–570 (2021)
13. Mehta, S., Rastegari, M.: Separable self-attention for mobile vision transformers. arXiv preprint arXiv:2206.02680 (2022)
14. Paszke, A., et al.: Pytorch: an imperative style, high-performance deep learning library. Adv. Neural Inform. Process. Syst. **32** (2019)
15. Reisenbüchler, D., Wagner, S.J., Boxberg, M., Peng, T.: Local attention graph-based transformer for multi-target genetic alteration prediction. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part II, pp. 377–386. Springer Nature Switzerland, Cham (2022). https://doi.org/10.1007/978-3-031-16434-7_37
16. Riasatian, A., et al.: Fine-tuning and training of densenet for histopathology image representation using TCGA diagnostic slides. Med. Image Anal. **70**, 102032 (2021)
17. Wu, Z., Jain, P., Wright, M., Mirhoseini, A., Gonzalez, J.E., Stoica, I.: Representing long-range context for graph neural networks with global attention. Adv. Neural. Inf. Process. Syst. **34**, 13266–13279 (2021)
18. 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. Med. Image Anal. **65**, 101789 (2020)
19. Yao, X.H., et al.: Pathological evidence for residual SARS-CoV-2 in pulmonary tissues of a ready-for-discharge patient. Cell Res. **30**(6), 541–543 (2020)
20. Zheng, Y., et al.: A graph-transformer for whole slide image classification. IEEE Trans. Med. Imaging **41**(11), 3003–3015 (2022)