Sparse Multi-Modal Graph Transformer with Shared-Context Processing for Representation Learning of Giga-pixel Images

Ramin Nakhli
University of British Columbia
ramin.nakhli@ubc.ca

Puria Azadi Moghadam
University of British Columbia
puria.azadi@ubc.ca

Haoyang Mi
Johns Hopkins University
hmi1@jhmi.edu

Hossein Farahani
University of British Columbia
h.farahani@ubc.ca

Alexander Baras
Johns Hopkins University
baras@jhmi.edu

Blake Gilks
University of British Columbia
blake.gilks@vch.ca

Ali Bashashati
University of British Columbia
ali.bashashati@ubc.ca

Abstract

Processing giga-pixel whole slide histopathology images (WSI) is a computationally expensive task. Multiple instance learning (MIL) has become the conventional approach to process WSIs, in which these images are split into smaller patches for further processing. However, MIL-based techniques ignore explicit information about the individual cells within a patch. In this paper, by defining the novel concept of shared-context processing, we designed a multi-modal Graph Transformer (AMIGO) that uses the cellular graph within the tissue to provide a single representation for a patient while taking advantage of the hierarchical structure of the tissue, enabling a dynamic focus between cell-level and tissue-level information. We benchmarked the performance of our model against multiple state-of-the-art methods in survival prediction and showed that ours can significantly outperform all of them including hierarchical Vision Transformer (ViT). More importantly, we show that our model is strongly robust to missing information to an extent that it can achieve the same performance with as low as 20% of the data. Finally, in two different cancer datasets, we demonstrated that our model was able to stratify the patients into low-risk and high-risk groups while other state-of-the-art methods failed to achieve this goal. We also publish a large dataset of immunohistochemistry images (InUIT) containing 1,600 tissue microarray (TMA) cores from 188 patients along with their survival information, making it one of the largest publicly available datasets in this context.

1. Introduction

Digital processing of medical images has recently attracted significant attention in computer vision communities, and the applications of deep learning models in this domain span across various image types (*e.g.*, histopathology images, CT scans, and MRI scans) and numerous tasks (*e.g.*, classification, segmentation, and survival prediction) [6, 11, 27, 28, 30, 36, 38, 44]. The paradigm-shifting ability of these models to learn predictive features directly from raw images has presented exciting opportunities in medical imaging. This has especially become more important for digitized histopathology images where each data point is a multi-gigapixel image (also referred to as a Whole Slide Image or WSI). Unlike natural images, each WSI has high granularity at different levels of magnification and a size reaching 100,000×100,000 pixels, posing exciting challenges in computer vision.

The typical approach to cope with the computational complexities of WSI processing is to use the Multiple Instance Learning (MIL) technique [31]. More specifically, this approach divides each slide into smaller patches (*e.g.*, 256×256 pixels), passes them through a feature extractor, and represents the slide with an aggregation of these representations. This technique has shown promising results in a variety of tasks, including cancer subtype classification and survival prediction. However, it suffers from several major issues. Firstly, considering the high resolution of WSIs, even a non-overlapping 256×256 window generates a huge number of patches. Therefore, the subsequent aggregation method of MIL has to perform either a simple pooling operation [3, 17] or a hierarchical aggregation to

add more flexibility [6]. Nevertheless, the former limits the representative power of the aggregator drastically, and the latter requires a significant amount of computational power. Secondly, this approach is strongly dependent on the size of the dataset, which causes the over-fitting of the model in scenarios where a few data points (*e.g.*, hundreds) are available. Lastly, despite the fact that cells are the main components of the tissue, the MIL approach primarily focuses on patches, which limits the resolution of the model to a snapshot of a population of cells rather than a single cell. Consequently, the final representation of the slide lacks the mutual interactions of individual cells.

Multiple clinical studies have strongly established that the heterogeneity of the tissue has a crucial impact on the outcome of cancer [32, 46]. For instance, high levels of immune infiltration in the tumor stroma were shown to correlate with longer survival and positive therapy response in breast cancer patients [46]. Therefore, machine learning methods for histopathology image analysis are required to account for tumor heterogeneity and cell-cell interactions. Nonetheless, the majority of the studies in this domain deal with a single image highlighting cell nuclei (regardless of cell type) and extra cellular matrix. Recently, few studies have investigated pathology images where various cell types were identified using different protein markers [25, 42]. However, they still utilized a single-modal approach (i.e., one cell type in an image), ignoring the multi-modal context (i.e., several cell types within the tissue) of these images.

In this study, we explore the application of graph neural networks (GNN) for the processing of cellular graphs (i.e., a graph constructed by connecting adjacent cells to each other) generated from histopathology images (Fig. 1). In particular, we are interested in the cellular graph because it gives us the opportunity to focus on cell-level information as well as their mutual interactions. By delivering an adaptable focus at different scales, from cell level to tissue level, such information allows the model to have a multi-scale view of the tissue, whereas MIL models concentrate on patches with a preset resolution and optical magnification. The availability of cell types and their spatial location helps the model to find regions of the tissue that have more importance for its representation (*e.g.*, tumor regions or immune cells infiltrating into tumor cells). In contrast to the expensive hierarchical pooling in MIL methods [6], the message-passing nature of GNNs offers an efficient approach to process the vast scale of WSIs as a result of weight sharing across all the graph nodes. This approach also reduces the need for a large number of WSIs during training as the number of parameters is reduced.

In this work, we introduce a spArse MultI-modal Graph transfOrmer model (AMIGO) for the representation learning of histopathology images by using cells as the main building blocks. Starting from the cell level, our model gradually propagates information to a larger neighborhood of cells, which inherently encodes the hierarchical structure of the tissues. More importantly, in contrast to other works, we approach this problem in a multi-modal manner, where we can get a broader understanding of the tissue structure, which is quite critical for context-based tasks such as survival prediction. In particular, for a single patient, there can be multiple histopathology images available, each highlighting cells of a certain type (by staining cells with specific protein markers), and resulting in a separate cellular graph (Fig. 2). Therefore, using a multi-modal approach, we combine the cellular graphs of different modalities together to obtain a unified representation for a single patient. This also affirms our stance regarding the importance of cell type and the distinction between different cellular connectivity types. Aside from achieving state-of-the-art results, we notice that, surprisingly, our multi-modal approach is strongly robust to missing information, and this enables us to perform more efficient training by relying on this reconstruction ability of the network. Our work advances the frontiers of MIL, Vision Transformer (ViT), and GNNs in multiple directions:

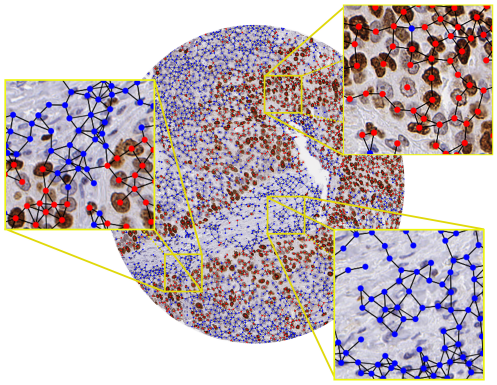


Figure 1. Cellular graph built from a $4,000 \times 4,000$ pixel TMA core stained with Ki67 biomarker. Each red point demonstrates a cell that has a positive response to Ki67 while the blue points show cells that had a negative response to this biomarker. The highlighted patches show representative areas of the tissue where the spatial distribution of cells and the structure of the tissue are different. A typical MIL method cannot capture this heterogeneity as it does not take into account the location of the patches and lacks explicit information about the specific cells present within a patch.

- We introduce the first multi-modal cellular graph processing model that performs survival prediction based on the multi-modal histopathology images with shared contextual information.

- Our model eliminates the critical barriers of MIL mod-

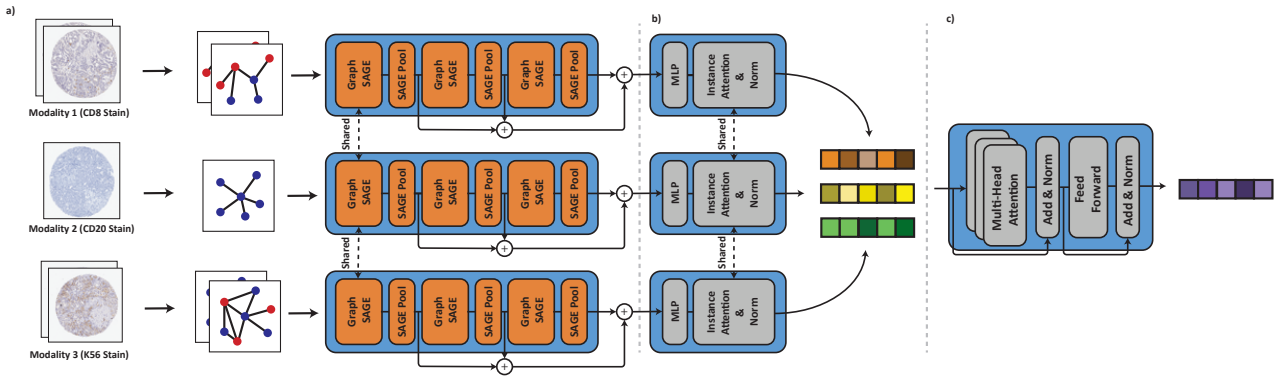


Figure 2. Overview of our proposed method. a) The Cellular graphs are first extracted from histopathology images stained with different biomarkers (*e.g.*, CD8, CD20, and Ki67) and are fed into the encoder corresponding to their modality. The initial layer of encoders is shared, allowing further generalization, while the following layers pick up functionalities unique to each modality. The graphs at the top depict the hierarchical pooling mechanism of the model. b) The representations obtained from multiple graph instances in each modality are combined via a shared instance attention layer (shared-context processing), providing a single representation vector. c) A Transformer is used to merge the resultant vectors to create a patient-level embedding that will be used for downstream tasks such as survival prediction.

els, enabling efficient training of multi-gigapixel images on a single GPU and outperforming all the baselines including ViT. It also implements the hierarchical structure of Vision Transformer while keeping the number of parameters significantly lower during end-to-end training.

- We also publish a large dataset of IHC images containing $1,600$ tissue microarray (TMA) cores from 188 patients along with their survival information, making it one of the largest datasets in this context.

2. Related Work

2.1. Multiple Instance Learning in Histopathology

Inspired by the bag-of-words idea, Zaheer *et al*. [47] and Brendel *et al*. [4] are two pioneers of MIL models that propose the permutation-invariant bag-of-features for image representation learning. Similarly, the early works of MIL in digital pathology follow the same approach to learning representations for WSIs by relying on simple algorithms for patch-level aggregation [16, 19]. However, the later works adopt more flexible designs for this purpose. For instance, Ilse *et al*. [17] use an attention-based operation to pool the representations across all patches, and Campanella *et al*. [5] aggregate the representation of the top-ranked patches using a Recurrent Neural Network (RNN). Li *et al*. [23] introduce the idea of multi-resolution MIL by proposing a two-stage model, clustering patches at a $5\times$ magnification and using an attention pooling on the most informative patches at the $10\times$ magnification. Likewise, another concurrent work implements the multi-resolution idea using self-supervised learning while the authors consider the spatial positioning of the patches as well [22]. In a more recent study, Zhang *et al*. [48] introduce pseudo-bagging in a double-tier setting, while Chen *et al*. [6] propose using Vision Transformers for hierarchical pooling of WSIs. Nevertheless, all the aforementioned studies ignore the cell-level details residing in the images and require a large amount of data (thousands of patients) for the training of the model. In this paper, we focus on resolving these shortcomings by using a cell-centric method while performing a hierarchical pooling of information across different sections of the image.

2.2. Graph Neural Networks in Histopathology

Graph neural networks have recently drawn significant attention as they have led to outstanding performance in various tasks, mainly due to their structure-preserving ability [33, 37]. Since this type of model works based on the foundation of local message passing, it is suitable for capturing spatial information in histopathology images [13]. Adnan *et al*. [1] selected the most important patches from the WSI, created a fully-connected graph from them, and processed it using a GNN to obtain a representation for the whole graph. On the other hand, Lu *et al*. [24] combine the adjacent similar patches into a node in the graph and then apply a GNN. In another application, Zheng *et al*. [49] use GNNs to provide a hashing mechanism for retrieving regions of interest that are contextually similar to the query image. Similar to our work, Chen *et al*. [8] and Wang *et al*. [42] use cellular graphs for survival prediction. However, unlike our proposed work, these studies ignore the type of the cell and use a single-modal setting to perform the prediction.

2.3. Multi-Modal Image Analysis in Histopathology

Using histopathology images along with omics data (*e.g.*, transcriptomics and mutation) is very well studied in histopathology. For instance, Vale-Silva *et al.* [40] use a combination of histopathology images, clinical information, and RNA data to perform survival prediction. Chen *et al.* [8] utilize the Kronecker product to fuse the processed histopathology image data with genomics information for survival prediction, and the same authors [9] add more analysis to their work to link the results to interpretable features in pancreatic cancer. Although multi-modal learning of histopathology images with genomic data is extensively studied, the applications of multi-modal learning on images with different stains is vastly ignored. To the best of our knowledge, Dwivedi *et al.*'s work [10] is the only available study that does so to fuse different staining images for grade prediction. However, unlike our proposed design, they approach this problem in an MIL design.

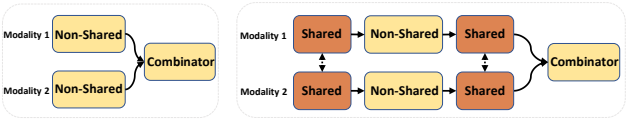
3. Method

3.1. Problem Formulation

In this part, we introduce the notations used in the remainder of the paper. Consider $\{x^m_{n,i} | n = 0, ..., N; m = 0, ..., M; i = 0, ..., C(n, m)\}$ to be the collection of images in a dataset, where n is the patient number, m is the modality number, and i is the image identifier. In this setting, N shows the total number of patients, M is the total number of modalities, and $C(n, m)$ is the number of images available for patient n from the modality of m. Our goal is to predict the estimated survival time of each patient, also called outcome. More specifically, we use all the available images for a patient (across different modalities) to obtain a unified representation of $R_n \in \mathbb{R}^{1\times d}$ based on which a survival time can be predicted. To avoid duplication, in the rest of this paper, we assume $x^m_{n,i}$ refers to both the image and the cellular graph generated from it. We will explain this preprocessing step in section Sec. 4.1.

3.2. Multi-Modal Shared-Context Processing

Before delving into the specifics of our model, we must first introduce a new notion that we refer to as shared-context processing. The common strategy for processing multi-modal data is to encode each modality using a separate encoder (Fig. 3a). However, we contend that combining shared and non-shared processing steps can be beneficial when dealing with different modalities containing comparable context (*e.g.*, cellular graphs from different stains). In particular, we believe a 3-step procedure (Fig. 3b) is necessary for such scenarios: 1) using a shared model to extract basic features from all of the modalities to help with the generalization; 2) performing modality-specific analysis using separate models for each modality; 3) unifying the high-level representations by a model shared across all modalities. This shared-context processing approach enables low-level and high-level feature unification while allowing flexibility for mid-level feature processing. In the next section, we will explain how our model benefits from this design.



(a) Non-shared-context (b) Shared-context

Figure 3. Shared-context vs non-shared-context multi-modal processing architectures. The shared-context benefits from generalization at low-level and high-level features while allowing modality-specific processing at mid-level.

3.3. Sparse Multi-modal Cellular Graph Neural Network

3.3.1 Modality Encoding

In general, there are two degrees of variability for each patient: 1) a variety of images from different modalities; 2) a variety of images within each modality. The first stage of our method deals with the second type of variability and involves a processing branch specific to each modality. Each branch includes a single-modal encoder followed by an instance attention aggregator, and given $\{x^m_{n,i} | i = 0, ..., C(n, m)\}$ as the input, it generates a single representation vector of $R^m_n \in \mathbb{R}^{1\times d}$.

The encoder of the branch is designed to be a GNN model consisting of three GraphSAGE layers [14], each of which is followed by a SAGPool [21]. The SAGPooling layers enable the model to perform hierarchical pooling by selecting the most important nodes in the graph. Subsequently, the average and max pooling embeddings of the graph nodes after each SAGPool are concatenated, added together for different pooling layers, and passed through a 2-layer MLP (multi-layer perceptron).

Considering that the input graphs from different modalities have comparable context, they can benefit from the shared-context processing explained in Sec. 3.2. To perform low-level feature unification, we couple the first layer of each branch using matrix factorization. More specifically, our GraphSAGE layers follow Eq. (1)

$$\hat{h}^m_k = W_s W_m [h^m_k, \frac{1}{K} \sum_{j\in\mathcal{N}_k} h^m_j]. \quad (1)$$

In this equation, h^m_k and $\hat{h}^m_k$ are the embeddings of the node k before and after the layer, W_s is the weight shared across modality branches, W_m is the weight specific to the modality branch of m, $[.,.]$ is the concatenation operation, $\mathcal{N}_k$ is the set of nodes that are connected to node k, and K

is the size of $\mathcal{N}_k$. For the first GraphSAGE layer of each modality branch, we set the W_s to be a learnable matrix, while it will be an all-ones matrix for other layers.

In order to combine the embeddings of multiple graphs of a patient within a specific modality, an instance attention, followed by an instance normalization [39], is used similar to Eq. (2)

$$R_n^m = InstanceNorm(\sum_{i=0}^{C(n,m)} \sigma(W R_{n,i}^m) R_{n,i}^m), \quad (2)$$

where $R_{n,i}^m$ is the corresponding representation of $x_{n,i}^m$, W is a learnable matrix, and σ is the sigmoid function. The instance attention layer performs the high-level aggregation part of the shared-context processing (Sec. 3.2) by sharing W across all modalities.

3.3.2 Cross-modal Aggregation

In order to combine the representations from different modalities (R_n^m), we adopted a Transformer model [41]. The equation of each attention head follows Eq. (3)

$$H_n = softmax(\frac{Q_n K_n^T}{\sqrt{d}}) V_n, \quad (3)$$

where Q_n, K_n, V_n are $M \times d$ matrices resulting from applying linear transformations over the representations matrix of the patient n (i.e., $concat\{R_n^1, ..., R_n^M\}$). Finally, all of the heads are concatenated, passed through an MLP, and averaged across modalities to generate an embedding for the patient.

By obtaining a general understanding of the representations, the Transformer enables our model to do a cross-attention across all the stains to emphasise the most informative ones.

3.3.3 Sparse Processing

Despite the fact that previous studies emphasize on learning the precise topological structure of the input graph for various graph-related applications [45], we find that our multimodal approach is strongly robust to missing information. Similar to recent works in the computer vision domain (*e.g.*, MAE [15]), we use this finding to further reduce the computational complexity of our model. More specifically, in each modality, we perform a masking operation over the feature and adjacency matrices of the input graph as shown in Eq. (4)

$$\hat{X} = MX, \hat{A} = MAM^T, M = P\mathcal{I}, \quad (4)$$

where $X \in \mathbb{R}^{c \times d}$ is the feature matrix of the nodes, $A \in \mathbb{R}^{c \times c}$ is the adjacency matrix of the graph, c is the number nodes in the graph, $\mathcal{I} \in \mathbb{R}^{1 \times c}$ is an all-ones matrix, and $P \in \mathbb{R}^{c \times 1}$ is the mask matrix where each element comes from the Bernoulli distribution with the parameters of $1 - s$, where s is the sparsity ratio. As s increases, the number of non-zero elements in both $\hat{X}$ and $\hat{A}$ decreases, resulting in the reduction of the subsequent computational operations. We refer to this as sparse processing and will demonstrate that our model can maintain its performance even with large sparsity ratios.

3.3.4 Loss Function and BCP Technique

Survival prediction is a challenging task that includes the estimating of the failure time (death) as a continuous variable [20]. In a maximum likelihood estimation terminology, this means that, for a subject failed at a specific time, we have to maximize the failure probability of that subject relative to the other unfailed subjects. Consider t_j and $R(t_j)$ to be the time of failure for subject j and the set of subjects who have survived until time t_j, respectively. The probability of failure for subject j is calculated using Eq. (5)

$$P_j(T = t_j | R(t_j)) = \frac{P_j(T = t_j | T \geq t_j)}{\sum_{i:t_i \geq t_j} P_i(T = t_j | T \geq t_j)}. \quad (5)$$

Our training goal is to maximize this probability for each j. In particular, the expectation of the total loss over a minibatch of $\mathcal{B}$ will be calculated as Eq. (6), in which $\mathcal{U}(.)$ is a uniform distribution over the subjects

$$L_{batch} = -E_{i \sim \mathcal{U}(.)}[\log P_i(T = t_i | R(t_i))]. \quad (6)$$

However, the above loss has a practical issue. The problem rises from the fact that the loss in Eq. (5) is only defined for subjects who have a certain time of failure, and it is undefined for subjects with a survived status in their latest follow-up (we conventionally refer to such subjects as censored data). Therefore, the censored subjects do not provide any gradient in the backpropagation step of Eq. (6) as a separate data point due to their undefined loss (explicit gradient), which interferes with the proper training of the model. Nonetheless, one must note that such subjects still participate in the back-propagation via the denominator of Eq. (5) of non-censored subjects' loss (implicit gradient).

To mitigate this issue, we reformulate the loss function as Eq. (7), in which $\mathcal{U}_C(.)$ and $\mathcal{U}_N(.)$ are uniform distributions over the censored and non-censored subjects, respectively, and k comes from a Bernoulli distribution with the parameter of α

$$L_{batch} = -E_{i \sim k\mathcal{U}_C(.) + (1-k)\mathcal{U}_N(.)}[\log P_i(T = t_i | R(t_i))]. \quad (7)$$

One can note that for an α equal to the percentage of the censored cases, these two equations are equal. However, we will show that the selection of an appropriate value for this parameter results in a balanced trade-off between the implicit and explicit gradient of the censored data. We refer to α as batch censored portion (BCP) and show that it can have a substantial impact on the results.

4. Experiments

4.1. Datasets and Pre-Processing Steps

We used two immunohistochemistry (IHC) datasets in this study: (1) InUIT: internal high-grade serous ovarian cancer cohort with 1,600 TMA cores stained with Ki67, CD8, and CD20 biomarkers collected from 188 patients, (2) MIBC: muscle-invasive bladder cancer cohort with 585 TMA cores stained with Ki67, CK20, and P16 collected from 58 patients [26]. Each patient has at least one TMA core stained with each biomarker, and the latest survival status (alive or dead) along with the overall survival time (since diagnosis) is available for all the patients.

Each cell in a TMA core would appear in either a red or blue color, which shows whether the cell is positive or negative for the corresponding biomarker (Fig. 1). A cell segmentation algorithm was applied to each TMA core to locate and identify the type of cell (positive/negative) [12]. Then, each cell was considered as a node in the graph, and these nodes were connected to each other using a K-nearest neighbor algorithm (K=5). Similar to [18], we hypothesize that there is a biological restriction on the distance of intercellular communications. Therefore, we removed the edges with a length of more than 60 pixels. Afterward, the embedding representation of each node was obtained by applying a pre-trained ResNet34 on a 72×72 crop centered on the corresponding cell in the image. Additionally, the type of node (positive/negative) and its location (relative to the size of the image) were added to the embedding as well.

4.2. Implementation Details

Please refer to the supplemental material.

4.3. Survival Prediction

The summary of survival prediction results can be found in Tab. 1. To compare different models, similar to previous works [8], we used the concordance index (C-Index) which measures the quality of survival ranking of the patients [34]. Also, we used the p-value of the LogRank test to demonstrate the ability of the models in separating the high- and low-risk patients (see Sec. 4.7 for more information). All the experiments were performed in the 3-fold patient-wise cross-validation setting, and in contrast to the previous works, we conducted each experiment with 3 different seeds to account for the initialization variability. The results confirm that our model can outperform all of the baselines, including ViT, and has a consistent performance in both metrics across both datasets, unlike the baselines. More specifically, our model reaches the c-index of 0.57 and 0.61 for InUIT and MIBC datasets while the closest baseline performances are 0.55 and 0.59, respectively. Additionally, our model can separate the low- and high-risk patients on both datasets significantly (p-value < 0.01) while being the only method to do so on the InUIT dataset. It is worth mentioning that a few of the baseline models achieve a C-Index of 0.5 (equivalent to random prediction), that could be attributed to the aforementioned issues of MIL-based techniques. We also notice that our model has less performance variation and number of parameters compared to the baselines, and this observation shows the generalizability and efficiency of our model, which can be linked to its cellular foundation.

The setting for the baseline models was set similar to that of [7]. Although the original setting used ResNet50 with a dimension of 1024 for feature extraction, we also conducted our experiments with ResNet34 to ensure a fair comparison between our model and the baselines (more results and visualizations in the supplementary).

4.4. Ablation Study

We conducted ablation studies on different parts of our model, the results of which can be found in Tab. 2. These experiments included the removal of the instance normalization after the instance attention (no instance norm), decoupling the modality branch weights (no weight sharing), fully coupling the modality branch weights (full weight sharing), decoupling the weights of instance attention layers (non-shared attention), no consideration of batch censored portion (no BCP), using Transformer instead of instance attention (Transformer attention), and applying sparsity at inference time (inference-time sparsity). As can be seen, depending on the dataset, each ablated feature shows a noticeable reduction in the performance of the model, emphasizing the importance of each of our design choices (more results in the supplementary).

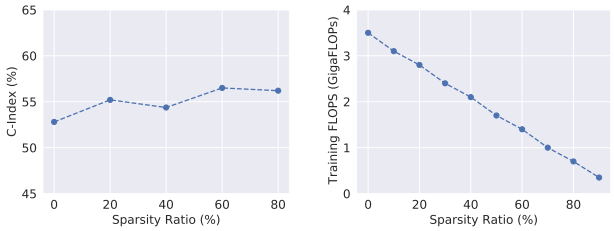
As was elaborated, although different modalities of our data represent different stains, we believe there is a shared contextual information in all of these modalities. As a result, we can take advantage of it by employing the previously presented idea of shared-context processing. Our ablation experiments on the elimination of this step (no weight sharing and non-shared attention rows of Tab. 2) confirm this hypothesis. On the other hand, one might argue that using the same network for all of the modalities might achieve this purpose as well. However, the corresponding ablation study (full weight sharing row of the table) invalidates this argument. Additionally, since all the modalities are processed using a shared instance attention, our model could

Method	Feature Extractor	Parameters	InUIT		MIBC	
			C-Index (↑)	P-value (↓)	C-Index (↑)	P-value (↓)
DeepSet	ResNet34	$395K$	0.50 ± 0.0	0.43	0.50 ± 0.001	–
	ResNet50	$657K$	0.53 ± 0.007	0.40	0.45 ± 0.004	0.28
Attention MIL	ResNet34	$657K$	0.51 ± 0.004	0.62	0.59 ± 0.007	0.04
	ResNet50	$920K$	0.55 ± 0.004	0.65	0.55 ± 0.004	0.57
DGC	ResNet34	$658K$	0.53 ± 0.007	0.46	0.58 ± 0.007	< 0.01
	ResNet50	$790K$	0.55 ± 0.005	0.31	0.54 ± 0.007	0.64
Patch-GCN	ResNet34	$1.3M$	0.53 ± 0.008	0.45	0.50 ± 0.004	< 0.01
	ResNet50	$1.4M$	0.50 ± 0.004	0.25	0.46 ± 0.009	0.33
Pathomic Fusion	CPC	$368K$	0.51 ± 0.001	0.43	0.52 ± 0.003	0.56
HIPT	Hierarchical ViT	$23.8M$	0.50 ± 0.002	0.18	0.53 ± 0.010	0.10
AMIGO (Ours)	ResNet34	$451K$	$\mathbf{0.57 \pm 0.002}$	**0.01**	$\mathbf{0.61 \pm 0.004}$	$\mathbf{< 0.01}$

Table 1. Survival prediction performance (average and variance) comparison of our model with all the baselines on two datasets.

Ablated Feature	InUIT		MIBC	
	C-Index (↑)	P-value (↓)	C-Index (↑)	P-value (↓)
No instance Norm	0.53 ± 0.001	0.15	0.54 ± 0.005	0.03
No weight sharing	0.56 ± 0.001	0.06	0.54 ± 0.016	0.001
Full weight sharing	0.53 ± 0.001	0.46	0.51 ± 0.005	0.78
No BCP	0.54 ± 0.001	0.07	0.58 ± 0.013	0.38
Transformer attention	0.53 ± 0.002	0.05	0.55 ± 0.011	0.90
Inference-time sparsity	0.56 ± 0.001	0.04	0.55 ± 0.004	0.08
Non-shared attention	0.54 ± 0.001	0.13	0.58 ± 0.003	0.06
AMIGO (Ours)	$\mathbf{0.57 \pm 0.002}$	**0.01**	$\mathbf{0.61 \pm 0.004}$	$\mathbf{< 0.001}$

Table 2. Ablation Studies.



(a) Sparsity C-Index (b) Sparsity Flops

Figure 4. Fig. 4a shows that the final performance of our model on the InUIT dataset is robust to the sparsity of training data. Fig. 4b demonstrates that the computational cost of our model (FLOPs) has a reverse linear relationship with the sparsity ratio.

benefit from a normalization layer before passing the embeddings to the cross-modal aggregator (no instance norm).

Our result with the removal of the BCP also demonstrates that a trade-off between the portion of the censored and non-censored data is important as it can improve the gradient signals in the backpropagation. Finally, avoiding adding sparsity at inference time results in a higher performance as the model would have access to all of the information needed for making a prediction.

4.5. Sparsity Robustness and Computational Efficiency

One of the most important findings of our study is the robustness of our model against training data sparsity. More specifically, we realized that our model's final performance is stable regardless of the sparsity ratio of the input graph. In particular, although previous digital histopathology studies [29] suggest that the learning of the complete topological structure of the cellular graph is critical for the downstream tasks, we noticed that a small sparsity ratio (20%) can increase the model performance (Fig. 4a). This observation is consistent with previous findings where they show that deep learning models can benefit from data augmentation due to the prevention of over-fitting [35]. However, the performance of our model surprisingly stays almost the same as we increase the sparsity ratio. On the other side, this sparsity ratio has a reverse linear relationship with the computational cost of the model (Fig. 4b), suggesting that higher sparsity ratios lead to a lower number of computational operations. As a result, the computational cost of our model can be significantly reduced (from 3.5 to 0.7 GigaFlops), while achieving the same performance.

4.6. BCP Effect

We also measured the effect of BCP on the performance of our model. As can be seen in Fig. 5, a BCP of 0 (no censored data in the batch) can results in a better performance compared to the typical uniform batching as it increases the explicit gradient signals during training. On the other hand, high values of BCP result in a lower performance compared to the uniform batching as it eliminates the explicit gradient. However, the results depicted in this figure confirm our hypothesis regarding achieving the highest performance by selecting a suitable value of BCP (0.1) due to the trade-off between the implicit and explicit gradients.

4.7. Patient Stratification

While c-index is a common measure to benchmark various survival prediction models, it is not particularly informative for patient management. For ML-based survival prediction models to become applicable in the clinic, one would need to show their utility in stratifying patients into

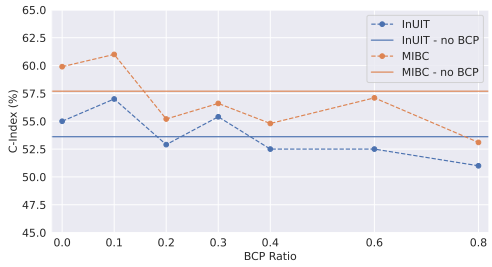
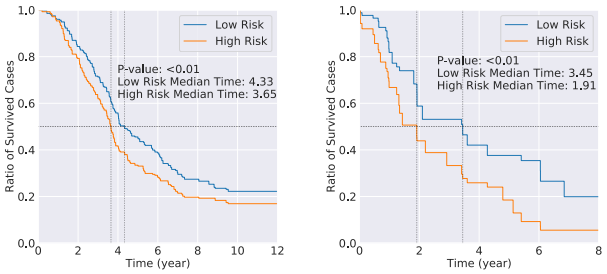


Figure 5. The performance of the model based on the ratio of BCP. Low BCP values lead to an increase in the explicit gradient and a decrease in the implicit gradient of the censored cases. A trade-off between these two types of gradients can produce the highest performance.



(a) InUIT (b) MIBC

Figure 6. Survival curves for cohorts of patients identified as low-risk (predicted hazard $<$ median of hazards) and high-risk (prediction hazard $>$ median of hazards) by our model.

various risk groups. Therefore, we divide the patients in each dataset into two groups (i.e., low- and high-risk) based on the predicted risk scores of our model. Kaplan-Meier (KM) survival curves of these cohorts are shown in Fig. 6. To test the significance of the difference between the KM curves of the low- and high-risk patient categories, we employ log-rank test [2]. All the patients in both datasets are treated uniformly, and to the best of our knowledge, there are no clinical parameters that can perform this separation. However, as shown in Fig. 6, our model can successfully separate both ovarian cancer and bladder cancer cases into low- and high-risk cohorts, highlighting the ability of the model in picking meaningful contextual insights from the histopathology images. In particular, the median survival time for the high-risk and low-risk cohorts are 3.65 and 4.33 years for the InUIT dataset (log-rank p-value = 0.01) and 1.91 and 3.45 years for the MIBC dataset (log-rank p-value < 0.001), respectively. It is of note to mention that our findings comply with a previous study on the bladder cancer dataset, where Mi *et al*. [26] showed similar separation for the muscle-invasive bladder cancer patients. Although they used manually engineered features from the cells, we approached this problem in an end-to-end trainable manner while considering cellular interactions. Our ovarian cancer dataset represents a highly aggressive subtype (i.e., high-grade serous), and the majority of the efforts (though mainly unsuccessful) in the last few decades have focused on identifying biomarkers of therapy response for these patients. A study by Wang *et al*. [43] demonstrated that such markers could be found from global genomic aberration profiles and our study is the first that has led to promising results based on routine histopathology slide images.

5. Conclusion

In this work, we developed, for the first time, a multi-modal GNN for the processing of histopathology images by focusing on cells and their interactions. Alongside introducing new techniques such as Batch Censored Portion (BCP) and shared-context processing, we showed that our proposed model can outperform all of its counterparts in two datasets representing ovarian and muscle-invasive bladder cancer. More importantly, we demonstrated that the proposed model is strongly robust to the sparsity of the data, to the extent that it still achieves relatively similar performance with as low as 20% of the data during training. By taking advantage of this observation, we were able to reduce the computational costs of the model even further. We also evaluated the applicability of our model as a tool for patient stratification, where it could split the patients into statistically significant low-risk and high-risk groups.

We believe that our proposed model highlights the importance of heterogeneity, spatial positioning, and mutual interactions of the cells for image representation across different cancer types. We hope this work can open new interesting pathways toward the efficient cell-based processing of histopathology images. Considering the success of our model in stratifying cohorts of patients that can only be separated using genomic information, we can use it as an engine to link histopathology images to gene expression, mutation, and genomic traits, where deeper analysis and biological interrogations can be performed. Furthermore, the cell-centricity of our approach offers an opportunity to identify visually-interpretable biological entities that play a key role in predicting outcomes and could be used in clinics.

Acknowledgement

This work was supported by the Canadian Institute of Health Research, Natural Sciences and Engineering Research Council of Canada, Michael Smith Foundation for Health Research, OVCARE Carraresi, VGH UBC Hospital Foundation, and the National Institutes of Health of United States (grant number R01CA138264).

References

[1] Mohammed Adnan, Shivam Kalra, and Hamid R Tizhoosh. Representation learning of histopathology images using graph neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 988–989, 2020. 3

[2] J Martin Bland and Douglas G Altman. The logrank test. *Bmj*, 328(7447):1073, 2004. 8

[3] Joseph Boyd, Mykola Liashuha, Eric Deutsch, Nikos Paragios, Stergios Christodoulidis, and Maria Vakalopoulou. Self-supervised representation learning using visual field expansion on digital pathology. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 639–647, 2021. 1

[4] Wieland Brendel and Matthias Bethge. Approximating cnns with bag-of-local-features models works surprisingly well on imagenet. *arXiv preprint arXiv:1904.00760*, 2019. 3

[5] Gabriele Campanella, Matthew G Hanna, Luke Geneslaw, Allen Miraflor, Vitor Werneck Krauss Silva, Klaus J Busam, Edi Brogi, Victor E Reuter, David S Klimstra, and Thomas J Fuchs. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature medicine*, 25(8):1301–1309, 2019. 3

[6] Richard J Chen, Chengkuan Chen, Yicong Li, Tiffany Y Chen, Andrew D Trister, Rahul G Krishnan, and Faisal Mahmood. Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16144–16155, 2022. 1, 2, 3

[7] 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. 6

[8] Richard J Chen, Ming Y Lu, Jingwen Wang, Drew FK Williamson, Scott J Rodig, Neal I Lindeman, and Faisal Mahmood. Pathomic fusion: an integrated framework for fusing histopathology and genomic features for cancer diagnosis and prognosis. *IEEE Transactions on Medical Imaging*, 2020. 3, 4, 6

[9] Richard J Chen, Ming Y Lu, Drew FK Williamson, Tiffany Y Chen, Jana Lipkova, Zahra Noor, Muhammad Shaban, Maha Shady, Mane Williams, Bumjin Joo, et al. Pan-cancer integrative histology-genomic analysis via multimodal deep learning. *Cancer Cell*, 40(8):865–878, 2022. 4

[10] Chaitanya Dwivedi, Shima Nofallah, Maryam Pouryahya, Janani Iyer, Kenneth Leidal, Chuhan Chung, Timothy Watkins, Andrew Billin, Robert Myers, John Abel, et al. Multi stain graph fusion for multimodal integration in pathology. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1835–1845, 2022. 4

[11] Jevgenij Gamper and Nasir Rajpoot. Multiple instance captioning: Learning representations from histopathology textbooks and articles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16549–16559, 2021. 1

[12] Parmida Ghahremani, Yanyun Li, Arie Kaufman, Rami Vanguri, Noah Greenwald, Michael Angelo, Travis J Hollmann, and Saad Nadeem. Deepliif: Deep learning-inferred multiplex immunofluorescence for ihc image quantification. *bioRxiv*, 2021. 6

[13] Yonghang Guan, Jun Zhang, Kuan Tian, Sen Yang, Pei Dong, Jinxi Xiang, Wei Yang, Junzhou Huang, Yuyao Zhang, and Xiao Han. Node-aligned graph convolutional network for whole-slide image representation and classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18813–18823, 2022. 3

[14] Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. *Advances in neural information processing systems*, 30, 2017. 4

[15] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022. 5

[16] Le Hou, Dimitris Samaras, Tahsin M Kurc, Yi Gao, James E Davis, and Joel H Saltz. Efficient multiple instance convolutional neural networks for gigapixel resolution image classification. *arXiv preprint arXiv:1504.07947*, 7:174–182, 2015. 3

[17] Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *International conference on machine learning*, pages 2127–2136. PMLR, 2018. 1, 3

[18] Sajid Javed, Arif Mahmood, Muhammad Moazam Fraz, Navid Alemi Koohbanani, Ksenija Benes, Yee-Wah Tsang, Katherine Hewitt, David Epstein, David Snead, and Nasir Rajpoot. Cellular community detection for tissue phenotyping in colorectal cancer histology images. *Medical image analysis*, 63:101696, 2020. 6

[19] Zhipeng Jia, Xingyi Huang, I Eric, Chao Chang, and Yan Xu. Constrained deep weak supervision for histopathology image segmentation. *IEEE transactions on medical imaging*, 36(11):2376–2388, 2017. 3

[20] Jared L Katzman, Uri Shaham, Alexander Cloninger, Jonathan Bates, Tingting Jiang, and Yuval Kluger. Deepsurv: personalized treatment recommender system using a cox proportional hazards deep neural network. *BMC medical research methodology*, 18(1):1–12, 2018. 5

[21] Junhyun Lee, Inyeop Lee, and Jaewoo Kang. Self-attention graph pooling. In *International conference on machine learning*, pages 3734–3743. PMLR, 2019. 4

[22] 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. 3

[23] Jiayun Li, Wenyuan Li, Anthony Sisk, Huihui Ye, W Dean Wallace, William Speier, and Corey W Arnold. A multi-resolution model for histopathology image classification and

localization with multiple instance learning. *Computers in biology and medicine*, 131:104253, 2021. 3

[24] Wenqi Lu, Simon Graham, Mohsin Bilal, Nasir Rajpoot, and Fayyaz Minhas. Capturing cellular topology in multi-gigapixel pathology images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 260–261, 2020. 3

[25] Armin Meier, Katharina Nekolla, Lindsay C Hewitt, Sophie Earle, Takaki Yoshikawa, Takashi Oshima, Yohei Miyagi, Ralf Huss, Günter Schmidt, and Heike I Grabsch. Hypothesis-free deep survival learning applied to the tumour microenvironment in gastric cancer. *The Journal of Pathology: Clinical Research*, 6(4):273–282, 2020. 2

[26] Haoyang Mi, Trinity J Bivalacqua, Max Kates, Roland Seiler, Peter C Black, Aleksander S Popel, and Alexander S Baras. Predictive models of response to neoadjuvant chemotherapy in muscle-invasive bladder cancer using nuclear morphology and tissue architecture. *Cell Reports Medicine*, 2(9):100382, 2021. 6, 8

[27] Ramin Nakhli, Amirali Darbandsari, Hossein Farahani, and Ali Bashashati. Ccrl: Contrastive cell representation learning. In *Computer Vision–ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VII*, pages 397–407. Springer, 2023. 1

[28] Ramin Nakhli, Allen Zhang, Hossein Farahani, Amirali Darbandsari, Elahe Shenasa, Sidney Thiessen, Katy Milne, Jessica McAlpine, Brad Nelson, C Blake Gilks, et al. Volta: an environment-aware contrastive cell representation learning for histopathology. *arXiv preprint arXiv:2303.04696*, 2023. 1

[29] Pushpak Pati, Guillaume Jaume, Antonio Foncubierta-Rodríguez, Florinda Feroce, Anna Maria Anniciello, Giosue Scognamiglio, Nadia Brancati, Maryse Fiche, Estelle Dubruc, Daniel Riccio, et al. Hierarchical graph representations in digital pathology. *Medical image analysis*, 75:102264, 2022. 7

[30] Simon Reiß, Constantin Seibold, Alexander Freytag, Erik Rodner, and Rainer Stiefelhagen. Every annotation counts: Multi-label deep supervision for medical image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9532–9542, 2021. 1

[31] Yash Sharma, Aman Shrivastava, Lubaina Ehsan, Christopher A Moskaluk, Sana Syed, and Donald Brown. Cluster-to-conquer: A framework for end-to-end multi-instance learning for whole slide image classification. In *Medical Imaging with Deep Learning*, pages 682–698. PMLR, 2021. 1

[32] Beomseok Son, Sungmin Lee, HyeSook Youn, EunGi Kim, Wanyeon Kim, and BuHyun Youn. The role of tumor microenvironment in therapeutic resistance. *Oncotarget*, 8(3):3933, 2017. 2

[33] Maximilian Stadler, Bertrand Charpentier, Simon Geisler, Daniel Zügner, and Stephan Günnemann. Graph posterior network: Bayesian predictive uncertainty for node classification. *Advances in Neural Information Processing Systems*, 34:18033–18048, 2021. 3

[34] Harald Steck, Balaji Krishnapuram, Cary Dehing-Oberije, Philippe Lambin, and Vikas C Raykar. On ranking in survival analysis: Bounds on the concordance index. *Advances in neural information processing systems*, 20, 2007. 6

[35] Susheel Suresh, Pan Li, Cong Hao, and Jennifer Neville. Adversarial graph augmentation to improve graph contrastive learning. *Advances in Neural Information Processing Systems*, 34:15920–15933, 2021. 7

[36] Aiham Taleb, Matthias Kirchler, Remo Monti, and Christoph Lippert. Contig: Self-supervised multimodal contrastive learning for medical imaging with genetics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20908–20921, 2022. 1

[37] Shixiang Tang, Dapeng Chen, Lei Bai, Kaijian Liu, Yixiao Ge, and Wanli Ouyang. Mutual crf-gnn for few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2329–2339, 2021. 3

[38] Yucheng Tang, Dong Yang, Wenqi Li, Holger R Roth, Bennett Landman, Daguang Xu, Vishwesh Nath, and Ali Hatamizadeh. Self-supervised pre-training of swin transformers for 3d medical image analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20730–20740, 2022. 1

[39] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, 2016. 5

[40] Luís A Vale-Silva and Karl Rohr. Long-term cancer survival prediction using multimodal deep learning. *Scientific Reports*, 11(1):1–12, 2021. 4

[41] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 5

[42] Yanan Wang, Yu Guang Wang, Changyuan Hu, Ming Li, Yanan Fan, Nina Otter, Ikuan Sam, Hongquan Gou, Yiqun Hu, Terry Kwok, et al. Cell graph neural networks enable the precise prediction of patient survival in gastric cancer. *NPJ precision oncology*, 6(1):1–12, 2022. 2, 3

[43] Yi Kan Wang, Ali Bashashati, Michael S Anglesio, Dawn R Cochrane, Diljot S Grewal, Gavin Ha, Andrew McPherson, Hugo M Horlings, Janine Senz, Leah M Prentice, et al. Genomic consequences of aberrant dna repair mechanisms stratify ovarian cancer histotypes. *Nature genetics*, 49(6):856–865, 2017. 8

[44] An Xu, Wenqi Li, Pengfei Guo, Dong Yang, Holger R Roth, Ali Hatamizadeh, Can Zhao, Daguang Xu, Heng Huang, and Ziyue Xu. Closing the generalization gap of cross-silo federated medical image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20866–20875, 2022. 1

[45] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? *arXiv preprint arXiv:1810.00826*, 2018. 5

[46] Yinyin Yuan. Spatial heterogeneity in the tumor microenvironment. *Cold Spring Harbor perspectives in medicine*, 6(8):a026583, 2016. 2

[47] Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Russ R Salakhutdinov, and Alexander J Smola. Deep sets. *Advances in neural information processing systems*, 30, 2017. 3

[48] Hongrun Zhang, Yanda Meng, Yitian Zhao, Yihong Qiao, Xiaoyun Yang, Sarah E Coupland, and Yalin Zheng. Dtfd-mil: Double-tier feature distillation multiple instance learning for histopathology whole slide image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18802–18812, 2022. 3

[49] Yushan Zheng, Bonan Jiang, Jun Shi, Haopeng Zhang, and Fengying Xie. Encoding histopathological wsis using gnn for scalable diagnostically relevant regions retrieval. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 550–558. Springer, 2019. 3