CO-PILOT: Dynamic Top-Down Point Cloud with Conditional Neighborhood Aggregation for Multi-Gigapixel Histopathology Image Representation

Ramin Nakhli, Allen Zhang, Ali Mirabadi, Katherine Rich,
Maryam Asadi, Blake Gilks, Hossein Farahani*, Ali Bashashati*

University of British Columbia

Abstract

Predicting survival rates based on multi-gigapixel histopathology images is one of the most challenging tasks in digital pathology. Due to the computational complexities, Multiple Instance Learning (MIL) has become the conventional approach for this process as it breaks the image into smaller patches. However, this technique fails to account for the individual cells present in each patch, while they are the fundamental part of the tissue. In this work, we developed a novel dynamic and hierarchical point-cloud-based method (CO-PILOT) for the processing of cellular graphs extracted from routine histopathology images. By using bottom-up information propagation and top-down conditional attention, our model gains access to an adaptive focus across different levels of tissue hierarchy. Through comprehensive experiments, we demonstrate that our model can outperform all the state-of-the-art methods in survival prediction, including the hierarchical Vision Transformer (ViT), across three datasets and four metrics with only half of the parameters of the closest baseline. Importantly, our model is able to stratify the patients into different risk cohorts with statistically different outcomes across three large datasets, a task that was previously achievable only using genomic information. Furthermore, we publish a large dataset containing 873 cellular graphs from 188 patients, along with their survival information, making it one of the largest publicly available datasets in this context.

1. Introduction

The utilization of deep learning models in the digital processing of medical images has garnered substantial interest in the computer vision community, where these models have been used for a wide range of image types (*e.g.*, histopathology images and CT scans) and tasks (*e.g.*, classification, segmentation, and survival prediction) [41, 6, 49, 28, 47, 30, 37, 40, 29]. The ability of these models to learn meaningful features from raw images with little to no supervision has created exciting opportunities, especially in digitized histopathology where unique challenges are posed due to the large scale and high granularity of input images, also known as Whole-Slide Images or WSIs (Fig. 1).

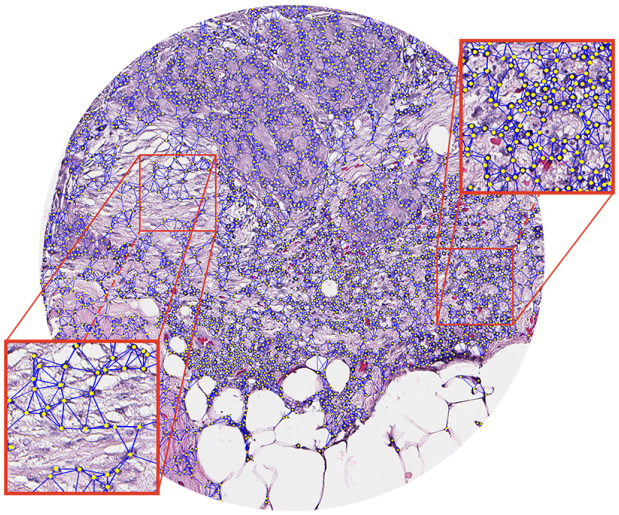


Figure 1: Cellular graph constructed by connecting the adjacent nodes within a $4,000 \times 4,000$ pixels image. The two enlarged windows demonstrated two different tissue types with distinct spatial positioning and composition of the cells. The window on the right demonstrates a high-density area while the one on the left is associated with a low-density tissue region.

The high resolution and intricate details of WSIs (each image reaching up to 150,000×150,000 pixels in size) can pose intriguing challenges in computer vision such as memory limit issues during end-to-end training. To address the computational difficulties, Multiple Instance Learning (MIL) techniques are often used as the main training strategy. These techniques divide the WSI into smaller patches, pass them through a pre-trained feature extractor, and aggregate the patch embeddings to provide a representation for the whole slide. Although MIL has shown promising results in several tasks, including cancer subtype classi-

fication and survival prediction, it has several significant shortcomings [20, 2, 6]. Firstly, due to the large number of patches generated from the high-resolution WSIs, most studies utilize either a simple pooling [20] or hierarchical aggregation [6]. However, the former limits the representational capacity of the model, and the latter requires substantial computational power. Secondly, the bottom-up information flow of these methods prevents them from attending to high-granular details lying at higher resolutions. On the other hand, acquiring a top-down aggregation strategy along with the bottom-up information flow can potentially address this issue. Thirdly, the training process is heavily dependent on the number of available images, resulting in lower generalizability when only a limited number of data points are available. Lastly, focusing on patches rather than individual cells leads to missing the mutual interactions of cells, thereby reducing the representation power of the model toward the biological basis.

Various studies have shown that the spatial positioning of the cells and their mutual interactions can have a prominent impact on the progression of the tumor [35, 50, 34]. For instance, Sirinukunwattana et al. [34] have shown that quantitative statistics extracted from cell-cell connections can provide meaningful insights into cancer metastasis, and Son et al. [35] explained different roles of tumor-tumor, tumor-stromal, and tumor-extracellular matrix connections in the development of therapeutic resistance. Hence, directing attention toward the processing of cellular structures may boost models' performance by offering a comprehensive, multi-scaled perspective of the tissue.

In this study, with a point-cloud perspective, we investigate the utilization of graph neural networks (GNNs) for the representation learning of the histopathology images through the dynamic and hierarchical processing of the cellular graphs extracted from these images. Cellular graphs grant our model the ability to examine cell-level information and the interconnections between cells (Fig.1). The versatility in focus at different scales (ranging from cell to tissue level) permits the model to have a multi-faceted perspective of the tissue. This is in contrast to MIL models, which only examine patches with a pre-determined resolution and magnification. Furthermore, compared to the costly hierarchical pooling procedure in Visual-Transformer-based methods [6], GNNs offer a more efficient approach for the processing of WSIs due to the weight sharing across the graph nodes. Consequently, this could help with the mitigation of over-parameterization issues in low-data regimes.

We present a dynamic top-down point-cloud-based GNN with conditional neighborhood aggregation (CO-PILOT) for learning the representation of histopathology images. Our model begins by processing information at the cellular level and gradually expands to larger neighborhoods of cells, capturing the hierarchical structure of the tissue. Through a bottom-up hierarchical process, our model combines the representation of each cell with its surrounding neighbors using conditional and position-aware information propagation. However, it utilizes a top-down procedure to aggregate the representations from higher to lower levels. This enables our model to attend to finer details in the tissue, which is critical for challenging tasks like survival prediction. Our work advances the frontiers of MIL, Vision Transformer (ViT), and GNNs in multiple directions:

- We introduce the first dynamic top-down GNN on cellular graphs with conditional neighborhood aggregation that achieves state-of-the-art survival prediction results across three large datasets comprising 872 patients.
- CO-PILOT eliminates the critical barriers of MIL models, enabling efficient training of multi-gigapixel images on a single GPU and outperforming all the baselines including Vision Transformer (ViT). It also implements the hierarchical structure of ViT while keeping the number of parameters significantly lower during end-to-end training.
- For the first time, we demonstrate that it is possible to stratify high-grade serous patients (the most aggressive and common subtype of ovarian cancer) into different risk groups soley based on routine hematoxylin and eosin (H&E)-stained tissue slides.
- We will also publish a large cellular graph dataset, containing 873 graphs from 188 high-grade serous ovarian patients along with their survival information. To be best of our knowledge, this dataset is one of the largest datasets in this context.

2. Related Work

2.1. Multiple Instance Learning in Histopathology

The concept of permutation-invariant bag-of-features for image representation learning was first introduced by Zaheer *et al*. [52] and Brendel *et al*. [3]. Early works in digital histopathology employed similar techniques to learn representations of WSIs by aggregating information at the patch level [17, 21]. However, later works introduced more sophisticated methods. For instance, Ilse *et al*. [20] proposed the use of attention-based pooling, Campanella *et al*. [4] introduced RNN-based aggregation, Li *et al*. [25, 24] designed a self-supervised multi-resolution MIL, and Zhang *et al*. [53] utilized the pseudo-bagging in a double-tier setting. Chen *et al*. [6] also recently proposed using the large-scale Vision Transformers for the hierarchical pooling of WSIs, while Thandiackal *et al*. [41] designed a differentiable zooming strategy for multi-scale attention. Despite

these advancements, the aforementioned studies still fail to take into account the cell-level details present in the images and require a large amount of data for training. In this work, we aim to address these limitations by using a cell-centric method for hierarchical and dynamic information propagation across different sections of the image.

2.2. Graph Neural Networks in Histopathology

Recent achievements in graph neural networks (GNNs) have attracted significant attention, as they have obtained exceptional results in various tasks due to their capability in preserving the structural information of data [39, 36]. GNNs are well-suited for capturing spatial relationships in histopathology images as well [13]. Several studies in computational histopathology have leveraged GNNs as aggregator models to combine representations of WSIs at the patch level. For example, Adnan *et al*. [1] used a fully-connected graph constructed from the most informative patches to produce a graph-level representation via a GNN. Meanwhile, Lu *et al*. [27] combined similar patches into nodes in a graph and processed it with a GNN. Zheng *et al*. [55] applied GNNs to provide a hashing mechanism for retrieving contextually similar regions of interest in response to a query image. In line with our work, Chen *et al*. [8] and Wang *et al*. [45] used cellular graphs for survival prediction. However, these studies differ from our approach as they ignore the conditional dependency of cells during the message propagation and do not take the top-down dependency of the tissue structure into account. Additionally, cellular-based processing and dynamic graph construction along with the positional encoding are the two main differentiation factors of our work and the prior graph-based approaches.

2.3. Point-Clouds

Due to its ability in capturing fine details of complex structures, the processing of unordered Cartesian points (also known as point-clouds) has become a popular approach for 3D object representation. PointCNN [26] uses a convolutional neural network (CNN) that operates directly on point-clouds, enabling it to capture both local and global features. Similarly, KPConv [42] uses a kernel point convolution approach that is able to process large point clouds efficiently while also preserving fine details. Wang *et al*. [44] take an adaptive approach to the processing of the point cloud, and He *et al*. [16] propose a density-preserving architecture to improve the reconstruction ability of the network. On the other hand, Pang *et al*. [31] investigated the utility of masked autoencoders as a self-supervised pre-training tool for downstream tasks, and Choe *et al*. [9] explored the utility of MLP-Mixer in point cloud understanding. Even though these architectures have shown impressive performance on a variety of tasks, including segmentation [16, 9], classification [26, 44], and object detection [11], they either have to limit the number of input nodes in the point-cloud to a fixed number or reduce the batch size to 1 which subsequently deducts the performance of the model. As the first study to investigate the point-cloud utility in cellular graphs, we address this problem by using a dynamic GNN model, allowing the processing of arbitrary-size point-clouds.

3. Method

3.1. Problem Formulation

Let the set of images in a given dataset be denoted by $\{x_{n,k}|n = 0, ..., N; k = 0, ..., K(n)\}$. In this notation, n represents the unique patient number, k serves as the identifier for the images corresponding to patient n, N represents the total number of patients in the dataset, and $K(n)$ specifies the number of images available for patient n. Our aim is to obtain a representation vector of $R_n \in \mathbb{R}^{1\times d}$ given all the available images of patient n. Finally, we will use the aforementioned representation to predict the estimated survival time (outcome) of the patient. In the rest of the paper, we refer to $x_{n,k}$ as both the image and the point cloud of cells extracted from this image to prevent any duplications.

3.2. Model Architecture

Our framework, as depicted in Fig. 2, consists of six Dynamic Neighborhood Processing Units (DNPUs) that are grouped in pairs to form three main blocks. These blocks are responsible for low-, mid-, and high-level feature processing. To minimize the number of parameters, the DNPUs in each block are closely weight-coupled together.

Given $x_{n,k}$ as the input, all the points in the cloud are processed by the stack of DNPUs to provide a set of features for each point. These processed features are then passed to the designated Waterfall attention module, which performs top-down conditional attention at different scales of hierarchy to generate the final patient representation of $R_{n,k} \in \mathbb{R}^{1\times d}$. To combine all the representations of a patient, we employ an instance attention resulting in the final representation vector of $R_n \in \mathbb{R}^{1\times d}$, which is then used for downstream tasks such as survival prediction.

We will elaborate on each part of our model in more detail in the following sections.

3.2.1 Dynamic Neighborhood Processing Unit (DNPU)

Histopathology tissue samples often exhibit a complex, intricate hierarchical structure involving the interplay of cells and their neighbors at a local level, as well as interactions between different tissue types at a larger scale. In order to effectively capture this complex hierarchy, our design utilizes a dynamic graph construction, a neighborhood message passing operations, and a hierarchical node pooling

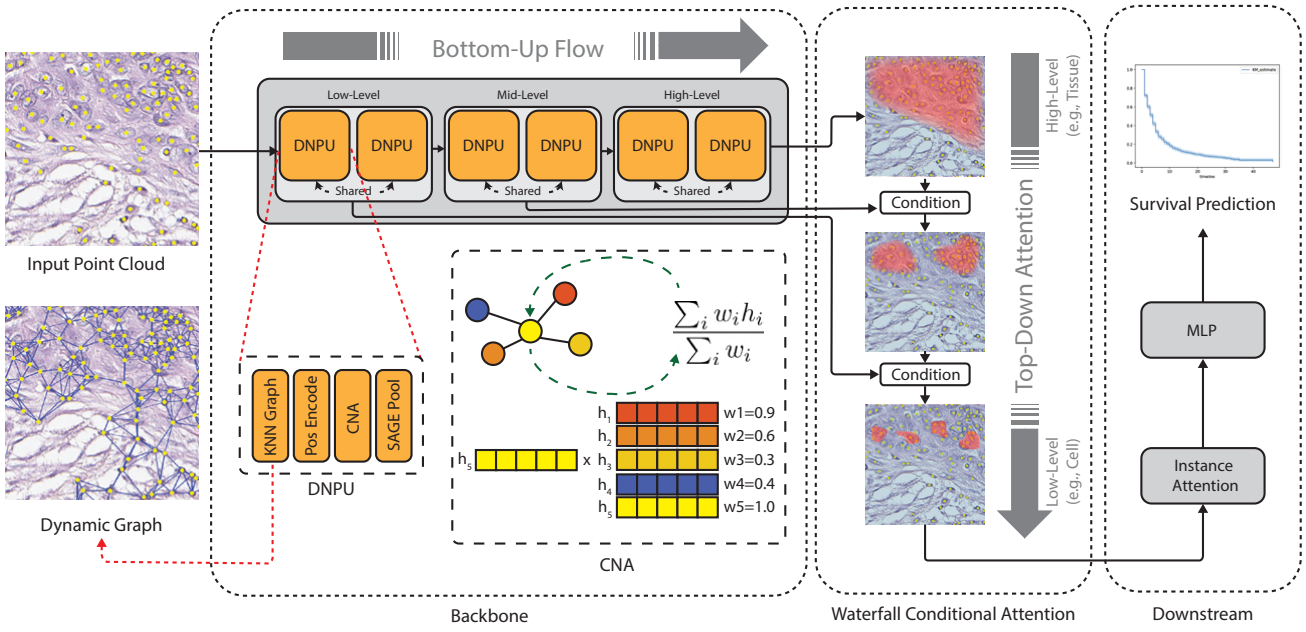


Figure 2: Our proposed framework, CO-PILOT, comprises of 6 DNPU layers that are weight-coupled in pairs, resulting in 3 low-, mid-, and high-level blocks. Each DNPU layer consists of a dynamic graph constructor, a position encoding module, a CNA (Conditional Neighborhood Aggregations) layer, and a SAGE pooling layer. The positional encoding module takes into consideration the spatial location and cellular density during aggregation, while the CNA layer combines the contextual information from neighbors on a conditional basis. Along with the bottom-up information flow within the backbone, the top-down conditional mechanism of the Waterfall attention ensures that the model can properly take the finer details into account. Finally, for each patient, the obtained representations are combined using the instance attention module, converted to hazard risks, and used for survival prediction.

mechanism. Specifically, each DNPU consists of four consecutive sub-modules:

Graph Construction: The first sub-module includes a graph construction layer that is responsible for creating a graph from the input nodes (i.e. points in the cloud) by connecting each node to its K nearest neighbors. This dynamic construction guarantees the information flow at different levels of hierarchy by preventing node isolation in the graph. This step also adds a self-loop for each node, which will be used for information balancing between the node and its neighbors in the message-passing process.

Position Encoding: The second sub-module is a position encoding layer that encodes the spatial positioning of the neighbors. Graph neural networks are designed to process unstructured data using permutation-invariant operations, regardless of the positional arrangement of the nodes. Even though this is a desirable characteristic for graphs such as social networks, it can drastically affect the interpretation of cellular graphs. For instance, despite the fact that each graph in Fig. 3 has its own unique structure, they are both represented in the same manner by graph convolutional layers such as GCN [22], GAT [43], and SAGE [14]. To overcome this challenge, a positional encoding mechanism is employed to integrate the arrangement of the nodes into the graph representation. In this regard, the positional encoding of a given node can be derived through the computation of its Cartesian coordinate, denoted as p_i, using Eq. 1

(a) Sample Graph 1 (b) Sample Graph 2

Figure 3: Different positioning of nodes in graphs. Even though these two graphs have different node positioning, they are represented in the same way by the conventional graph neural networks. Using a relative positional encoding, CO-PILOT is able to address this issue.

$$\mathcal{P}_i = \sum_{j \in \mathcal{N}_i} \alpha_j(\{p_j, |p_j - p_i|, \frac{p_j - p_i}{|p_j - p_i|}\} W_{pos} + b_{pos}). \tag{1}$$

In this equation, $\mathcal{P}_i$ is the positional encoding of node i,

$|.|$ is the second order norm, $\{.,.\}$ is concatenation, $\mathcal{N}_i$ is the set of nodes connected to node i in the graph, $W_{pos} \in \mathbb{R}^{5\times d}$ and $b_{pos} \in \mathbb{R}^{1\times d}$ are learnable matrices, d is the embedding dimension of node features, and α_j is the normalized attention weight obtained by a linear layer. This structure of the position encoding enables us to attend to both the density and geometry of the neighborhood.

In contrast to previous works [18, 16], we simply add the positional encoding of the node to its feature. This approach helps the model to reduce the number of parameters compared to a feature concatenation process while maintaining a comparable performance (see Sec. 4.6). More specifically, given a node feature of h'^{l}_{i} for node i at layer l, the positional encoding aggregation follows Eq. 2

$$h^{l}_{i} = h'^{l}_{i} + \mathcal{P}_i. \tag{2}$$

It is worth mentioning that despite the significant role played by positional encoding in capturing the spatial localization of cells, it is imperative that the model remains impartial towards the absolute value of the position. To ensure this, we have incorporated a random rotation augmentation that uniformly rotates all the nodes in each graph with the same angle of rotation and around the center of $(0, 0)$.

We have integrated the positional encoding only in the first DNPU layer of the model, as the introduction of extra positional encoding in subsequent layers exhibited similar performance (see Sec. 4.6).

Conditional Neighborhood Aggregation (CNA): The contextual composition of the neighborhood, particularly the specific types of cells present, is a crucial determinant of tissue characteristics. For instance, the presence of lymphocytes in close proximity to tumor cells has been correlated with better outcomes in ovarian cancer [33]. While graph neural networks can integrate this information via message-passing operations, they work on an unconditional likelihood basis relative to the current node. This approach can result in the over-smoothing of the representation and lead to inferior performance. To surmount this limitation, we have designed a novel graph convolutional layer, called Conditional Neighborhood Aggregations (CNA), which aggregates the representations of neighboring nodes given that of the current node. The formulation of CNA is articulated by Eq. 3

$$h^{l}_{i} = \sum_{j\in\mathcal{N}_i} softmax(\langle |h^{l-1}_{i} W^{l}_{1}|, |h^{l-1}_{j} W^{l}_{1}| \rangle) h^{l-1}_{j} W^{l}_{2}, \tag{3}$$

where h^{l}_{i} is the representation of node i at layer l, $\mathcal{N}_j$ is the collection of nodes connected to the node i in the graph at layer l, W^{l}_{1} and $W^{l}_{2} \in \mathbb{R}^{d\times d}$ are learnable matrices, $|.|$ is the second order norm, and $\langle . \rangle$ is the inner product. In contrast to the conventional GNN layers such as GAT [43], SAGE [14], and GIN [48], CNA offers the required flexibility for the model to propagate the information depending on the composition of the environment while maintaining computational efficiency (see Sec. 4.6).

Node Pooling: Cellular graphs typically include large numbers of nodes which can lead to computational issues in deep models. To make the computations manageable on large graphs, a hierarchical SAGE pooling operation [23] is performed to select the most important nodes at each level.

3.2.2 Waterfall Attention

When it comes to downstream tasks, the processed representations in the final layer of the network are generally utilized. This leads to using high-level representations of the image as a result of the bottom-up information propagation of the model. However, we hypothesize that the lower-level features could potentially reveal more valuable information if they were conditioned on the high-level features. This is also supported by clinical studies, such as Huang *et al.*'s work [19], which shows that cellular statistics (such as the proportion of different cell types) can be utilized to predict the therapeutic response, especially when they are classified based on their corresponding tissue types. To put our hypothesis to the test, we created a top-down attention aggregation module, called Waterfall Attention, which is illustrated in Fig. 2. More specifically, Waterfall Attention adheres to Eq. 4

$$\begin{aligned} g^{l} &= \sum_{i\in\mathcal{G}^{l}} softmax(c^{l} MLP^{l}(h^{l}_{i})) h^{l}_{i}, \\ c^{l} &= \sigma(g^{l+1} W^{l}_{wf} + b^{l}_{wf}). \end{aligned} \tag{4}$$

In this equation, $g^{l} \in \mathbb{R}^{1\times d}$ is the aggregated representation of the graph at layer l, $\mathcal{G}^{l}$ is the collection of nodes in the graph at layer l, h^{l}_{i} is the representation of node i at layer l, $W^{l}_{wf} \in \mathbb{R}^{d\times 1}$ and $b_{wf} \in \mathbb{R}$ are learnable parameters, and MLP^{l} is a two-layer MLP with hidden size of d and output size of 1. To calculate g^{L-1}, where L is the total number of DNPU layers, we set the c^{L-1}_{i} to $\mathbf{1}$ which is equivalent to an unconditional aggregation at the last layer. Finally, the aggregated g^{0} (equivalent to $R_{n,k}$ mentioned in Sec. 3.2) is fed into an instance attention to combine all of the image representations for the patient.

3.2.3 Instance Attention & Loss Function

In order to combine embeddings from different images belonging to the same patient, an instance attention is used similar to Eq. 5

$$R_n = \sum_{k=0}^{K(n)} softmax(R_{n,k} W_{inst}) R_{n,k} + b_{inst}, \tag{5}$$

Type	Method	Parameters	HGSOC 1				HGSOC 2				TCGA-OV			
			C-Index (↑)	RMST (↑)	SRD@10 (↑)	P-value (↓)	C-Index (↑)	RMST (↑)	SRD@10 (↑)	P-value (↓)	C-Index (↑)	RMST (↑)	SRD@10 (↑)	P-value (↓)
Patch-Based	DeepSet [52]	$395K$	0.475 ± 0.020	1.06	-4.6%	0.52	0.505 ± 0.019	1.29	$+0.7\%$	0.04	0.500 ± 0.000	1.0	0%	1.0
	Attention MIL [20]	$657K$	0.541 ± 0.040	1.88	$+11.9\%$	0.05	0.535 ± 0.025	1.52	$+3.4\%$	< 0.01	0.532 ± 0.078	0.80	$+11.0\%$	0.04
	Variance MIL [5]	$789K$	0.539 ± 0.051	0.86	$+3.5\%$	0.51	0.518 ± 0.017	1.49	$+6.9\%$	< 0.01	0.538 ± 0.072	0.64	$+11.4\%$	0.04
	DGC [54]	$658K$	0.502 ± 0.074	1.55	$+3.7\%$	0.41	0.524 ± 0.017	1.27	$+0.7\%$	0.05	0.538 ± 0.061	0.48	$+8.8\%$	0.10
	Patch-GCN [7]	$1.3M$	0.522 ± 0.086	0.75	-4.4%	0.93	0.531 ± 0.017	1.49	$+6.1\%$	< 0.01	0.517 ± 0.043	0.35	$+11.5\%$	0.03
	HIPT [6]	$24M$	0.477 ± 0.034	0.66	$+1.6\%$	0.51	0.486 ± 0.007	0.95	$+3.3\%$	0.45	0.486 ± 0.027	1.16	$+8.8\%$	0.96
Point Cloud-Based	PointNet [32]	$710K$	0.513 ± 0.033	1.18	$+12.2\%$	0.16	0.500 ± 0.012	0.77	-8.2%	< 0.01	0.500 ± 0.012	0.77	-8.2%	< 0.01
	DGCNN [44]	$1.9M$	0.519 ± 0.027	0.73	-3.9%	0.64	0.506 ± 0.022	1.11	$+5.3\%$	0.15	0.506 ± 0.022	1.11	$+5.3\%$	0.15
	DPCC [16]	$17M$	0.505 ± 0.040	0.67	-6.5%	0.36	0.514 ± 0.025	1.14	$+5.7\%$	0.21	0.514 ± 0.025	1.14	$+5.7\%$	0.21
Graph-Based	HGSurvNet [10]	$431K$	0.513 ± 0.043	0.74	-3.1%	0.71	0.450 ± 0.164	0.59	$+3.4\%$	0.38	0.520 ± 0.069	0.75	$+6.1\%$	0.23
	CO-PILOT (Ours)	$356K$	$\mathbf{0.568 \pm 0.027}$	**2.29**	$\mathbf{+12.6\%}$	$\mathbf{< 0.01}$	$\mathbf{0.558 \pm 0.033}$	**1.61**	$\mathbf{+10.2\%}$	$\mathbf{< 0.01}$	$\mathbf{0.557 \pm 0.040}$	**1.44**	$\mathbf{+13.8\%}$	**0.03**

Table 1: Survival prediction performance comparison of our model with all the baselines on two datasets.

where $R_{n,k}$ is the corresponding representation of $x_{n,k}$ obtained from the Waterfall attention, while $W_{inst} \in \mathbb{R}_{d\times 1}$ and $b_{inst} \in \mathbb{R}$ are learnable parameters. Finally, R_n is passed through a fully connected layer to generate the hazard, and the network is trained using the Negative Log Likelihood (NLL) loss [51].

4. Experiments

4.1. Data

We utilized two tissue microarray (TMA) datasets from high-grade serous ovarian cancer: HGSOC 1 represents 873 tissue samples from 188 patients, and HGSOC 2 contains $1,348$ samples from 684 patients. All the tissue samples were stained with H&E and scanned to generate $4,000 \times 4,000$-pixel images at 40x magnification. Each patient has multiple TMA cores, and the latest survival status (alive or dead) along with the overall survival time (since diagnosis) is available for all of the patients. We also used the publicly available ovarian TCGA as part of our experiments. Even though the pre-processing of TCGA-OV was the same as that of the other datasets, we used

4.2. Preprocessing Steps

We first obtained an instance segmentation mask for each of the images using HoVer-Net [12]. The embedding representation of each cell was extracted by applying ResNet34 (pre-trained on ImageNet) on a 72×72 pixels window around that cell. These embeddings served as the node features in the cellular graph (point cloud). In addition, we normalized the cell coordinates within each image to a range of 0 to 1 and used them for positional encoding.

4.3. Implementation Details

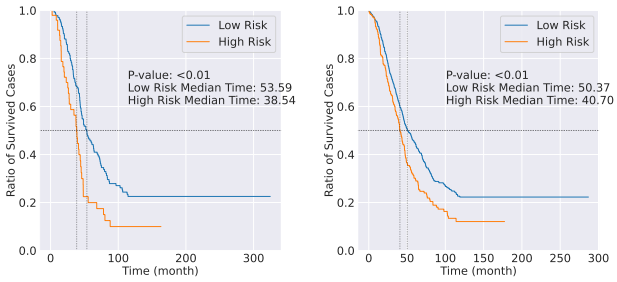
The Pytorch and DGL python libraries were utilized to develop the model, and all experiments were conducted on RTX 3090 GPUs with 3 distinct initialization seeds. During training, the adam optimizer was used in conjunction with a cosine scheduler and a weight decay of 0.00001, and a batch size of 64. The learning rate was set to 0.002 for the HGSOC 1 dataset, 0.01 for the HGSOC 2 dataset, and 0.02 for the TCGA-OV. The hidden dimension of all CNA modules was set to 128. Additionally, K was set to 6, and a 0.4 ratio was utilized in the SAGE pooling layers. However, we reduced the ratio to 0.3 for the TCGA-OV to accommodate the computations.

4.4. Survival Outcome Prediction

The results of the survival prediction are available in Tab. 1. Our model's performance was compared to that of the baselines using four metrics. The C-index (ranging from 0 to 1) measures the consistency between predicted and actual survival times across all patients. A value of 1 indicates complete agreement, while a value of 0 indicates complete disagreement. Although the C-index is commonly used to compare machine learning models, patient stratification into risk groups is more important for clinical decision-making. Therefore, the p-value of the log-rank test was used as the main metric for clinical utility (see Sec. 4.5). We also utilized the restricted mean survival time (RMST) [15] as the third evaluation metric. This metric quantifies the ratio of the area under the survival curves of the low- and high-risk cohorts. Additionally, we employed the survival rate difference (SRD) as the fourth measure. This metric was calculated by subtracting the 10-year survival rate of the high-risk cohort from that of the low-risk. Larger values of SRD indicate that the identified low-risk group has a higher survival rate compared to the high-risk patients in long term.

We conducted all experiments using a 3-fold cross-validation approach and ran each fold using 3 seeds to account for initialization variability, resulting in a total of 9 runs for each experiment. The results of our experiments confirm that CO-PILOT is able to outperform all of the baselines, including Vision Transformer, and has consistent performance improvement across both datasets and all four metrics, unlike the baselines. In particular, our model achieves a c-index of 0.568, 0.558, and 0.557 on the HGSOC 1, HGSOC 2, and TCGA-OV datasets while those of the closest baselines are 0.541, 0.535, and 0.544. Our results are also consistent with previous studies, where Nakhli *et al*. [29] demonstrate a c-index of 0.550 on ovarian cancer patients using immunohistochemistry images.

CO-PILOT's results indicate that the cellular graph alone



(a) HGSOC 1 (b) HGSOC 2

Figure 4: Survival curves for cohorts of patients identified as low-risk (predicted hazard $<$ hazard threshold) and high-risk (prediction hazard $>$ hazard threshold) by our model. Our model can predict patients who have significantly higher risks and shorter survival times.

captures enough information to surpass the patch-based methods, implying the model's potential to deduce the micro-environment structure from the cellular graph. Furthermore, our model reduces the number of parameters by half compared to the closest baseline, which can be linked to its cellular foundation. More results and visualizations can be found in the supplementary material.

Additionally, our model can separate the low- and high-risk patients significantly on both datasets, being the only one to do so on the HGSOC 1. Our study is the first to show the possibility of this separation on the high-grade serous ovarian patients using only the routine H&E images.

4.5. Patient Risk Stratification

C-index is a common metric for comparison between machine learning models in the survival prediction domain [6, 8]. However, from the clinical utility perspective, it is more crucial for the model to stratify patients into groups with statistically significant differences in outcome. Such patient groups can then be managed differently in the clinic. For example, low-risk patients may be managed with less aggressive treatments (e.g., radiotherapy) while high-risk patients may be treated with more aggressive regimens (e.g., systemic chemotherapy, combination therapy). This evaluation is usually achieved through the use of Kaplan-Meier (KM) curves and log-rank test p-values. To demonstrate the potential of our model in this regard, we segregated patients into low- and high-risk groups based on the predicted hazards by the model for all patients. Fig. 4 presents the KM plots of our model across both datasets along with the corresponding log-rank test p-value. As shown, our model effectively stratifies patients into low- and high-risk cohorts, with high-risk patients exhibiting statistically significant shorter survival times. Specifically, the median survival times of high-risk patients in the HGSOC 1 and HGSOC 2 were 38.5 and 40.7 months, respectively, while the median times for low-risk cohorts were 53.6 and 50.4 months.

Our datasets represent a highly aggressive subtype of ovarian cancer patients (high-grade serous) that are uniformly treated. To the best of our knowledge, there are no clinical parameters that can separate the low- and high-risk patients in this subtype, and despite numerous attempts over the past few decades, most research has been unsuccessful in identifying biomarkers that indicate response to treatment in these patients. Studies conducted by Wang *et al*. [46] and Talhouk *et al*. [38] found that such markers could be identified through global genomics and transcriptomics aberration profiles, whereas our study is the first to yield promising results based on routine H&E histopathology images.

4.6. Ablation Studies

We validated our design choices through extensive ablation experiments (Tab. 2). Our results confirm the importance of the low-, mid-, and high-level blocks as their ablation (rows 1 & 2) resulted in reduced c-index and insignificant separation of the patient cohorts. We also observed that our CNA layer outperformed the state-of-the-art GAT (row 3), SAGE (row 4), and GIN (row 5) layers, with better efficiency in terms of memory consumption (CNA requires only 1 GPU while the others require 2). The elimination of conditional aggregation of the neighboring nodes in our CNA layer resulted in a performance drop similar to that of the SAGE layer, while having better patient stratification capability (row 6). This performance drop illustrates the importance of conditional aggregation for micro-environment representation in cellular graphs.

The conditional top-down aggregation using our Waterfall attention proved crucial, as its ablation (row 7) led to the failure of patient stratification and a reduction in other metrics. The replacement of the waterfall attention with averaging operation also significantly affected the model's performance (row 8), reinforcing our hypothesis regarding the key role of hierarchical top-down dependency in histopathology applications.

Our experiments on positional encoding concatenation (row 9) validated our previous hypothesis concerning the over-parametrization of the network and the challenges of its dynamics during training. Additionally, we observed that rotational augmentations of node positions could significantly impact both the c-index and the stratification p-values, consistent with our earlier assumption (row 10). Conversely, we found that including extra positional encoding in all DNPU layers (row 11) did not lead to any significant differences, indicating the model's robustness in conveying positional information throughout.

Row	Ablated Feature	HGSOC 1				HGSOC 2			
		C-Index (↑)	RMST (↑)	SRD@10 (↑)	P-value (↓)	C-Index (↑)	RMST (↑)	SRD@10 (↑)	P-value (↓)
1	Weight sharing of all DNPUs	0.559 ± 0.032	1.56	$+3.8\%$	0.27	0.544 ± 0.029	1.08	-0.1%	0.97
2	No DNPU weight sharing	0.564 ± 0.037	1.71	$+5.9\%$	0.07	0.546 ± 0.026	0.98	$+2.9\%$	0.47
3	Replacing CNA with GAT	0.530 ± 0.036	1.14	-6.7%	0.39	0.532 ± 0.031	0.97	$+1.4\%$	0.88
4	Replacing CNA with SAGE	0.564 ± 0.030	1.41	-2.0%	0.53	0.548 ± 0.030	1.06	$+3.3\%$	0.23
5	Replacing CNA with GIN	0.552 ± 0.022	1.69	$+7.3\%$	0.01	0.520 ± 0.033	1.12	-0.8%	0.36
6	Unconditional CNA	0.567 ± 0.033	1.75	$+6.0\%$	0.05	0.547 ± 0.030	1.43	$+8.0\%$	0.04
7	Unconditional Waterfall attention	0.551 ± 0.030	1.27	-5.6%	0.96	0.548 ± 0.034	0.98	$+1.6\%$	0.94
8	No waterfall attention	0.565 ± 0.017	1.53	$+1.8\%$	0.27	0.546 ± 0.037	0.95	-1.4%	0.85
9	Position encoding concatenation	0.564 ± 0.038	0.66	-9.0%	0.33	0.548 ± 0.033	1.42	$+4.4\%$	0.02
10	No rotation augmentation	0.523 ± 0.056	1.12	-12.3%	0.57	0.558 ± 0.039	1.42	$+7.0\%$	0.04
11	Position encoding in all DNPUs	0.568 ± 0.027	2.29	$+12.6\%$	< 0.01	0.558 ± 0.032	1.61	$+10.2\%$	< 0.01
	CO-PILOT (Ours)	$\mathbf{0.568 \pm 0.027}$	**2.29**	$\mathbf{+12.6\%}$	$\mathbf{< 0.01}$	$\mathbf{0.558 \pm 0.033}$	**1.61**	$\mathbf{+10.2\%}$	$\mathbf{< 0.01}$

Table 2: Ablation studies of our model on both datasets.

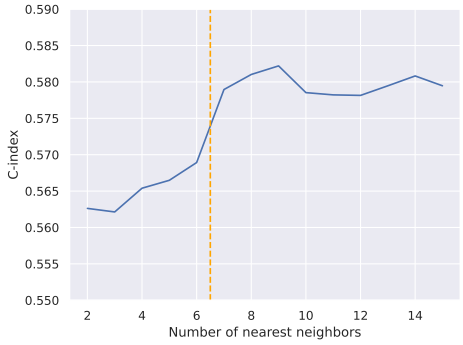


Figure 5: Performance of the model with respect to the number of nearest neighbors used in the graph construction. As the number of nearest neighbors increases, the c-index rises as well. This is mainly because the model receives more information about the neighborhood to perform the prediction. On the other hand, a number larger than $K = 9$ results in a performance drop, which can be attributed to the over-smoothing problem. The experiments on the right side of the vertical line require 2 GPUs.

4.7. Impact of the Number of Cell Neighbors (K)

We analyzed how the performance of the model was affected by the number of neighbors, which is the only hyperparameter of our proposed model. The results are shown in Fig. 5, where we present the model's performance on the HGSOC 1 dataset as a function of the number of nearest neighbors in the dynamic graph construction. As expected, the model's performance increases as the number of nearest neighbors grows, since it can capture more information about each cell's neighborhood. However, after a certain point ($k = 9$), a drop in the performance can be seen. We attribute this to over-smoothing issues and improper information processing caused by an excessively large neighborhood in the graph neural network.

5. Conclusion

This study, for the first time, introduces a dynamic point-cloud approach for the processing of cellular graphs extracted from histopathology images. Our model (CO-PILOT) utilizes a bottom-up conditional neighborhood aggregation for information propagation and a hierarchical top-down aggregation to combine representations at different scales. CO-PILOT outperforms the existing patch-based survival prediction methods, including the Hierarchical Vision Transformer [6], demonstrating the potential of the model in deducing tissue environment from the cellular graphs.

Most importantly, our model stratifies high-grade serous ovarian cancer patients into low- and high-risk cohorts using routine H&E images across two different datasets, a task that previously required global genomic and transcriptomics profiles or immunohistochemistry images, neither of which is routinely used in practice. This underscores the importance of cellular heterogeneity, spatial positioning, and mutual interactions in histopathology image representation learning.

Our work offers promising new pathways for efficient cell-based processing of histopathology images, potentially leading to new applications in massive archives of histopathology images already available in clinics. Integrating our model with additional clinico-pathological variables, such as stage and age, may enable us to establish connections between histopathology images and genomic mutational profiles, thereby facilitating in-depth analyses and biological inquiries. Even though we have confined our experiments to histopathology data only, the application of our model in other imaging domains such as fluorescent images can be interesting as well. Furthermore, our model's emphasis on cellular-based methodologies opens doors for identifying visually understandable biological components that have a significant impact on predicting outcomes, and which may prove beneficial for clinical and deep biological interrogations.

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] 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. 2

[3] Wieland Brendel and Matthias Bethge. Approximating cnns with bag-of-local-features models works surprisingly well on imagenet. *arXiv preprint arXiv:1904.00760*, 2019. 2

[4] 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. 2

[5] Iain Carmichael, Andrew H Song, Richard J Chen, Drew FK Williamson, Tiffany Y Chen, and Faisal Mahmood. Incorporating intratumoral heterogeneity into weakly-supervised deep learning models via variance pooling. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 387–397. Springer, 2022. 6

[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, 6, 7, 8

[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 *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VIII 24*, 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, 7

[9] Jaesung Choe, Chunghyun Park, Francois Rameau, Jaesik Park, and In So Kweon. Pointmixer: Mlp-mixer for point cloud understanding. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVII*, pages 620–640. Springer, 2022. 3

[10] Donglin Di, Changqing Zou, Yifan Feng, Haiyan Zhou, Rongrong Ji, Qionghai Dai, and Yue Gao. Generating hypergraph-based high-order representations of whole-slide histopathological images for survival prediction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(5):5800–5815, 2022. 6

[11] Songlin Fan, Wei Gao, and Ge Li. Salient object detection for point clouds. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVIII*, pages 1–19. Springer, 2022. 3

[12] Simon Graham, Quoc Dang Vu, Shan E Ahmed Raza, Ayesha Azam, Yee Wah Tsang, Jin Tae Kwak, and Nasir Rajpoot. Hover-net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images. *Medical Image Analysis*, page 101563, 2019. 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, 5

[15] Kyunghwa Han and Inkyung Jung. Restricted mean survival time for survival analysis: a quick guide for clinical researchers. *Korean Journal of Radiology*, 23(5):495, 2022. 6

[16] Yun He, Xinlin Ren, Danhang Tang, Yinda Zhang, Xiangyang Xue, and Yanwei Fu. Density-preserving deep point cloud compression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2333–2342, 2022. 3, 5, 6

[17] 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. 2

[18] Qingyong Hu, Bo Yang, Linhai Xie, Stefano Rosa, Yulan Guo, Zhihua Wang, Niki Trigoni, and Andrew Markham. Randla-net: Efficient semantic segmentation of large-scale point clouds. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11108–11117, 2020. 5

[19] Zhi Huang, Wei Shao, Zhi Han, Ahmad Mahmoud Alkashash, Carlo De la Sancha, Anil V Parwani, Hiroaki Nitta, Yanjun Hou, Tongxin Wang, Paul Salama, et al. Artificial intelligence reveals features associated with breast cancer neoadjuvant chemotherapy responses from multi-stain histopathologic images. *NPJ Precision Oncology*, 7(1):14, 2023. 5

[20] Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *International conference on machine learning*, pages 2127–2136. PMLR, 2018. 2, 6

[21] 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. 2

[22] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016. 4

[23] Junhyun Lee, Inyeop Lee, and Jaewoo Kang. Self-attention graph pooling. In *International conference on machine learning*, pages 3734–3743. PMLR, 2019. 5

[24] 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. 2

[25] 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. 2

[26] Yangyan Li, Rui Bu, Mingchao Sun, Wei Wu, Xinhan Di, and Baoquan Chen. Pointcnn: Convolution on x-transformed points. *Advances in neural information processing systems*, 31, 2018. 3

[27] 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

[28] 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

[29] Ramin Nakhli, Puria Azadi Moghadam, Haoyang Mi, Hossein Farahani, Alexander Baras, Blake Gilks, and Ali Bashashati. Sparse multi-modal graph transformer with shared-context processing for representation learning of giga-pixel images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11547–11557, 2023. 1, 6

[30] 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

[31] Yatian Pang, Wenxiao Wang, Francis EH Tay, Wei Liu, Yonghong Tian, and Li Yuan. Masked autoencoders for point cloud self-supervised learning. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part II*, pages 604–621. Springer, 2022. 3

[32] C Qi, H Su, K Mo, and LJ Guibas. Pointnet: deep learning on point sets for3d classification and segmentation. 2017. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 77–85, 2016. 6

[33] Phillip P Santoiemma and Daniel J Powell Jr. Tumor infiltrating lymphocytes in ovarian cancer. *Cancer biology & therapy*, 16(6):807–820, 2015. 5

[34] Korsuk Sirinukunwattana, David Snead, David Epstein, Zia Aftab, Imaad Mujeeb, Yee Wah Tsang, Ian Cree, and Nasir Rajpoot. Novel digital signatures of tissue phenotypes for predicting distant metastasis in colorectal cancer. *Scientific reports*, 8(1):1–13, 2018. 2

[35] 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

[36] 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

[37] 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

[38] Aline Talhouk, Joshy George, Chen Wang, Timothy Budden, Tuan Zea Tan, Derek S Chiu, Stefan Kommoss, Huei San Leong, Stephanie Chen, Maria P Intermaggio, et al. Development and validation of the gene expression predictor of high-grade serous ovarian carcinoma molecular subtype (protype). *Clinical Cancer Research*, 26(20):5411–5423, 2020. 7

[39] 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

[40] 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

[41] Kevin Thandiackal, Boqi Chen, Pushpak Pati, Guillaume Jaume, Drew FK Williamson, Maria Gabrani, and Orcun Goksel. Differentiable zooming for multiple instance learning on whole-slide images. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXI*, pages 699–715. Springer, 2022. 1, 2

[42] Hugues Thomas, Charles R Qi, Jean-Emmanuel Deschaud, Beatriz Marcotegui, François Goulette, and Leonidas J Guibas. Kpconv: Flexible and deformable convolution for point clouds. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6411–6420, 2019. 3

[43] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017. 4, 5

[44] Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E Sarma, Michael M Bronstein, and Justin M Solomon. Dynamic graph cnn for learning on point clouds. *ACM Transactions on Graphics (tog)*, 38(5):1–12, 2019. 3, 6

[45] 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. 3

[46] 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. 7

[47] 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

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

[49] Jiawei Yang, Hanbo Chen, Yuan Liang, Junzhou Huang, Lei He, and Jianhua Yao. Concl: Concept contrastive learning for dense prediction pre-training in pathology images. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXI*, pages 523–539. Springer, 2022. 1

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

[51] Shekoufeh Gorgi Zadeh and Matthias Schmid. Bias in cross-entropy-based training of deep survival networks. *IEEE transactions on pattern analysis and machine intelligence*, 43(9):3126–3137, 2020. 6

[52] 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. 2, 6

[53] 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. 2

[54] Yu Zhao, Fan Yang, Yuqi Fang, Hailing Liu, Niyun Zhou, Jun Zhang, Jiarui Sun, Sen Yang, Bjoern Menze, Xinjuan Fan, et al. Predicting lymph node metastasis using histopathological images based on multiple instance learning with deep graph convolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4837–4846, 2020. 6

[55] 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