GRASP: GRAPH-STRUCTURED PYRAMIDAL WHOLE SLIDE IMAGE REPRESENTATION

Ali Khajegili Mirabadi[1]✉, Graham Archibald[1], Amirali Darbandsari[1], Alberto Contreras-Sanz[1,2], Ramin Ebrahim Nakhli[1], Maryam Asadi[1], Allen Zhang[1], C. Blake Gilks[2], Peter Black[2], Gang Wang[3], Hossein Farahani[1], Ali Bashashati[1]✉
[1]School of Biomedical Engineering & Department of Pathology and Laboratory Medicine
[2]Vancouver Prostate Centre, [3]BC Cancer Institute
The University of British Columbia

ABSTRACT

Cancer subtyping is one of the most challenging tasks in digital pathology, where Multiple Instance Learning (MIL) by processing gigapixel whole slide images (WSIs) has been in the spotlight of recent research. However, MIL approaches do not take advantage of inter- and intra-magnification information contained in WSIs. In this work, we present GRASP, a novel lightweight graph-structured multi-magnification framework for processing WSIs in digital pathology. Our approach is designed to dynamically emulate the pathologist's behavior in handling WSIs and benefits from the hierarchical structure of WSIs. GRASP, which introduces a convergence-based node aggregation mechanism replacing traditional pooling mechanisms, outperforms state-of-the-art methods by a high margin in terms of balanced accuracy, while being significantly smaller than the closest-performing state-of-the-art models in terms of the number of parameters. Our results show that GRASP is dynamic in finding and consulting with different magnifications for subtyping cancers, is reliable and stable across different hyperparameters, and can generalize when using features from different backbones. The model's behavior has been evaluated by two expert pathologists confirming the interpretability of the model's dynamic. We also provide a theoretical foundation, along with empirical evidence, for our work, explaining how GRASP interacts with different magnifications and nodes in the graph to make predictions. We believe that the strong characteristics yet simple structure of GRASP will encourage the development of interpretable, structure-based designs for WSI representation in digital pathology. Data and code can be found in `https://github.com/AIMLab-UBC/GRASP`

1 INTRODUCTION

Though deep learning has revolutionized computer vision in many fields, digital pathology tasks such as cancer classification remain a complex problem in the domain. For natural images, the task usually relates to assigning a label to an image with an approximate size of 256×256 pixels, with the label being clearly visible and well-represented in the image. Gigapixel tissue whole-slide images (WSIs) break this assumption in digital pathology as images exhibit enormous heterogeneity and can be as large as $150{,}000 \times 150{,}000$ pixels. Further, labels are provided at the slide level and may be descriptive of a small region of pixels occupying a minuscule portion of the total image, or they may be descriptive of complex interactions between the substructures within the entire composition of the WSI Ehteshami Bejnordi et al. (2017); Zhang et al. (2015); Pawlowski et al. (2019).

Multiple Instance Learning (MIL) has become the prominent approach to address the computational complexity of WSI; however, the majority of methods in the literature focus only on a single level of magnification, usually $20\times$, due to the computational cost of including other magnifications Lu et al. (2021); Ilse et al. (2018); Schirris et al. (2022); Zheng et al. (2021); Chen et al. (2021); Shao et al. (2021); Zhou et al. (2019); Guan et al. (2022). Using this magnification, a set of patches from each WSI are extracted and used as an instance-level representation. This neither captures the biological structure of the data nor does it follow the diagnostic protocols of pathologists. That is to say, WSIs at higher magnifications reveal finer details—such as the structure of the cell nucleus and the intra/extracellular matrix—whereas lower magnifications enable the identification of larger

structures like blood vessels, connective tissue, or muscle fibers. Further, these structures are inconsistent from patient to patient, slide to slide, and subtype to subtype Morovic et al. (2021). To capture this variability, pathologists generally use a variety of lenses in their inspection of a tissue sample under the microscope, switching between different magnifications as needed. They generally begin with low magnifications to identify regions of interest for making preliminary decisions before increasing magnifications to confirm or rule out diagnoses Rasoolijaberi et al. (2022).

To address this challenge, several multi-magnification approaches have recently been introduced for various tasks such as cancer subtyping, survival analysis, and image retrieval. However, these models often possess millions of parameters, as briefly illustrated in Figure 1, and suffer from interpretability issues due to their modular complexity Thandiackal et al. (2022); Li et al. (2021); Riasatian et al. (2021); Chen et al. (2022b); D'Alfonso et al. (2021); Hashimoto et al. (2020). Although these models have demonstrated promise across different tasks, they are not well-suited for low-resource clinical settings, where computational resources are often limited and the infrastructure may not support large-scale computational clusters. Therefore, there is a critical need to develop approaches and models specifically designed for use in smaller clinics, where the hardware may consist of small GPUs with limited memory. These lightweight models must balance accuracy with efficiency, enabling reliable deployment on standard devices while ensuring real-time performance and ease of integration within existing clinical workflows.

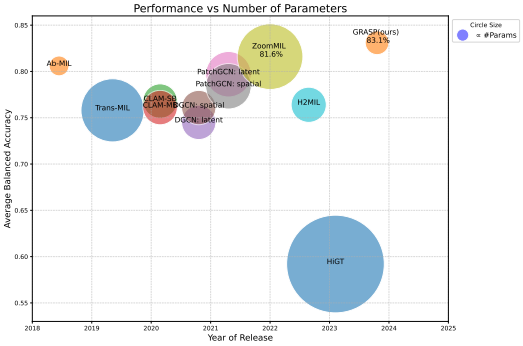

Figure 1: A chronological overview of different WSI representation methods and their performance compared to the size of the model.

In this research, we aim to further the progress of deep learning in this context by introducing a pre-defined fixed structure for a lightweight model that reduces the complexity while maintaining efficacy and interoperability. Therefore, our contribution is as follows:

1. Introducing GRASP to capture pyramidal information contained in WSIs, as the first *lightweight* multi-magnification model in computational pathology.
2. GRASP introduces a novel convergence-based mechanism instead of traditional pooling layers to capture intra-magnification information.
3. We provide a solid theoretical foundation of the model's functionality and its interpretability from both technical and pathological perspectives, as well as providing empirical evidence for the model's efficacy concerning hyperparameters.
4. An extensive comparison with eleven state-of-the-art models across three different cancers, ranging from two to five histotypes, using two popular backbones demonstrates the generalizability of the proposed method.

2 Related Work

2.1 Patch-Level Encoding

With recent progress in deep learning, deep features, i.e. high-level embeddings from a deep network, have advanced past handcrafted features and are considered the most robust sources for image representation. Pre-trained networks such as DenseNet Huang et al. (2017), ResNet He et al. (2016), or Swin Liu et al. (2021) draw their features from millions of non-medical and non-histopathological images, where they cannot necessarily produce high-level embeddings for complex images, especially rare cancers Wang et al. (2023); Riasatian et al. (2021); Ciga et al. (2022); Wang et al. (2022). In this context, the use of Variational Autoencoders (VAEs) has been evaluated in Chen et al. (2022a), where the authors show that DenseNet pre-trained on ImageNet performs better for extracting semantic features from WSIs than VAEs. However, domain-specific vision encoders such as KimiaNet Riasatian et al. (2021), CTransPath Wang et al. (2022), PLIP Huang et al. (2023), UNI

Chen et al. (2023), Virchow Vorontsov et al. (2023), etc. were developed and trained on large sets of histopathology images (patches) outperforming pre-trained models on ImageNet across various tasks.

2.2 Weak Supervision in Gigapixel WSIs

MIL Approaches: Several domains of deep learning have been explored in an attempt to effectively address the task of classification in digital pathology. Models such as AB-MIL Ilse et al. (2018), CLAM Lu et al. (2021), and Trans-MIL Shao et al. (2021) have utilized MIL with promising results. Such approaches have generally focused only on instance-level feature extraction and have not yet explored modeling global, long-range interactions within and across different magnifications. In Waqas et al. (2024), a detailed overview of different MIL methods has been provided.
Graph-based Approaches: To incorporate contextual information and long-range interactions, models such as PatchGCN Chen et al. (2021) and DGCN Zheng et al. (2021) have been designed with a graph structure that can capture and learn context-aware features from interactions across the WSI. These models represent WSIs as graphs where the nodes are usually embeddings and edges are defined based on clustering or neighborhood node similarity, which in turn adds new hyperparameters and increases inference time. The similarity between nodes can be measured in terms of spatial or latent space, leading to the construction of different graphs for each WSI.
Multi-Magnification Approaches: Multiple efforts have been made to incorporate multi-magnification information in the context of gigapixel histopathology subtyping tasks. Models such as HiGT Guo et al. (2023), ZoomMIL Thandiackal et al. (2022), CSMIL Deng et al. (2023), H^2-MIL Hou et al. (2022), and DSMIL Li et al. (2021) address this by aggregating contextual tissue information using features from multiple magnifications in WSIs. DSMIL concatenates embeddings from different magnifications together by duplicating lower-magnification features, making the model biased toward lower magnifications and unable to look into inter-magnification information. On the other hand, ZoomMIL aggregates information from $5x$ to $20x$ in a fixed hierarchy with no interaction in the opposite direction. Chen et. al explore this in the context of vision transformers with their Hierarchical Image Pyramid Transformer (HIPT) Chen et al. (2022b). Their architecture incorporates regions of size 256×256 and 4096×4096 pixels to leverage the natural hierarchical structure of WSIs. H^2-MIL also adopts a graph-based approach, where it pools the nodes in each magnification using an Iterative Hierarchical Pooling module. Our proposed model, on the other hand, is designed to dynamically aggregate information within and across different magnifications without using traditional pooling layers in its intra-magnification interactions. It also employs a similar mechanism to zoom-in and zoom-out through its inter-magnification interactions, from lower to higher magnifications and vice versa.

3 Method

This section introduces the GRAph-Structured Pyramidal (GRASP) WSI Representation, a framework for subtype recognition using multi-magnification weakly-supervised learning, illustrated in Figure 2.

3.1 Problem Formulation

Contrary to Multiple Instance Learning (MIL) approaches, which use a bag of instances to represent a given WSI, GRASP benefits from a graph-based, multi-magnification structure to objectively represent connections between different instances across and within different magnifications. To build a graph and learn a graph-based function $\mathcal{F}$ that predicts slide-level labels with no knowledge of patch labels, the following formulation is adopted.

For a given WSI, $W_r \in \mathbb{R}^{N \times M \times 3}$ with label $\mathcal{Y}$, three sets of m patches, $\{p_i \in \mathbb{R}^{n \times n \times 3} : \forall i \in [1, ..., m]\}$, $\{p'_i \in \mathbb{R}^{n \times n \times 3} : \forall i \in [1, ..., m]\}$, and $\{p''_i \in \mathbb{R}^{n \times n \times 3} : \forall i \in [1, ..., m]\}$ are extracted for each magnification of $\mathbf{M}_1 = 5x$, $\mathbf{M}_2 = 10x$, and $\mathbf{M}_3 = 20x$, respectively. It is important to note that p''_i is the high-resolution window located at the center of p'_i, and p'_i is the high-resolution window located at the center of p_i. These patches provide $3m$ patches in total that are then fed into an encoder ϕ to encode extracted patches into a lower dimension space as follows:

$$\phi : p_i \longrightarrow h_i \in \mathbb{R}^{d \times 1}, \forall i \in [1, ..., m] \tag{1}$$

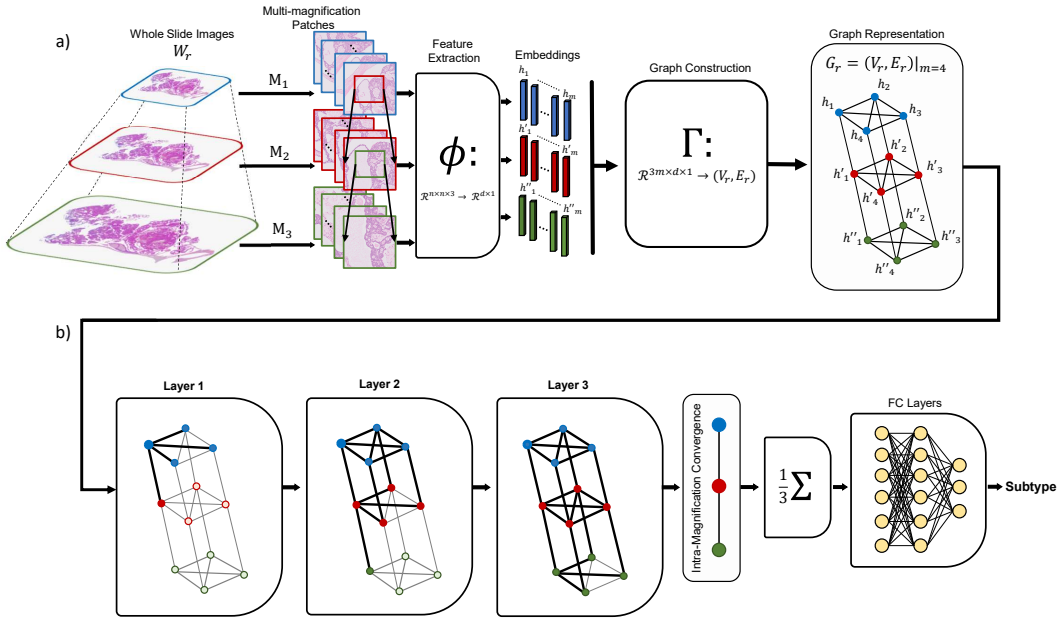

Figure 2: Overview of our workflow beginning with WSIs and outputting slide-level subtype predictions. **a)** shows the WSI being tiled into patches of varying magnification which are then embedded and assembled into a hierarchical graph. In **b)**, graph representations are fed into a three-layer GCN Kipf & Welling (2016) and subsequently, a two-layer MLP to predict graph-level (slide-level) subtypes. As shown in the message passing steps in **b)**, nodes in the first GCN layer interact with their immediate neighbors; those in the second GCN layer can interact with their second neighbors; and nodes in the final GCN layer can interact with all nodes in the graph. Then, the inter-magnification convergence causes the nodes within each magnification to converge, which is an intrinsic property of the architecture. In the end, the three converged nodes are passed through an average readout module. This dynamic helps the model to look for important messages in the entire graph, and if a node contains important information, it will be broadcast to all other nodes in the graph. The output of the GCN layers is then averaged by the *readout module* and passed to the FC layers. (For the sake of illustration, $m = 4$ is used to show the structure of GRASP).

where h_i is the feature vector corresponding to the patch p_i. Correspondingly, h'_i represents p'_i, and so h''_i does p''_i. Using all the feature vectors for each W_r, graph $\mathbb{G}_r$ is constructed using the transformation Γ:

$$\Gamma : \begin{bmatrix} \{h_1, ..., h_m\} \\ \{h'_1, ..., h'_m\} \\ \{h''_1, ..., h''_m\} \end{bmatrix} \in \mathbb{R}^{3m\times d\times 1} \longrightarrow \mathbb{G}_r = (V_r, E_r) \tag{2}$$

Eventually, classifier $\mathcal{C}$ is applied on top of graph convolutional layers $\mathcal{G}$ to build the graph-based function $\mathcal{F}$ to predict slide-level label $\mathcal{Y}$ as follows:

$$\mathcal{Y} = \mathcal{F}(W_r) = \mathcal{C}(\mathcal{G}(V_r, E_r)) \tag{3}$$

3.2 GRASP

We start by extracting multi-magnification patches as described earlier. Then, for any i, we use the same encoder to encode p_i, p'_i, and p''_i into features h_i, h'_i, and h''_i respectively. Having instances features, we use the transformation Γ to build $\mathbb{G}_r$ as introduced in Eq. 2.

The mechanism of connecting every two nodes in $\mathbb{G}_r$ through Γ is premised upon an intuition of the pyramidal nature of WSIs as well as the way in which a conventional light microscope works when one switches from one magnification to another. When using a microscope, increasing magnification preserves the size of the image yet increases resolution by showing the central window of the lower

magnification. This is the exact procedure we use to extract our patches in three magnifications. Therefore, for any i, h_i, h'_i, and h''_i are connected to each other via undirected edges, where this connection represents inter-magnification information contained in the features. On the other hand, for any i, all h_i's contain information in $\mathbf{M}_1$, such that they are connected to each other, forming a fully connected graph at $\mathbf{M}_1$ magnification to represent intra-magnification information. Similarly, all h'_is are connected to each other and also all the h''_is to represent intra-magnification information contained in $\mathbf{M}_2$ and $\mathbf{M}_3$, respectively.

Figure 2 shows a small example of such a graph for $\mathbb{G}_r = (V_r, E_r)|_{m=4}$ where blue, red, and green nodes each form a fully connected graph of size m; the inter-magnification relationship can also be seen via the edges between blue & red nodes as well as the red & green nodes. So far, each WSI, W_r, has been represented by a fixed graph $\mathbb{G}_r$ with $3m$ nodes and $\frac{(3m+1)m}{2}$ edges. These graphs are thus deployed to train the GCNs and predict the label $\mathcal{Y}$ at the output.

3.3 Graph Convolutional Layers

Following Eq. 3, we are defining $\mathcal{G}$ which includes three GCN layers. The intuition behind using three layers is that as a pathologist begins to look for a tumor in a given WSI, they use an initial magnification to find the region of interest; Once found, they consult with other magnifications, which may require zooming in and out back and forth, to confirm their final decision. Therefore, as shown in Figure 2, all nodes in the graph interact with one another in a hierarchical fashion through the GCN layers. Consequently, each node gradually gathers information from all other nodes, and therefore, if there are any important messages carried by some nodes, it is guaranteed to be broadcast to all other nodes, which is the equivalent of zoom-in and zoom-out mechanism. This dynamic and hierarchical structure imposes theoretical properties on the model which is going to be discussed here. Following the graph convolutional layer introduced in Kipf & Welling (2016), the graph nodes are updated as follows:

$$h_i^{(l+1)} = \alpha(b^{(l)} + \sum_{j \in \mathcal{N}(i)} \frac{1}{c_{ji}} h_j^{(l)} W^{(l)}), \tag{4}$$

where $b^{(l)}$ is bias; $h_i^{(l+1)}$ is the node feature's update of the graph at $(l+1)$-th step at $\mathbf{M}_1$; $\mathcal{N}(i)$ is the set of neighbors of node i, $c_{ji} = \sqrt{|\mathcal{N}(j)||\mathcal{N}(i)|}$, where given the symmetry of the graph, all c_{ji}s are equal; and $\alpha(.)$ is the activation function, which is ReLU in our implementation. $h_i^{\prime(l+1)}$ and $h_i^{\prime\prime(l+1)}$ expressions follow the same logic as $h_i^{(l+1)}$ in terms of the parameters mentioned above. After the last graph convolutional layer, where the intra-magnification convergence happens, the graph is passed through an average readout module to pool the three-node graph mean embedding and then is fed into the two-layer classifier $\mathcal{C}$ to predict $\mathcal{Y}$.

3.4 Intra-Magnification Convergence

Based on the idea of capturing the information within and across magnifications, we now show that node features converge in each magnification in the graph to one node Having this, GRASP essentially encodes a graph of size $3m$ nodes to only 3 nodes. We interpret this as the model learning the information contained in the magnification through interaction with other magnifications without a need for traditional pooling layers.

Theorem 1 *Supposing the graph convolutional layers have L_2-bounded weights, and the graph node features at $l = 0$ are L_2-bounded. Therefore, $\forall i, j \in [1, ..., m]$,*

$$\lim_{m\to\infty} ||h_i^{(3)} - h_j^{(3)}||_2 = 0; \lim_{m\to\infty} ||h_i^{\prime(3)} - h_j^{\prime(3)}||_2 = 0; \textit{and } \lim_{m\to\infty} ||h_i^{\prime\prime(3)} - h_j^{\prime\prime(3)}||_2 = 0. \tag{5}$$

Proof: Please see Section 7.3 (Theoretical Analysis).

Note: It is worth mentioning that the assumptions made in Theorem 1 are minimal. In the case of L_2-bounded weights, it has been further explained in Wu et al. (2023); Cai & Wang (2020). The assumption that graph node features at $l = 0$ are bounded is also minimal as we use a frozen features

extractor, which ideally should not generate nondefinitive values for a finite feature vector. With that, we show

$$\left\|h_i^{(3)} - h_j^{(3)}\right\|_2 \leq (\frac{1}{m+1})^3 \left\|h_i'^{(0)} - h_j'^{(0)}\right\|_2 \left\|W^{(2)}\right\|_2 \left\|W^{(1)}\right\|_2 \left\|W^{(0)}\right\|_2,$$

which essentially means that $\left\|h_i^{(3)} - h_j^{(3)}\right\|_2$ is upper bounded inversely with the reciprocal of $(m+1)^3$. Thus, as m increases, the upper-bound gets tighter and eventually leads to $\lim_{m\to\infty} ||h_i^{(3)} - h_j^{(3)}||_2 = 0$. As a result, we can conclude the following corollary.

Corollary 1 *$\forall i \in [1, ..., m]$ and m sufficiently large, $h_i^{(3)} \longrightarrow h^*$, $h_i'^{(3)} \longrightarrow h'^*$, and $h_i''^{(3)} \longrightarrow h''^*$, where h^*, h'^*, and h''^* are functions of m; h^*, h'^*, and h''^* are the convergence node for each magnification.*

This is necessarily equivalent to pooling the nodes in magnification level, yet with a completely new approach than the traditional pooling layer, without further imposing computational load on the network for pooling. Taking into account the fact that h^*, h'^*, and h''^* are not necessarily equal, our model is fusing node features in each magnification while it consults with other magnifications, and draw the conclusion via averaging nodes across three magnifications at the end of the convolutional layers by means of the readout module. This means that the final embedding of the graph is $\frac{h^* + h'^* + h''^*}{3}$. We believe that this process helps the model reduce variance and uncertainty in making predictions as m grows. To support this claim, we provide empirical evidence detailed in Section 7.5.1 (Monte Carlo Test).

The structure of the graph has been designed in such a way that it does not get stuck in the bottleneck of over-smoothing, a common issue in deep GCNs Cai & Wang (2020). Our intuition is that nodes in $\mathbf{M}_1$, $\mathbf{M}_2$, and $\mathbf{M}_3$ are interacting via message passing, and the flow of inter-magnification information helps the model to keep its balance and continue the process of learning. Nevertheless, by increasing the number of GCN layers for the model to four or higher, the so-called over-smoothing problem will take place, which can possibly deteriorate the model's performance. On the other hand, less than three layers of GCNs might not be able to fully capture the inter-magnification interactions. This leaves us with three layers of GCNs, which is equal to the graph's diameter. In addition to this theoretical description, we empirically support our claim in Section 7.5.5 (Graph Depth).

4 Experiments

4.1 Data Preparation

We utilize three datasets: Esophageal Carcinoma (ESCA) from The Cancer Genome Atlas (TCGA), which includes 135 WSIs across two subtypes, and Ovarian Carcinoma and Bladder Cancer, where Ovarian Carcinoma consists of 948 WSIs with five histotypes, while Bladder Cancer contains 262 WSIs with two histotypes. These datasets were curated using HistoQC Janowczyk et al. (2019). A detailed breakdown of each dataset is available in Table 3.

4.2 Comparisons with State-of-the-Art

To compare with state-of-the-art approaches, we repeated these experiments with the same cross-validation folds and random seeds to have a fair comparison; the choice of ten random seeds is to capture statistical significance and reliability. For evaluating the models, we adopt Balanced Accuracy and F1 Score since these metrics show how reliable a model performs on imbalanced data, and more importantly on clinical applications. We compare our proposed model, GRASP, with models using different approaches to have a broad spectrum of evaluation. These models include Ab-MIL Ilse et al. (2018), Trans-MIL Shao et al. (2021), CLAM-SB Lu et al. (2021), and CLAM-MB Lu et al. (2021) from the attention/transformer-based family; ZoomMIL (2021) Thandiackal et al. (2022), H2MIL (2022) Hou et al. (2022), and HiGT (2023) Guo et al. (2023) from multi-magnification approaches since they have a hierarchical structure and are compatible with our patch extraction paradigm; and PatchGCN: latent & spatial Chen et al. (2021) and DGCN: latent & spatial Zheng et al. (2021) from graph-based learning approaches.

4.3 SUBTYPE PREDICTION

Table 1 shows the comparison between our model and state-of-the-art methods based on Swin features, where GRASP outperforms all the competing methods on the Ovarian and Bladder datasets. Interestingly, Ab-MIL and CLAMs are the closest-performing methods to GRASP on these two datasets. On the ESCA dataset, however, ZoomMIL is the superior model with GRASP being the closest counterpart. Overall, GRASP is the superior model among all other models based on the average Balanced Accuracy on the three datasets.
Table 2 shows the comparison between our model and state-of-the-art methods based on KimiaNet features, where GRASP outperforms all the competing methods by a margin of $2.6\% - 10.7\%$ Balanced Accuracy on the Ovarian dataset and $0.4\% - 10.0\%$ on the Bladder dataset. It is worth mentioning that ZoomMIL is the closest-performing model to GRASP, although ZoomMIL has 7 times more parameters than GRASP. PatchGCN and DGCN are not performing comparably to GRASP, even though they are using spatial information that GRASP does not. This implies that a multi-magnification graph structure can potentially show more capability compared to other state-of-the-art approaches in terms of representing gigapixel WSIs. Moreover, single-magnification approaches are faster in terms of inference time than other approaches, especially CLAM-SB which has the lowest inference time. Inference times (per slide) have been calculated on the same machine for all models.

Table 1: The average performance on 3 folds and 10 random seeds based on Swin's features. The **best** and second best average values are highlighted in **bold** and underlined, respectively.

Model	*Params.*	*Inference*	**Ovarian**: *Five subtypes*		**Bladder**: *Two subtypes*		**ESCA**: *Two subtypes*		**Average**
			Balanced Acc.	F1 Score	Balanced Acc.	F1 Score	Balanced Acc.	F1 Score	Balanced Acc.
Trans-MIL	2.672M	0.019 sec	0.297 ± 0.011	0.244 ± 0.011	0.830 ± 0.037	0.819 ± 0.030	0.626 ± 0.021	0.611 ± 0.021	0.584
Ab-MIL	0.263M	0.015 sec	0.643 ± 0.022	0.647 ± 0.020	0.900 ± 0.013	0.884 ± 0.023	0.818 ± 0.010	0.812 ± 0.004	0.787
CLAM-SB	0.795M	0.014 sec	0.546 ± 0.062	0.550 ± 0.065	0.903 ± 0.051	0.902 ± 0.044	0.877 ± 0.067	0.861 ± 0.056	0.775
CLAM-MB	0.796M	0.015 sec	0.558 ± 0.044	0.565 ± 0.042	0.903 ± 0.032	0.901 ± 0.028	0.848 ± 0.055	0.833 ± 0.051	0.769
DGCN: *latent*	0.790M	0.098 sec	0.224 ± 0.017	0.146 ± 0.017	0.725 ± 0.052	0.655 ± 0.108	0.763 ± 0.051	0.736 ± 0.059	0.570
DGCN: *spatial*	0.790M	0.086 sec	0.210 ± 0.011	0.133 ± 0.012	0.700 ± 0.044	0.620 ± 0.074	0.660 ± 0.049	0.606 ± 0.053	0.523
PatchGCN: *latent*	1.385M	0.099 sec	0.397 ± 0.039	0.362 ± 0.047	0.537 ± 0.011	0.351 ± 0.052	0.855 ± 0.076	0.847 ± 0.076	0.596
PatchGCN: *spatial*	1.385M	0.110 sec	0.423 ± 0.042	0.390 ± 0.053	0.527 ± 0.020	0.336 ± 0.017	0.864 ± 0.080	0.859 ± 0.077	0.605
ZoomMIL	2.891M	0.024 sec	0.640 ± 0.018	0.648 ± 0.011	0.899 ± 0.046	0.895 ± 0.037	$\mathbf{0.889 \pm 0.037}$	$\mathbf{0.895 \pm 0.040}$	0.809
HiGT	6.388M	0.148 sec	0.251 ± 0.037	0.184 ± 0.049	0.755 ± 0.041	0.717 ± 0.023	0.760 ± 0.050	0.744 ± 0.061	0.588
H2MIL	0.829M	0.092 sec	$\mathbf{0.671 \pm 0.008}$	$\mathbf{0.667 \pm 0.024}$	0.900 ± 0.054	0.899 ± 0.044	0.854 ± 0.072	0.845 ± 0.084	0.808
GRASP *(ours)*	0.378M	0.024 sec	0.669 ± 0.029	0.654 ± 0.041	$\mathbf{0.905 \pm 0.058}$	$\mathbf{0.906 \pm 0.051}$	0.877 ± 0.111	0.872 ± 0.112	**0.817**

Comparing Tables 1 and 2, all the models performed better with KimiaNet embeddings than with Swin embeddings, which is mostly because KimiaNet has domain knowledge and can provide more contextual features than Swin. Furthermore, GRASP, H2MIL, ZoomMIL, Ab-MIL, and CLAMs showcase robust generalization and effective performance even when utilizing features from different backbones, especially with GRASP being the most robust model.

Although Deng et al. (2023) has used attention score distribution to show their model is reliable across different magnifications, we want to step further and adopt a similar logic as first introduced in Selvaraju et al. (2017) to define the concept of energy of gradients for graph nodes (*Please see Graph-Based Visualization 7.6 in Appendix*). Therefore, for the first time in the field, we show that an AI model such as GRASP can learn the concept of magnification and behave according to the subtype and slide characteristics. To this end, we formulate an experiment to obtain a sense of each magnification's influence on the model which leads to Figure 3. The main takeaway of this experiment is that depending on the subtype, the distribution of referenced magnifications by GRASP is different. From a pathological point of view, this finding fits our knowledge of the biological properties of each subtype. As an example, we conducted a case study on the bladder dataset, where the micropapillary subtype is known to be diagnosed generally in lower magnification owing to its morphological properties and the structure of micropapillary tumors, whereas UCC needs to be examined in higher magnifications due to its cell- and texture-dependent structure.

On the Ovarian dataset, for the subtype ENOC, endometrioids can often be recognized and identified at low power as they tend to have characteristic glandular architecture occupying contiguous, large areas. Low-grade serous carcinomas (LGSC) can be very difficult at low power due to the necessity of confirming low-grade cytology at high power. For CCOC, clear cell carcinomas have characteristic low-power architectural patterns but can also require high-power examinations to exclude high-grade serous carcinoma with clear cell features, meaning that important information is distributed on all magnifications. The other subtypes, MUC and HGSC, may either show pathognomic

Table 2: The average performance on 3 folds and 10 random seeds based on KimiaNet's features. The **best** and second best average values are highlighted in **bold** and underlined, respectively.

Model	*Params.*	*Inference*	**Ovarian**: *Five Subtypes*		**Bladder**: *Two Subtypes*		**Model's Average**
			Balanced Acc.	F1 Score	Balanced Acc.	F1 Score	Balanced Acc.
Trans-MIL	$2.672M$	0.019 sec	0.647 ± 0.007	0.632 ± 0.005	0.868 ± 0.023	0.877 ± 0.013	0.758
Ab-MIL	$\mathbf{0.263M}$	0.015 sec	0.692 ± 0.016	0.680 ± 0.014	0.919 ± 0.018	0.922 ± 0.016	0.806
CLAM-SB	$0.795M$	**0.014** sec	0.627 ± 0.015	0.623 ± 0.010	0.908 ± 0.026	0.911 ± 0.023	0.768
CLAM-MB	$0.796M$	0.015 sec	0.620 ± 0.035	0.609 ± 0.030	0.901 ± 0.039	0.906 ± 0.037	0.761
DGCN: *latent*	$0.790M$	0.098 sec	0.654 ± 0.017	0.652 ± 0.024	0.835 ± 0.034	0.841 ± 0.035	0.745
DGCN: *spatial*	$0.790M$	0.086 sec	0.654 ± 0.009	0.652 ± 0.009	0.867 ± 0.015	0.875 ± 0.007	0.761
PatchGCN: *latent*	$1.385M$	0.099 sec	0.683 ± 0.003	0.675 ± 0.005	0.911 ± 0.031	0.919 ± 0.020	0.797
PatchGCN: *spatial*	$1.385M$	0.110 sec	0.672 ± 0.002	0.662 ± 0.005	0.896 ± 0.033	0.905 ± 0.021	0.784
ZoomMIL	$2.891M$	0.024 sec	0.701 ± 0.020	$\mathbf{0.690 \pm 0.021}$	0.931 ± 0.008	0.933 ± 0.009	0.816
HiGT	6.388M	0.148 sec	0.337 ± 0.044	0.288 ± 0.054	0.847 ± 0.067	0.842 ± 0.055	0.592
H2MIL	0.829M	0.092 sec	0.653 ± 0.018	0.658 ± 0.032	0.876 ± 0.054	0.876 ± 0.048	0.764
GRASP *(ours)*	$0.378M$	0.024 sec	$\mathbf{0.727 \pm 0.036}$	0.689 ± 0.040	$\mathbf{0.935 \pm 0.011}$	$\mathbf{0.937 \pm 0.014}$	**0.831**

architectural features at low power or require high-power examination on a case-by-case basis. According to Figure 3, GRASP collects the information from all three magnifications for MUC and CCOC.

Furthermore, to examine whether GRASP understands the biological meaning of the data, i.e., differentiating between tumor vs non-tumor regions, we conduct a visualization experiment to plot the pixel-level heatmap of patches in multiple magnifications as depicted in Figure 4.

4.4 Ablation Study

Here, we design five experiments to evaluate our proposed model. Firstly, a Monte Carlo test on graph size, i.e., the number of nodes to investigate the impact of the number of nodes on the model's performance (Section Monte Carlo Test 7.5.1). Secondly, analyzing model performance on individual or pairs of magnifications to study the effectiveness of multi-magnification representation (Section Magnification Test 7.5.2). Thirdly, we investigate the effect of different graph convolution types (Section Graph Convolutions 7.5.3). In the fourth experiment, we study how different models perform when all the patches from a WSI are used (Section Patch Number 7.5.4). Lastly, we empirically show that the graph depth of $d = 3$ is the appropriate choice for our design (Section Graph Depth 7.5.5).

5 Conclusion

In this work, we developed GRASP, the first *lightweight* multi-magnification framework for processing gigapixel WSIs. GRASP is a *fixed-structure* model that can learn multi-magnification interactions in the data based on the idea of capturing both the inter- and intra-magnification information. This relies on the theoretical property of the model, where it benefits from intra-magnification convergence to pool the nodes rather than conventional pooling layers. GRASP, with its pre-defined fixed structure, has comparably fewer parameters than other state-of-the-art multi-magnification models in the field and outperforms the competing models in terms of average *Balanced Accuracy* over three complex cancer datasets using two different backbones. For the first time in the field, confirmed by two expert genitourinary pathologists, we showed that our model is dynamic in finding and consulting with different magnifications for subtyping two challenging cancers. We also evaluated the model's decision-making to show that the model is learning semantics by highlighting tumorous regions in patches. Furthermore, we not only run extensive experiments to show the model's reliability and stabilization in terms of its different hyperparameters, but we also provide the theoretical foundation of our work to shed light on the dynamics of GRASP in interacting with different nodes and magnifications in the graph. To conclude, we hope that the strong characteristics of GRASP and its straightforward structure, along with the theoretical basis, will encourage the modeling of lightweight structure-based design in the field of digital pathology for WSI representation.

6 MEANINGFULNESS STATEMENT

In digital pathology, a meaningful representation of life involves capturing the intricate, multi-scale structures of biological tissues, similar to how pathologists operate. GRASP (GRAph-Structured Pyramidal Whole Slide Image Representation) aids this process by modeling whole slide images as hierarchical graphs that integrate information across different microscopic magnification levels. This method, though lightweight, improves cancer subtyping accuracy and aligns computational analysis with human diagnostic processes, promoting deeper insights into tissue architecture and disease mechanisms.

7 APPENDIX

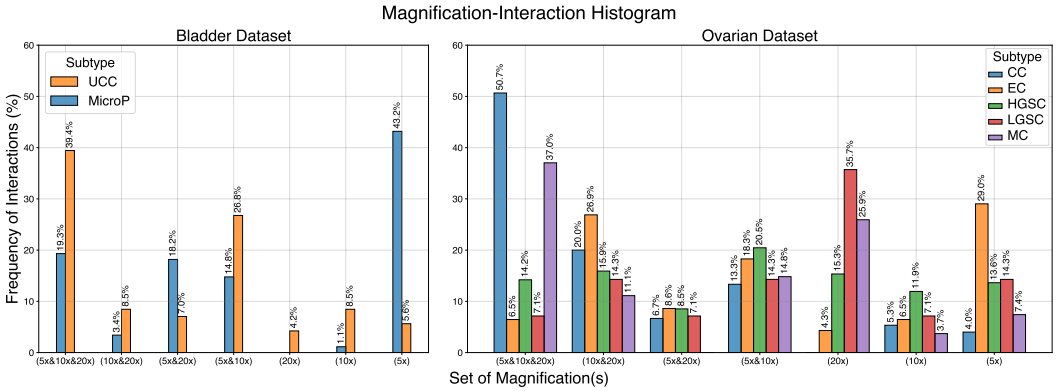

Figure 3: The histogram of consultations conducted by GRASP with different magnifications. First, this shows GRASP is actively dynamic in terms of capturing information from different magnifications benefiting from its multi-magnification structure. Second, information is distributed differently over magnifications depending on the subtype and slide, and there is no optimal magnification for a subtype. For example, in the Bladder dataset, '$(5x\&10x\&20x)$' shows that the model needed to consult with all three magnifications for 19.3% and 39.4% of slides for MicroP and UCC, respectively; '$(5x)$' shows that the model has mostly focused on only $5x$ magnification for 43.2% and 5.6% of slides for MicroP and UCC, respectively. This behavior is similar to pathologists, where they can diagnose massive MicroP tumors with lower magnifications, while they need to consult with higher magnifications to confirm a minuscule mass of MicroP tumors. On the other hand, UCC is hard to diagnose at lower magnifications and requires careful examination with different magnifications due to its morphological complexity, which fits the model behavior in proclivity to highlight more than one magnification for the majority of cases.

7.1 DATASETS

Table 3: Summary of the datasets used in this study.

Dataset	Source	No. of WSIs	Histotypes/Subtypes
Ovarian Carcinoma	Private Dataset	948	High-Grade Serous Carcinoma (HGSC): 410
			Clear Cell Ovarian Carcinoma (CCOC): 167
			Endometrioid Carcinoma (ENOC): 237
			Low-Grade Serous Carcinoma (LGSC): 69
			Mucinous Carcinoma (MUC): 65
Esophageal Carcinoma (ESCA)	TCGA	135	Adenocarcinoma: 86
			Squamous Cell Carcinoma: 49
Bladder Cancer	Private Dataset	262	Micropapillary (MicroP): 128
			Conventional Urothelial Carcinomas (UCC): 134

A total of 1,133,388 patches of size 1000×1000 pixels for the Ovarian dataset, 602,874 patches of size 224×224 from the ESCA dataset, and 313,191 patches of size 1000×1000 pixels for the Bladder dataset are extracted in multi-magnification setting (approximately 2 TB of Gigapixel WSIs). Patches being extracted such that they do not overlap at $\mathbf{M}_3$ while overlapping at $\mathbf{M}_2$ and $\mathbf{M}_1$ is inevitable. From each magnification, $m \leq 400$ patches (note that we use a large field of view, meaning this number eventually covers much of tissue regions) have been extracted per slide in both Ovarian and Bladder datasets, as it's been shown in Chen et al. (2022a); Wang et al. (2023); Rasoolijaberi et al. (2022) that a subset of patches is enough to represent WSIs.

For patch-level feature extraction, we utilized two backbones: KimiaNet and Swin_base. Given that KimiaNet was trained on TCGA data in a supervised fashion, we intentionally refrained from extracting features from the ESCA dataset using this backbone to ensure an unbiased and leakage-free comparison. Conversely, we employed Swin, pre-trained on ImageNet, to extract features from

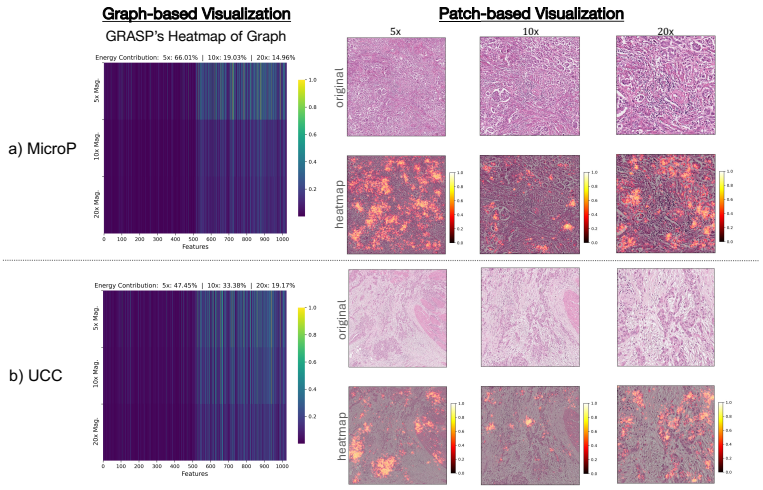

Figure 4: A case study on the Bladder dataset using KimiaNet features. **a)** Graph-based visualization: a random case from the subtype MicroP in the test data was selected to visualize its magnification heatmap where we show the absolute gradient in terms of each node. The $5x$ magnification contributes to 66.01% of the whole energy model spent on this slide, meaning GRASP overall emphasizes more on $5x$ on this slide. Patch-based visualization: GRASP highlights patches of the three magnifications of a region of interest. In the second row, highlighted regions show the model has identified those areas as important while paying minimal attention to other regions. As confirmed by an expert pathologist, the model's highlights on the three patches are tumors. The model can thus differentiate MicroP tumors from other tissue textures despite being trained for separating MicroP vs UCC. **b)** shows a similar case yet on the subtype UCC from a random slide in the test data. In this case, GRASP focuses on both $5x(47.45\%)$ and $10x(33.38\%)$ but is more interested in $5x$. As confirmed by the expert pathologist, the regions highlighted (yellowish areas in the second row) by the model are tumorous neighborhoods. Therefore, GRASP can differentiate UCC tumors from other textures and healthy cells across multiple magnifications.

all three datasets.
For each cancer dataset, we trained our proposed method in a 3-fold cross-validation and repeated the experiments *ten* times with different random seeds, where random seeds were randomly generated, to ensure a rigorous comparison. In order to prevent data leakage in our cross-validation splits, we split the slides based on their patients, since some patients have more than one slide, meaning that slides were split in a way that all slides from the same patient remain in the same set.

7.2 Training and Inference

To tackle the data imbalance problem, for all models in the study, we deployed a weighted cross entropy loss. A learning rate of 0.001 and a weight decay of 0.01 for Adam optimizer have been adopted, and in case competing models were not converging, learning rate of 0.0001 resolved the problem. Models were trained for 100, 50, and 10 epochs for the Ovarian, ESCA, and Bladder datasets, respectively. Specific to GRASP, the first two layers are of size 256 and the last layer output is of size 128. For all training and testing, the GPU hardware used was either a GeForce GTX 3090 Ti-24 GB (Nvidia), Quadro RTX 5000-16 GB (Nvidia), RTX 6000-48 GB (Nvidia) based on availability. Deep Graph Library (DGL), PyTorch, NumPy, SciPy, PyGeometric, and Scikit-Learn libraries have been used to perform the experiments.

7.3 Theoretical Analysis

Here, we prove Theorem 1 for any $h_i^{(3)}$ and $h_j^{(3)}$, and conclude the case for $h_i'^{(3)}$s and $h_i''^{(3)}$s similarly. To start with, we demonstrate Lemma 1.

Lemma 1 *For any given vectors x and y, and having $\|.\|_2$ as the L_2 norm, the following inequality holds,*

$$\|\alpha(x) - \alpha(y)\|_2 \leq \|x - y\|_2 \,, \tag{6}$$

where $\alpha(.) = \mathrm{ReLU}(.)$

proof: Let's reformulate $\mathrm{ReLU}(x)$ as $\frac{x+|x|}{2}$ where operator $|x|$ is the element-wise absolute value of the vector x. Thus,

$$\begin{aligned}
\|\alpha(x) - \alpha(y)\|_2 &= \left\| \frac{x + |x|}{2} - \frac{y + |y|}{2} \right\|_2 \\
&= \left\| \frac{x - y}{2} + \frac{|x| - |y|}{2} \right\|_2 \\
&\leq \left\| \frac{x - y}{2} \right\|_2 + \left\| \frac{|x| - |y|}{2} \right\|_2
\end{aligned} \tag{7}$$

using the reverse triangle inequality, $\left\| \frac{|x|-|y|}{2} \right\|_2 \leq \left\| \frac{x-y}{2} \right\|_2$, which gives result to

$$\begin{aligned}
\|\alpha(x) - \alpha(y)\|_2 &\leq \left\| \frac{x - y}{2} \right\|_2 + \left\| \frac{x - y}{2} \right\|_2 \\
&= \|x - y\|_2
\end{aligned} \tag{8}$$

□

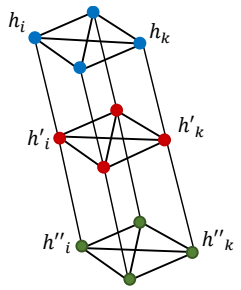

Figure 5: The structure of our hierarchical graph and the relationship between two given nodes h_i and h_k within and across different magnifications.

Recalling the main GCN formula for any l in 4, $\mathcal{N}(i)$ is the neighborhood size, and taking self-loops (*see Note 1*) into account, $\forall i \in [1, ..., m]$, $|\mathcal{N}(i)| = m + 1$. With the graph being symmetric, we deduce that $\forall j \in [1, ..., m]$, $|\mathcal{N}(j)| = m + 1$. These result in

$$c_{ji} = \sqrt{|\mathcal{N}(j)|\,|\mathcal{N}(i)|} = m + 1$$

hence, Eq. 4 is simplified as follows,

$$h_i^{(l+1)} = \alpha(b^{(l)} + \frac{1}{m+1} \sum_{j \in \mathcal{N}(i)} h_j^{(l)} W^{(l)}). \tag{9}$$

Now, for any i, we can partition the set of all nodes in $\mathcal{N}(i)$ into two partitions $\{h_1^{(l)}, ..., h_m^{(l)}\}$ and $\{h_i'^{(l)}\}$ based on the relationship between nodes as shown in Figure 5. Therefore, $\sum_{j \in \mathcal{N}(i)} h_j^{(l)}$ can be rewritten as,

$$\sum_{j \in \mathcal{N}(i)} h_j^{(l)} = \left(\sum_{j \in [1,...,m]} h_j^{(l)} \right) + h_i'^{(l)}. \tag{10}$$

The first term in Eq. 10 is common among all nodes at a given magnification, so we call it $\mathcal{H}^{(l)}$ leading to Eq. 11,

$$\sum_{j \in \mathcal{N}(i)} h_j^{(l)} = \mathcal{H}^{(l)} + h_i'^{(l)} \tag{11}$$

as a result, we combine Eq. 9 with Eq. 11 which yields

$$h_i^{(l+1)} = \alpha(b^{(l)} + \frac{1}{m+1}\left(\mathcal{H}^{(l)} + h_i'^{(l)}\right) W^{(l)}) \tag{12}$$

and similarly for any $j \neq i$ as well,

$$h_j^{(l+1)} = \alpha(b^{(l)} + \frac{1}{m+1}\left(\mathcal{H}^{(l)} + h_j'^{(l)}\right) W^{(l)}). \tag{13}$$

By using Lemma 1 and with combination with Eq. 12 and 13,

$$\begin{aligned}
&\left\| h_i^{(l+1)} - h_j^{(l+1)} \right\|_2 \\
&\leq \left\| \frac{1}{m+1} h_i'^{(l)} W^{(l)} - \frac{1}{m+1} h_j'^{(l)} W^{(l)} \right\|_2 \\
&= \frac{1}{m+1} \left\| \left(h_i'^{(l)} - h_j'^{(l)} \right) W^{(l)} \right\|_2 \\
&\leq \frac{1}{m+1} \left\| h_i'^{(l)} - h_j'^{(l)} \right\|_2 \left\| W^{(l)} \right\|_2 .
\end{aligned} \tag{14}$$

Therefore, we reach the inequality below,

$$\left\| h_i^{(l+1)} - h_j^{(l+1)} \right\|_2 \leq \frac{1}{m+1} \left\| h_i'^{(l)} - h_j'^{(l)} \right\|_2 \left\| W^{(l)} \right\|_2 . \tag{15}$$

Now, by going recursively over $l = 0, 1, 2$, we have

$$\begin{aligned}
&\left\| h_i^{(3)} - h_j^{(3)} \right\|_2 \leq \\
&(\frac{1}{m+1})^3 \left\| h_i'^{(0)} - h_j'^{(0)} \right\|_2 \left\| W^{(2)} \right\|_2 \left\| W^{(1)} \right\|_2 \left\| W^{(0)} \right\|_2 .
\end{aligned} \tag{16}$$

Since $\left\|W^{(2)}\right\|_2$, $\left\|W^{(1)}\right\|_2$, and $\left\|W^{(0)}\right\|_2$ are $L_2 - bounded$ based on our assumption. Also, $\left\| h_i'^{(0)} - h_j'^{(0)} \right\|_2$ is an $L_2 - bounded$ value based on our assumption (*as input image data is $L_2 - bounded$, and also our encoder ϕ is a bounded encoder: features are not scattered in an infinite space, rather they are encoded in a finite space*). Given these, by approaching $m \to \infty$ (*see remark 2*), the right side of the Eq. 16 approaches 0. Therefore,

$$\lim_{m \to \infty} \left\| h_i^{(3)} - h_j^{(3)} \right\|_2 = 0 \tag{17}$$

similar to this case, it can be proved that

$$\lim_{m \to \infty} \left\| h_i'^{(3)} - h_j'^{(3)} \right\|_2 = 0 \tag{18}$$

$$\lim_{m \to \infty} \left\| h_i''^{(3)} - h_j''^{(3)} \right\|_2 = 0 \tag{19}$$

□

remark 1 *To implement GCNs, self-loops are considered to represent the relationship between each node with itself, and it is also part of the technical implementation of the models. Thus, we consider this fact in our theoretical discussion.*

remark 2 *Empirically, reaching sufficiently large m can guarantee the convergence. For example, $m = 10$ can affect the upper bound in Eq. 16 with an order of $\frac{1}{10^3}$, while $m = 100$ can affect the upper bound with an order of $\frac{1}{10^6}$. In our experiments, $m = 400$ has been adopted that guarantees the convergence with an order of $\frac{1}{64 \times 10^6}$. Therefore, the larger m, the tighter together node features at the last GCN layer.*

Corollary 2 *$\forall i \in [1, ..., m]$ and m sufficiently large, $h_i^{(3)} \longrightarrow h^*$, $h_i'^{(3)} \longrightarrow h'^*$, and $h_i''^{(3)} \longrightarrow h''^*$, where h^*, h'^*, and h''^* are functions of m; h^*, h'^*, and h''^* are the convergence node for each magnification.*

Description: *given Eq. 17, every two arbitrary nodes $h_i^{(3)}$ and $h_j^{(3)}$ in one level of magnification are converging to each other. This means that all nodes are converging to the same value, which we name h^*. Thus, $\forall i \in [1, ..., m]$, $\lim_{m\to\infty} \left\| h_i^{(3)} - h^* \right\|_2 = 0$ or equivalently $h_i^{(3)} \longrightarrow h^*$. Using the same logic as above, one can conclude $h_i'^{(3)} \longrightarrow h'^*$ and $h_i''^{(3)} \longrightarrow h''^*$. Since each of h^*, h'^*, and h''^* are a function of m, increasing m would result in them being a better estimation/representation for the intra-magnification information.*

7.4 Empirical Proof

In addition to the theoretical analysis in Section 7.3, we empirically demonstrate the intra-magnification convergence in Figure 6. In this experiment, we plot the mean squared error between all nodes in a magnification and the corresponding convergence node. As shown, the mean squared error at the third layer ($\ell = 3$) is nearly zero, providing empirical evidence for the convergence of the nodes, i.e., the nodes being pooled without the need for a pooling layer.

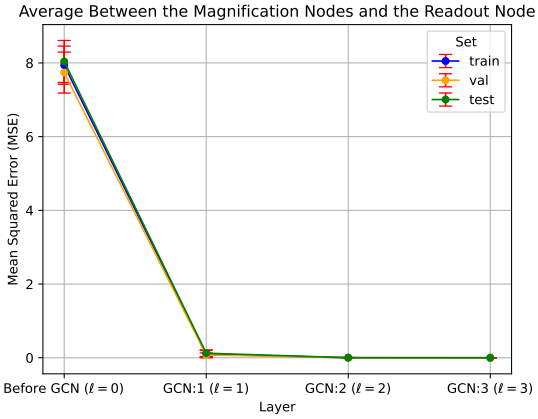

Figure 6: Empirical proof for intra-magnification convergence at $\ell = 3$.

7.5 Ablation Study

7.5.1 Monte Carlo Test

As described in Algorithm 1, this test requires $m = 400$. Therefore, we removed four slides that did not have 400 non-overlapping patches in $20x$ from the Bladder dataset and ran this test. Because of removing these four slides from the dataset, the result of Figure 4 in the paper is not comparable with any of the tables in the paper. In this test, a subset of the nodes is randomly dropped from each magnification layer (corresponding nodes in each magnification are removed), to create independent graphs. Then, training and inference happens and the results are reported at the end.

In this experiment, we take a graph of size $3m = 1200$, and randomly drop a set of its nodes, along with their multi-magnification correspondences, to build a new graph with a smaller size. For example, when $count = 10$, we randomly drop 10 triplets of $(h_{index}, h'_{index}, h''_{index})$ from the graph to create a new graph Q_r of the size 1170. Similarly, when $count = 200$, we randomly drop 200 triplets of $(h_{index}, h'_{index}, h''_{index})$ from the graph to create a new graph Q_r of the size 600. This is an aggressive way to create statistically independent graphs of smaller sizes. To capture statistical variance, we repeat the experiment 10 times to create 40 graphs with different sizes and report the model performance in Figure 7. To accomplish this, the same 3-fold cross-validation sets

Algorithm 1 Monte Carlo Test on Graph Size

Data: $m = 400$
Result: *Balanced Accuracy and Standard Deviation*
$step \leftarrow 10$
Load DATASET
$D \leftarrow \{\emptyset\}$ { Node indices to be dropped}
for *iter* $\leftarrow 1$ **to** 10 **do**
  for G_r *in DATASET* **do**
    for *count* $\leftarrow 10$ **to** 390 $step$ **do**
      $D \leftarrow \text{RANDOM}([1, ..., m], \text{count})$
      $Q_{r,\text{count}} \leftarrow G_r$
      for *index in* D **do**
        $Q_{r,\text{count}} \leftarrow \text{DROP}(Q_{r,\text{count}}, h_{\text{index}})$
        $Q_{r,\text{count}} \leftarrow \text{DROP}(Q_{r,\text{count}}, h'_{\text{index}})$
        $Q_{r,\text{count}} \leftarrow \text{DROP}(Q_{r,\text{count}}, h''_{\text{index}})$
      end
      STORE$(Q_{r,\text{count}})$
    end
  end
end
Take $Q_{r,\text{count}}$ *for Cross-Validation and Inference*
report *Balanced Accuracy and Standard Deviation*

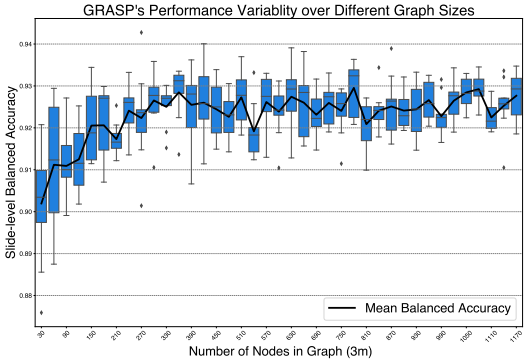

Figure 7: Monte Carlo experiments on the graph size. As the number of nodes increases, the uncertainty decreases and the model stabilizes.

and 10 random seeds have been used for all repetitions. Taking 10 times repetitions of 40 different graph sizes into account, we performed $12,000$ independent training and inference experiments. Algorithm 1 demonstrates this experiment.

As can be seen in Figure 7, the performance of GRASP increases and stabilizes as the number of nodes increases. Since the standard deviation decreases as the number of nodes increases, it brings

Table 4: Average Performance on 3-folds and 10 random seeds based on KimiaNet's features.

Model	Bladder Cancer	
	Balanced Acc.	F1 Score
Graph on $\mathbf{M}_1$	0.898 ± 0.052	0.890 ± 0.047
Graph on $\mathbf{M}_2$	0.927 ± 0.057	0.928 ± 0.051
Graph on $\mathbf{M}_3$	0.905 ± 0.035	0.913 ± 0.023
Graph on $\mathbf{M}_1$&$\mathbf{M}_2$	0.919 ± 0.032	0.919 ± 0.030
Graph on $\mathbf{M}_1$&$\mathbf{M}_3$	0.917 ± 0.031	0.922 ± 0.033
Graph on $\mathbf{M}_2$&$\mathbf{M}_3$	0.926 ± 0.024	0.934 ± 0.022
GRASP (ours)	$\mathbf{0.935 \pm 0.011}$	$\mathbf{0.937 \pm 0.014}$

Table 5: The average performance on 3 folds and 10 random seeds, based on Swin's features and the setting where all the patches were extracted from each WSI.

Model	Bladder: Two Subtypes		
	Balanced Acc.	F1 Score	AUC
ZoomMIL	0.879 ± 0.065	0.872 ± 0.060	0.951 ± 0.031
HiGT	0.720 ± 0.049	0.658 ± 0.042	0.819 ± 0.066
H2MIL	0.877 ± 0.050	0.871 ± 0.035	0.966 ± 0.022
GRASP (GCN)	$\mathbf{0.883 \pm 0.069}$	$\mathbf{0.879 \pm 0.065}$	$\mathbf{0.953 \pm 0.031}$
GRASP (GAT)	$\mathbf{0.917 \pm 0.013}$	$\mathbf{0.907 \pm 0.017}$	$\mathbf{0.978 \pm 0.007}$
GRASP (SAGEConv)	$\mathbf{0.936 \pm 0.023}$	$\mathbf{0.932 \pm 0.015}$	$\mathbf{0.988 \pm 0.008}$

to light the concept of variance convergence, meaning that the model with $m \geq 200$ is fairly generalizable over different cross-validation folds and is statistically reliable in terms of performance. This is also in agreement with our theoretical expectation based on inter-magnification convergence that as m grows, the model has better convergence resulting in more stability.

7.5.2 Magnification Test

To confirm that the idea of multi-magnification is valid and that multi-magnification is the cause for the model's performance, we design *6* different experiments (repeated on 10 random seeds and 3 folds) on the Bladder dataset, with KimiaNet as the backbone, as our empirical evidence. These include evaluating the same model on only $\mathbf{M}_1$, $\mathbf{M}_2$, and $\mathbf{M}_3$ fully connected graphs and on pairs of $\mathbf{M}_1\&\mathbf{M}_2$, $\mathbf{M}_1\&\mathbf{M}_3$, and $\mathbf{M}_2\&\mathbf{M}_3$. The results in Table 4 show that GRASP is superior to all other methods. One possible explanation is that for those single and paired graphs, three layers of GCNs most likely cause the aforementioned over-smoothing problem, which shows that GRASP can effectively capture the information contained in different magnifications and boost its performance.

7.5.3 Graph Convolutions

To study the impact of different graph convolutions on the performance of GRASP, we designed this experiment where we replaced the GCN layers in the original layers, with a newer version of graph convolutions. As seen in Figure 8, GAT and SAGEConv improve the performance of the primary GCN-based GRASP in terms of Balanced Accuracy over different batch sizes.

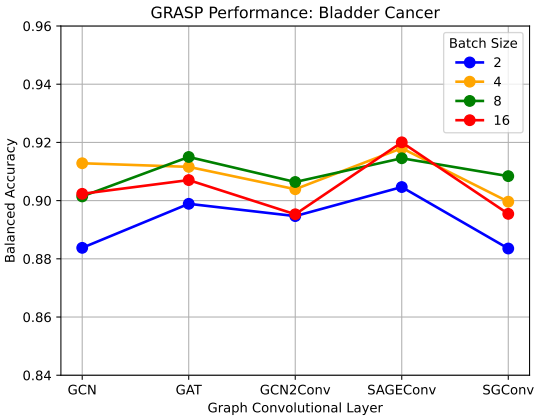

Figure 8: Ablation on the effect of different GNN structures benchmarked on 3 folds and 10 random seeds with different batch sizes. GNNs were taken from Deep Graph Library.

7.5.4 Patch Number

To compare the models' performance when all the patches from each slide are extracted, we designed this experiment and benchmarked the baseline against different variations of GRASP. GCN-based

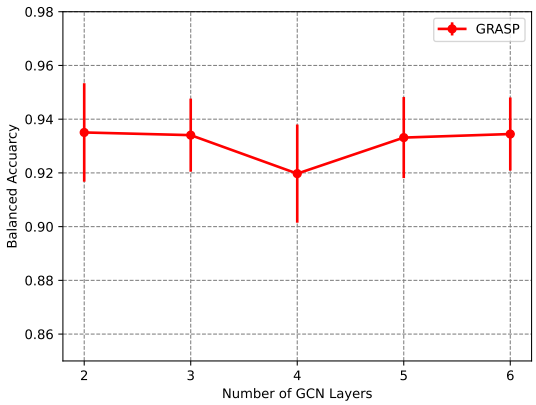

Figure 9: Our model's performance as the number of GCN layers increases. As can be seen, by increasing the number of layers from *three* to *four* a relatively large drop happens, showing that the model is being over-smoothed. However, the model is recovering this loss at $five$ layers or more yet with a relatively larger standard deviation compared to *three* layers.

GRASP is superior to other multi-magnification methods such as ZoomMIL, H2MIL, and HiGT. Compared to Table 5, the increased number of patches helps GRASP to improve, yet it degrades the performance of the other three models. However, single Magnification methods are more competitive with a higher density of patches. With this in mind, we followed the previous ablation study 7.5.3 and compared GRASP's performance with newer graph convolutions. Consequently, GRASP with SAGEConv outperforms all other models. This further emphasizes the flexibility of the graph structure that can easily employ different graph convolutions, which we belive is a unique advantage of GRASP.

7.5.5 Graph Depth

We experimented with GRASP with different numbers of GCN layers as shown in Figure 9. Firstly, *three* layers of GCNs show the same performance as *two* layers yet with lower standard deviation. Secondly, *four* layers of GCNs show a sudden drop in performance and increase in standard deviation, which can be attributed to over smoothing problem Cai & Wang (2020); Wu et al. (2023). In addition, with more than $five$ layers of GCNs, the network can recover the performance but with a slightly higher standard deviation and, clearly, an increased number of parameters. This shows that our original architecture of *three* layers is the best choice in the trade-off of average accuracy and the model's reliability. Based on the discussion in Wu et al. (2023), we expect the same phenomenon for different graph convolution types to happen.

7.6 Graph-Based Visualization

Let's call the output of the classifier, S, where the logit for correctly classified slides/graphs is S_c. To visualize the importance of magnifications, we compute the magnitude of the gradient of a graph with respect to its node features at $l = 0$ ($h_i^{(0)}$; for the sake of brevity, we drop the superscript (0) form $h_i^{(0)}$ and show it as h_i), which is $\left|\frac{\partial S_c}{\partial h_i}\right|$ for the magnification $\mathbf{M}_1$. $\left|\frac{\partial S_c}{\partial h_i}\right|$ is a vector of size 1024×1, and we have m nodes giving result to m such vectors. Likewise, we can define $\left|\frac{\partial S_c}{\partial h_i'}\right|$ and $\left|\frac{\partial S_c}{\partial h_i''}\right|$ for $\mathbf{M}_2$ and $\mathbf{M}_3$, respectively. Arranging these absolute gradients for each magnification in a

matrix of size $m \times 1024$ as follows,

$$\textit{Heatmap}_{\mathbf{M}_1} = \begin{bmatrix} \left|\frac{\partial S_c}{\partial h_1}\right| \\ \vdots \\ \left|\frac{\partial S_c}{\partial h_m}\right| \end{bmatrix} \tag{20}$$

$$\textit{Heatmap}_{\mathbf{M}_2} = \begin{bmatrix} \left|\frac{\partial S_c}{\partial h'_1}\right| \\ \vdots \\ \left|\frac{\partial S_c}{\partial h'_m}\right| \end{bmatrix} \tag{21}$$

$$\textit{Heatmap}_{\mathbf{M}_3} = \begin{bmatrix} \left|\frac{\partial S_c}{\partial h''_1}\right| \\ \vdots \\ \left|\frac{\partial S_c}{\partial h''_m}\right| \end{bmatrix} \tag{22}$$

As such, putting matrices in 20, 21, and 22 together in a matrix gives us the overall heatmap for the graph, of the size $3m \times 1024$, to compare the influence of each magnification:

$$\textit{Heatmap} = \begin{bmatrix} \textit{Heatmap}_{\mathbf{M}_1} \\ \textit{Heatmap}_{\mathbf{M}_2} \\ \textit{Heatmap}_{\mathbf{M}_3} \end{bmatrix} \tag{23}$$

This is the heatmap depicted in Figure 9 as a graph-based heatmap, which shows how model focuses on different magnifications.

Having the gradient for each node in the graph, we develop the concept of energy of gradients to find out which magnification(s) play a more important role in GRASP's final decision. To do so, we start by defining $\mathcal{E}_{\mathbf{M}_1}$ as follows,

$$\mathcal{E}_{\mathbf{M}_1} = \sum_{i \in [1,\dots,m]} \left\| \frac{\partial S_c}{\partial h_i} \right\|_2^2 \tag{24}$$

similarly, for $\mathbf{M}_2$ and $\mathbf{M}_3$,

$$\mathcal{E}_{\mathbf{M}_2} = \sum_{i \in [1,\dots,m]} \left\| \frac{\partial S_c}{\partial h'_i} \right\|_2^2 \tag{25}$$

$$\mathcal{E}_{\mathbf{M}_3} = \sum_{i \in [1,\dots,m]} \left\| \frac{\partial S_c}{\partial h''_i} \right\|_2^2 \tag{26}$$

Having these energies, the energy contribution of each magnification is calculated based on their relative share in the whole energy spent in the graph:

$$\mathbf{M}_1\textit{'s contribution} = \frac{\mathcal{E}_{\mathbf{M}_1}}{\mathcal{E}_{\mathbf{M}_1} + \mathcal{E}_{\mathbf{M}_2} + \mathcal{E}_{\mathbf{M}_3}} \tag{27}$$

$$\mathbf{M}_2\textit{'s contribution} = \frac{\mathcal{E}_{\mathbf{M}_2}}{\mathcal{E}_{\mathbf{M}_1} + \mathcal{E}_{\mathbf{M}_2} + \mathcal{E}_{\mathbf{M}_3}} \tag{28}$$

$$\mathbf{M}_3\textit{'s contribution} = \frac{\mathcal{E}_{\mathbf{M}_3}}{\mathcal{E}_{\mathbf{M}_1} + \mathcal{E}_{\mathbf{M}_2} + \mathcal{E}_{\mathbf{M}_3}} \tag{29}$$

Accordingly, the importance of each magnification can be quantified for further investigations. More samples are provides in Figure 3.

REFERENCES

Chen Cai and Yusu Wang. A note on over-smoothing for graph neural networks. *arXiv preprint arXiv:2006.13318*, 2020.

Chen Chen, Ming-Yuan Lu, David F. K. Williamson, Andrew D. Trister, Ravi G. Krishnan, and Faisal Mahmood. Fast and scalable search of whole-slide images via self-supervised deep learning. *Nature biomedical engineering*, 6(12):1420–1434, 2022a.

Richard J Chen, Ming-Yuan Lu, Muhammad Shaban, Chi Chen, Ting-Yun Chen, David F 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*, pp. 339–349. Springer, 2021.

Richard J Chen, Cheng Chen, Yuanfang Li, Tsung-Ying Chen, Andrew D Trister, Ravi 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*, pp. 16144–16155, 2022b.

Richard J Chen, Tong Ding, Ming Y Lu, Drew FK Williamson, Guillaume Jaume, Bowen Chen, Andrew Zhang, Daniel Shao, Andrew H Song, Muhammad Shaban, et al. A general-purpose self-supervised model for computational pathology. *arXiv preprint arXiv:2308.15474*, 2023.

Ozan Ciga, Tony Xu, and Anne Louise Martel. Self supervised contrastive learning for digital histopathology. *Machine Learning with Applications*, 7:100198, 2022. ISSN 2666-8270. doi: https://doi.org/10.1016/j.mlwa.2021.100198. URL https://www.sciencedirect.com/science/article/pii/S2666827021000992.

Ruiqi Deng, Chun Cui, Lester W. Remedios, Shunxing Bao, Ryan M. Womick, Sylvain Chiron, Jialun Li, Joseph T. Roland, Keith S. Lau, Qing Liu, and Keith T. Wilson. Cross-scale multi-instance learning for pathological image diagnosis. *arXiv preprint arXiv:2304.00216*, April 2023.

Timothy M D'Alfonso, David J Ho, Matthew G Hanna, Zhi Li, Hongyan Li, Limei Tang, Lei Zhang, Ziyang Li, Ruiyang Liu, Yiming Zheng, et al. Multi-magnification-based machine learning as an ancillary tool for the pathologic assessment of shaved margins for breast carcinoma lumpectomy specimens. *Mod Pathol*, 34:1487–1494, 2021. doi: 10.1038/s41379-021-00807-9.

Babak Ehteshami Bejnordi, Mitko Veta, Paul Johannes van Diest, and et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA*, 318(22):2199–2210, 2017. doi: 10.1001/jama.2017.14585.

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 *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18791–18801, 2022. doi: 10.1109/CVPR52688.2022.01825.

Ziyu Guo, Weiqin Zhao, Shujun Wang, and Lequan Yu. Higt: Hierarchical interaction graph-transformer for whole slide image analysis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 755–764. Springer, 2023.

Noriaki Hashimoto, Daisuke Fukushima, Ryoichi Koga, Yusuke Takagi, Kaho Ko, Kei Kohno, Masato Nakaguro, Shigeo Nakamura, Hidekata Hontani, and Ichiro Takeuchi. Multi-scale domain-adversarial multiple-instance cnn for cancer subtype classification with unannotated histopathological images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Wenjin Hou, Lu Yu, Chao Lin, Heng Huang, Rongtao Yu, Jing Qin, and Liang Wang. H^2-mil: Exploring hierarchical representation with heterogeneous multiple instance learning for whole slide image analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 933–941, 2022.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4700–4708. IEEE, 2017.

Zhi Huang, Federico Bianchi, Mert Yuksekgonul, Thomas Montine, and James Zou. Leveraging medical twitter to build a visual–language foundation model for pathology ai. *bioRxiv*, pp. 2023–03, 2023.

Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *International Conference on Machine Learning*, pp. 2123–2132. PMLR, 2018. URL http://proceedings.mlr.press/v80/ilse18a.html.

Andrew Janowczyk, Ren Zuo, Hannah Gilmore, Michael Feldman, and Anant Madabhushi. Histoqc: An open-source quality control tool for digital pathology slides. *JCO Clinical Cancer Informatics*, (3):1–7, 2019. doi: 10.1200/CCI.18.00157. URL https://doi.org/10.1200/CCI.18.00157. PMID: 30990737.

Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016. URL https://arxiv.org/abs/1609.02907.

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*, pp. 14318–14328, 2021.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10012–10022, 2021.

Michael Y. Lu, David F.K. Williamson, Tai-Yen Chen, and et al. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature Biomedical Engineering*, 5(6): 555–570, 2021. doi: 10.1038/s41551-020-00682-w.

Anamarija Morovic, Perry Damjanov, and Kyle Perry. *Pathology for the Health Professions-Sixth Edition*. Elsevier Health Sciences, 2021.

Nick Pawlowski, Saurabh Bhooshan, Nicolas Ballas, Francesco Ciompi, Ben Glocker, and Michal Drozdzal. Needles in haystacks: On classifying tiny objects in large images. *arXiv preprint arXiv:1908.06037*, 2019.

Maral Rasoolijaberi, Morteza Babaei, Abtin Riasatian, Sobhan Hemati, Parsa Ashrafi, Ricardo Gonzalez, and Hamid R. Tizhoosh. Multi-magnification image search in digital pathology. *IEEE Journal of Biomedical and Health Informatics*, 26(9):4611–4622, 2022. doi: 10.1109/JBHI.2022.3181531.

Abtin Riasatian, Morteza Babaie, Danial Maleki, Shivam Kalra, Mojtaba Valipour, Sobhan Hemati, Manit Zaveri, Amir Safarpoor, Sobhan Shafiei, Mehdi Afshari, Maral Rasoolijaberi, Milad Sikaroudi, Mohd Adnan, Sultaan Shah, Charles Choi, Savvas Damaskinos, Clinton JV Campbell, Phedias Diamandis, Liron Pantanowitz, Hany Kashani, Ali Ghodsi, and H. R. Tizhoosh. Fine-tuning and training of densenet for histopathology image representation using tcga diagnostic slides, 2021. URL https://arxiv.org/abs/2101.07903.

Yoni Schirris, Efstratios Gavves, Iris Nederlof, Hugo Mark Horlings, and Jonas Teuwen. Deepsmile: Contrastive self-supervised pre-training benefits msi and hrd classification directly from h&e whole-slide images in colorectal and breast cancer. *Medical Image Analysis*, 79:102464, 2022. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2022.102464. URL https://www.sciencedirect.com/science/article/pii/S1361841522001116.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

Zhuchen Shao, Hao Bian, Yang Chen, Yifeng Wang, Jian Zhang, Xiangyang Ji, and Yongbing Zhang. TransMIL: Transformer based correlated multiple instance learning for whole slide image classification. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=LKUfuWxajHc.

Kevin Thandiackal, Boqi Chen, Pushpak Pati, Guillaume Jaume, Drew F. K. Williamson, Maria Gabrani, and Orcun Goksel. Differentiable zooming for multiple instance learning on whole-slide images. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (eds.), *Computer Vision – ECCV 2022*, pp. 699–715, Cham, 2022. Springer Nature Switzerland. ISBN 978-3-031-19803-8.

Eugene Vorontsov, Alican Bozkurt, Adam Casson, George Shaikovski, Michal Zelechowski, Siqi Liu, Philippe Mathieu, Alexander van Eck, Donghun Lee, Julian Viret, et al. Virchow: A million-slide digital pathology foundation model. *arXiv preprint arXiv:2309.07778*, 2023.

Xiyue Wang, Sen Yang, Jun Zhang, Minghui Wang, Jing Zhang, Wei Yang, Junzhou Huang, and Xiao Han. Transformer-based unsupervised contrastive learning for histopathological image classification. *Medical image analysis*, 81:102559, 2022.

Xiyue Wang, Yuexi Du, Sen Yang, Jun Zhang, Minghui Wang, Jing Zhang, Wei Yang, Junzhou Huang, and Xiao Han. Retccl: Clustering-guided contrastive learning for whole-slide image retrieval. *Medical Image Analysis*, 83:102645, 2023. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2022.102645. URL https://www.sciencedirect.com/science/article/pii/S1361841522002730.

Muhammad Waqas, Syed Umaid Ahmed, Muhammad Atif Tahir, Jia Wu, and Rizwan Qureshi. Exploring multiple instance learning (mil): A brief survey. *Expert Systems with Applications*, 250: 123893, 2024. ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2024.123893. URL https://www.sciencedirect.com/science/article/pii/S0957417424007590.

Xinyi Wu, Amir Ajorlou, Zihui Wu, and Ali Jadbabaie. Demystifying oversmoothing in attention-based graph neural networks. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 35084–35106. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/6e4cdfdd909ea4e34bfc85a12774cba0-Paper-Conference.pdf.

Xingyuan Zhang, Hang Su, Lin Yang, and Shuicheng Zhang. Fine-grained histopathological image analysis via robust segmentation and large-scale retrieval. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5361–5368, Boston, MA, USA, 2015. doi: 10.1109/CVPR.2015.7299174.

Yu Zheng, Chen Gao, Liang Chen, Depeng Jin, and Yong Li. Dgcn: Diversified recommendation with graph convolutional networks. In *Proceedings of the Web Conference 2021*, WWW '21, pp. 401–412, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383127. doi: 10.1145/3442381.3449835. URL https://doi.org/10.1145/3442381.3449835.

Yanning Zhou, Simon Graham, Navid Alemi Koohbanani, Muhammad Shaban, Pheng-Ann Heng, and Nasir Rajpoot. Cgc-net: Cell graph convolutional network for grading of colorectal cancer histology images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, Oct 2019.