Dynamic Entity-Masked Graph Diffusion Model for Histopathological Image Representation Learning

Zhenfeng Zhuang[1*], Min Cen[2*], Yanfeng Li[1*], Fangyu Zhou[1], Lequan Yu[3], Baptiste Magnier[4,5], Liansheng Wang[1†]

[1]Department of Computer Science at School of Informatics, Xiamen University, Xiamen, China
[2]School of Artificial Intelligence and Data Science, University of Science and Technology of China, Hefei, Anhui, China
[3]School of Computing and Data Science, The University of Hong Kong, Hong Kong SAR, China
[4]EuroMov Digital Health in Motion, Univ Montpellier, IMT Mines Ales, Ales, France
[5]Service de Medecine Nucleaire, Centre Hospitalier Universitaire de Nimes, Universite de Montpellier, Nimes, France
{zhuangzhenfeng, liyanfeng, lswang}@stu.xmu.edu.cn, cenmin0127@mail.ustc.edu.cn

Abstract

Significant disparities between the features of natural images and those inherent to histopathological images make it challenging to directly apply and transfer pre-trained models from natural images to histopathology tasks. Moreover, the frequent lack of annotations in histopathology patch images has driven researchers to explore self-supervised learning methods like mask reconstruction for learning representations from large amounts of unlabeled data. Crucially, previous mask-based efforts in self-supervised learning have often overlooked the spatial interactions among entities, which are essential for constructing accurate representations of pathological entities. To address these challenges, constructing graphs of entities is a promising approach. In addition, the diffusion reconstruction strategy has recently shown superior performance through its random intensity noise addition technique to enhance the robust learned representation. Therefore, we introduce **H-MGDM**, a novel self-supervised **H**istopathology image representation learning method through the Dynamic Entity-**M**asked **G**raph **D**iffusion **M**odel. Specifically, we propose to use complementary subgraphs as latent diffusion conditions and self-supervised targets respectively during pre-training. We note that the graph can embed entities' topological relationships and enhance representation. Dynamic conditions and targets can improve pathological fine reconstruction. Our model has conducted pretraining experiments on three large histopathological datasets. The advanced predictive performance and interpretability of H-MGDM are clearly evaluated on comprehensive downstream tasks such as classification and survival analysis on six datasets.

Code — https://github.com/centurion-crawler/H-MGDM

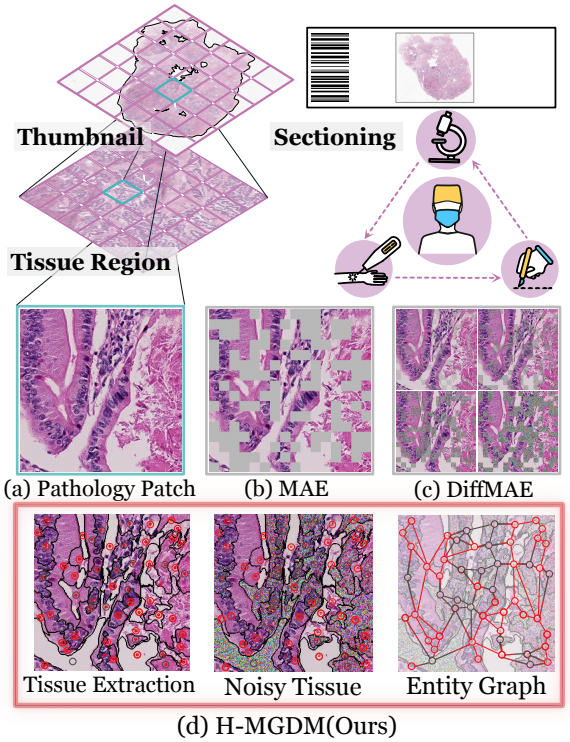

Figure 1: Pathological slide inspection process from the overall view to details. Unlike comparison methods, H-MGDM focuses on masked pathological tissue regions rather than grid tiles in patches, constructing the masked subgraph with varying intensities and noise for reconstruction by complementary conditional subgraph.

Introduction

Achieving concise and informative histopathology patch image representation is the cornerstone for solving many tasks in the field of computational histopathology analysis, such as cancer diagnosis, grading, segmentation, and prognosis tasks within whole Slide Image (WSI) (Panayides et al. 2020). There are many extensive investigations on pre-training features for pathology images, which can help alleviate the time-consuming and tedious task of manual slide inspection by pathologists (Song et al. 2023). Today, the representation of pathology image patches relies heavily on transfer learning methods (Dosovitskiy et al. 2020; Sharmay

*These authors contributed equally.
†Corresponding author.

et al. 2021; Li et al. 2022) and the supervised pathology classification models such as KimiaNet (Riasatian et al. 2021). The above approaches exist problems like the domain gap and category bias (Guan and Liu 2021), and scarce and high-cost annotations limit retraining. Therefore, unlabeled self-supervised learning (SSL) has emerged as a solution to alleviate these limitations by learning salient representations without using labels.

Previous SSL methods like MAE (He et al. 2022) have shown that using masks and reconstruction tasks in SSL effectively enables models to learn from unlabeled pathological data. In the context of pathological images, the topological connections among pathological tissues, including cellular interactions and their surrounding environment, are crucial for various tasks. Recent approaches have underscored the importance of structure function relationships by linking the spatial organization of cells within tissues through cell graphs. These methods enable the extraction of biomarker-based pathological features (Jaume et al. 2021b,a), capturing complex semantic associations that extend beyond pixel-level data to encompass tissues and cells. These approaches align more closely with pathological diagnostic procedures. Entity-based topological analysis provides enhanced control over tissue modeling and facilitates the integration of pathological priors into task-specific histopathological entity representations. This indicates that it is crucial to recognize the interactions among "Image-to-Graph" based pathological tissue entities. However, when SSL is applied to pathological images, recent studies often focus on using pathological grid tiles in patches as masks, while neglecting the impact of entities (e.g., cellular or tissue regions) mask strategy on the overall semantic representation. Therefore, we introduce a new method to convert images into graphics to capture the structure of pathological entities, such as tissues, in self-supervised learning.

On the other hand, diffusion strategies have recently improved robust learning representations as conditions through their technique of adding noise with varying intensities (Purma et al. 2023; Wei et al. 2023; Yang and Wang 2023). We propose using an entity-masked graph as the input for the diffusion process, with encoded features from different layers of the graph serving as conditions to maintain strong performance. This approach captures powerful and complex entity information, thereby enhancing the representation of pathological images.

In summary, the motivation of our paper is to better learn the knowledge of entity graphs under self-supervised reconstruction progress. We propose a novel approach that converts histopathological images into entity graphs with dynamic mask and noise for diffusion pre-training to obtain better pathological image representations in the pathology inspection process (see Fig.1). Our contributions are:

- We propose a novel framework the **H-MGDM, a novel self-supervised histopathological image representation learning method through Dynamic Entity-Masked Graph Diffusion Model**. A strategy for partially visible entities as conditioning to prompt masked noisy entities to graph diffusion. Random masks and dynamic intensity noises can enhance representations in histopathological images.
- In H-MGDM, we **convert pathology images into entity graphs of latent space to incorporate structural information of pathological entities**. This allows for more comprehensive spatial and semantic priors.
- We conducted pretraining experiments on three large histopathological datasets. The advanced predictive performance and interpretability of H-MGDM are clearly evaluated on comprehensive downstream tasks on six datasets. All performance **across several downstream tasks is averagely improved by 5.18%**. This demonstrates the effectiveness of the H-MGDM framework for pre-training in histopathological image analysis.

Related Works

Graph Representation for Digital Pathology

Recently, graph neural networks (GNN) have been employed to represent patches as graphs for pathology tasks. CGC-Net (Zhou et al. 2019) introduces a cell-graph convolutional neural network that converts large histology images into graphs, where vertices represent nuclei and edges denote cellular interactions based on vertex similarity. HACT (Pati et al. 2022) develops a hierarchical cell-to-tissue graph representation to jointly model both low-level and high-level graphs, incorporating intra- and inter-level coupling based on the topological distributions and interactions among entities. SHGNN (Hou et al. 2022a) proposes a novel spatial hierarchical GNN framework, equipped with a dynamic structure learning module, to capture entity location attributes and semantic representations, thereby extracting the characteristics of different entities in images. However, these methods focus supervision on global representation. Our proposed method introduces entities subgraph self-supervised targets. This enables entities to capture the contextual implications of local information.

Denoising Diffusion Models

Diffusion models (Rombach et al. 2022; Ho, Jain, and Abbeel 2020) are known for their ability to generate sophisticated images under conditional control. Diffusion models also exhibit variable masking capabilities and have also been used to enhance representation learning in the self-supervised learning domain, in learning paradigms such as DiffAE (Preechakul et al. 2022). Also, GenSelfDiff-HIS (Purma et al. 2023) proposes a diffusion-based generative pre-training process for self-supervision to learn efficient histopathological image representations. DiffMAE (Wei et al. 2023) integrates diffusion's nuanced detail reconstruction capabilities with MAE's comprehensive semantic representation capability, symbolizing a convergence of methodologies in pursuit of enhanced representation.

Mask for Self-Supervised Representation

Since the introduction of BERT (Devlin et al. 2018) in language models, masking prediction has reattracted attention. Lots of self-supervised masking methods have been

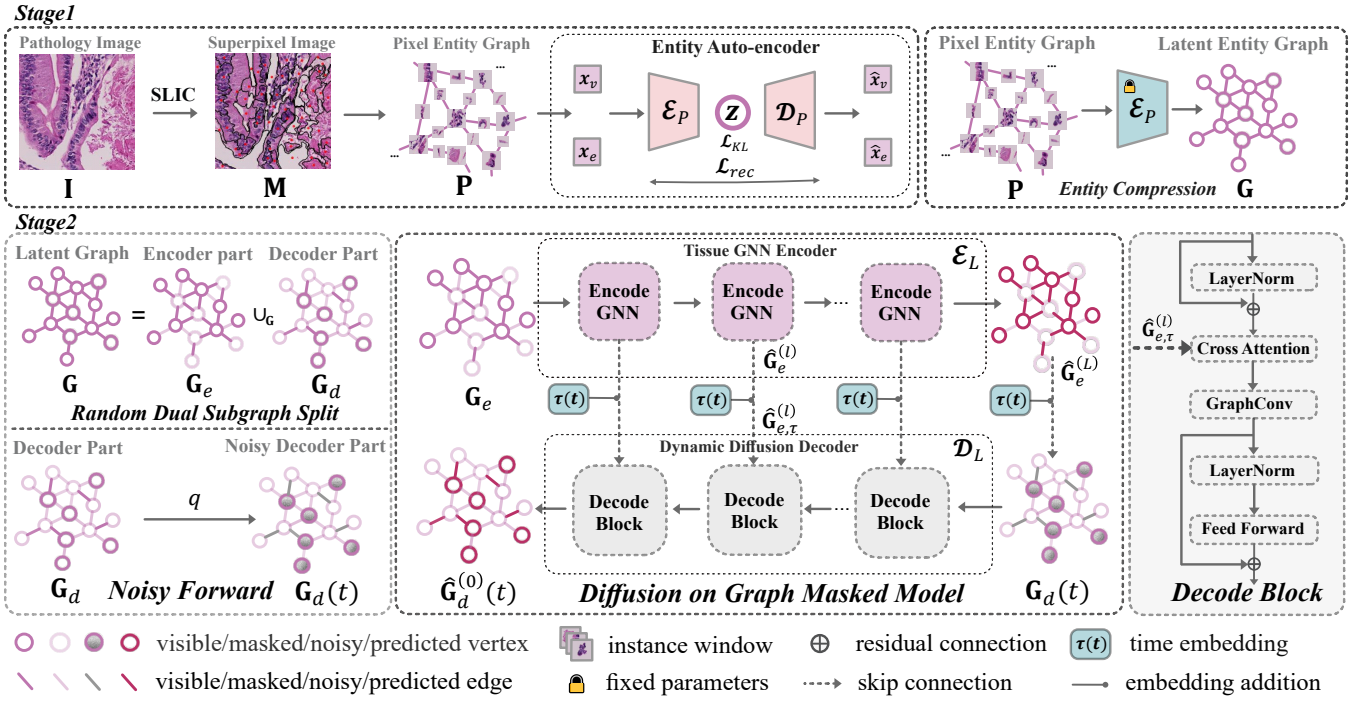

Figure 2: Overview of the H-MGDM pretraining stages. Conditional diffusion reverse process in the decoder. $\mathbf{G}_e$ and $\mathbf{G}_d$ are two complementary subgraphs of $\mathbf{G}$. $\mathbf{G}_d(t)$ are from the diffusion forward process q_L of $\mathbf{G}_d$. The target is to denoise $\mathbf{G}_d(t)$ to $\hat{\mathbf{G}}_d^{(0)}(t)$ close to $\mathbf{G}_d$ at sampling time t.

proposed in vision tasks, delineating enhanced representation techniques rooted in various theoretical frameworks: MAE (He et al. 2022) devised an asymmetric architecture tailored for pixel-level reconstruction, while MixMIM (Liu et al. 2022) endeavors to narrow the chasm between pre-training and fine-tuning through the stochastic amalgamation of masks. Additionally, GraphMAE (Hou et al. 2022b) and GraphMAE2 (Hou et al. 2023) advocate employing masked graph convolution to facilitate feature reconstruction, but the mentioned methods are confined to fixed-intensity masking approaches. The variation mechanism forces the features to adapt to each situation.

Methodology

Figure 2 illustrates the framework of H-MGDM. First, the histopathological entity graph is constructed using the superpixel algorithm SLIC (Achanta et al. 2012) to extract the topological relations among tissues in the image and compress entities to latent space. Then, the GNN encoder is used to encode the visible subgraph as a condition. The latent graph diffusion model is introduced to reconstruct the dynamic self-supervised target of the masked subgraph to obtain robust representations of patches.

Pathological Entity Graph Construction

To strengthen the concept of entity within limited structural constraints, we utilize priori pathological tissue superpixels as tissue entities when constructing the graph. First, a pathological image $\mathbf{I} \in \mathbb{N}^{h_\mathbf{I} \times w_\mathbf{I} \times 3}$ with the height $h_\mathbf{I}$ and the width $w_\mathbf{I}$ is partitioned via SLIC algorithms (Achanta et al. 2012) resulting in a set of superpixels. For each superpixel, s, a window of size $a \times a$ centered at s is considered as a vertex v of the pathological entity graph $\mathbf{P}$ in pixel space. Pixels in the window that do not belong to s are assigned the background color. Subsequently, edges will be established between every two vertices with adjacent boundaries, considering the local interactions between adjacent vertices more. The edge e originates from the region after the expansion operation along the boundary between s^i and another neighboring superpixel s^j. Thus, the image $\mathbf{I}$ is transformed into a pathological entity graph $\mathbf{P}(\mathbf{V_P}, \mathbf{E_P}, \mathbf{A}, \mathbf{D})$, where $\mathbf{V_P} = \{s_i\}_{i \in [0, N_\mathbf{V})}$, $\mathbf{E_P} = \{e_j\}_{j \in [0, N_\mathbf{E})}$ are sets of vertices and edges, respectively. And the adjacency matrix is $\mathbf{A}$ and the degree matrix of $\mathbf{A}$ is $\mathbf{D}$. $N_\mathbf{V}$ and $N_\mathbf{E}$ are the numbers of vertices and edges, respectively.

Entity Compression into Latent Space

Our compression model, based on previous work (Kingma and Welling 2013; Esser, Rombach, and Ommer 2021), utilizes an auto-encoder in stage 1. Given a pathological entity graph $\mathbf{P}$, the encoder $\mathcal{E}_\mathbf{P}$ transforms each entity $\boldsymbol{x} \in (\boldsymbol{X}_v \cup \boldsymbol{X}_e) \subset \mathbb{N}^{a \times a \times 3}$ in $\mathbf{P}$ into a latent representation $\boldsymbol{z} = \mathcal{E}_\mathbf{P}(\boldsymbol{x})$, where $\boldsymbol{z} \in \mathbb{R}^{l \times l \times c}$. The encoder learns an approximate posterior distriution $q(z|x) = \mathcal{N}(z; \mu(x), \sigma^2(x)I)$, with $\mu(x)$ and $\sigma(x)$ being learned mean and standard deviation of x. And the decoder $\mathcal{D}_\mathbf{P}$ then reconstructs the image from this latent space, $\hat{\boldsymbol{x}} = \mathcal{D}_\mathbf{P}(\boldsymbol{z}) = \mathcal{D}_\mathbf{P}(\mathcal{E}_\mathbf{P}(\boldsymbol{x}))$. The downsampling factor of the image is $f = a/l$ and we ex-

plore various downsampling factors f. After the first stage of training is completed, we infer $\mathcal{E}_{\mathbf{P}}$ to encode all entities into the latent space, resulting in sets $\mathbf{V_G}$ and $\mathbf{E_G}$. Those compose the latent space entity graph $\mathbf{G}(\mathbf{V_G}, \mathbf{E_G}, \mathbf{A}, \mathbf{D})$.

Latent Graph Diffusion Model

The forward process of the latent diffusion model can add noise to the graph entities, describing the degraded sequence caused by Gaussian noise on the latent space, which does not contain much semantics. Given a well-defined latent diffusion diffusion (Ho, Jain, and Abbeel 2020) forward process $q_L : \{\mathbf{G}_d(t)\}_{[0,T]}$ with variance time dependence, and a noise schedule $\{\beta(t)\}_{[0,T]}$ where the integer time $t \in [0, T]$, based on Markov chain and diffusion characteristics(Huang et al. 2023; Wei et al. 2023), we have:

$$\begin{cases} q_L(\mathbf{G}_d(t)|\mathbf{G}_d(t-1)) = \mathcal{N}(\mathbf{G}_d(t)|\sqrt{1-\beta(t)}\mathbf{G}_d(t-1), \beta(t)\boldsymbol{I}) \\ q_L(\mathbf{G}_d(t)|\mathbf{G}_d(0)) = \mathcal{N}(\mathbf{G}_d(t)|\sqrt{\overline{\alpha}(t)}\mathbf{G}_d(0), (1-\overline{\alpha}(t))\boldsymbol{I}) \end{cases}. \tag{1}$$

$\mathbf{G}_d(t)$ is reparameterized as $\mathbf{G}_d(t) = \sqrt{\overline{\alpha}(t)}\mathbf{G}_d(0) + \sqrt{1-\overline{\alpha}(t)}\epsilon$, noise follows a normal distribution: $\epsilon \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$, where $\alpha(t) = 1-\beta(t)$, $\overline{\alpha}(t) = \prod_{i=1}^{t}\alpha(i)$, signal-to-noise ratio $\{\frac{\alpha(t)}{\beta(t)}\}_{[0,T]}$ are chosen to noise gradually that $q_L(\mathbf{G}_d(T)) \approx \mathcal{N}(\mathbf{0}, \boldsymbol{I})$. The conditional diffusion reverse above by modeling the reverse distribution p_L which implies masked part conditioned on visible graph $\hat{\mathbf{G}}_e$ from encoder:

$$p_L(\mathbf{G}_d(t-1)|\mathbf{G}_d(t), \hat{\mathbf{G}}_e) \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}). \tag{2}$$

Then, a reverse diffusion network $\mathcal{D}_L$ with graph conditioning on $\{\beta(t)\}_{[0,T]}$ is introduced. Considering the graph structure, we can apply the continuous diffusion process to $\mathbf{V}_d(t)$ and $\mathbf{E}_d(t)$ respectively to facilitate the restoration of noisy latent within subgraph pathological entities.

Dynamic Diffusion on Masked Graph Model

In stage 2 illustrated in Fig. 2, an asymmetric auto-encoder mode is employed, utilizing GNN layers (Kipf and Welling 2016) as the encoder and ViT variants (Dosovitskiy et al. 2020) as the decoder. From the LDM perspective, the encoder also provides encoding conditions for the decoder's denoising process. For a graph input G, is dynamically randomly divided into two complementary subgraphs $\mathbf{G}_e(\mathbf{V}_e, \mathbf{E}_e, \mathbf{A}_e, \mathbf{D}_e)$ for encoder and $\mathbf{G}_d(\mathbf{V}_d, \mathbf{E}_d, \mathbf{A}_d, \mathbf{D}_d)$ for decoder according to the given masking ratio $r_m = \frac{N_{\mathbf{V}_d}}{N_{\mathbf{V}}} = \frac{N_{\mathbf{E}_d}}{\mathbf{N}_E}$. The noise-added $\mathbf{G}_d(t)$ serves as the initial input to the decoder. Furthermore, the topological information $\mathbf{A}_* \ \mathbf{D}_*(* = e, d)$ is maintained during training.

Tissue GNN Encoder The latent encoder $\mathcal{E}_L$ employs GNN that integrates tissue vertices and edges for L layers. In our pathological graph-based construction, the latent domains of graph vertices and edges are identical. Therefore, the vertex-based message passing can be used to forward edge latent for $\hat{\mathbf{G}}_e^{(l)}(V_e^{(l)}, E_e^{(l)}, A_e, D_e)$ in the l-th layer:

$$\mathcal{E}_L^{(l)} : \begin{cases} \mathbf{V}_e^{(l+1)} = \sigma(\tilde{\mathbf{A}}\mathbf{V}_e^{(l)}\mathbf{W}_{\mathbf{V}}^{(l)}) \\ \mathbf{E}_e^{(l+1)} = \sigma(\tilde{\mathbf{A}}^*\mathbf{E}_e^{(l)}\mathbf{W}_{\mathbf{E}}^{(l)}) \end{cases}, \tag{3}$$

where $\tilde{\mathbf{A}} = \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}$. $\tilde{\mathbf{A}}$ and $\tilde{\mathbf{A}}^*$ are the normalized symmetric adjacency matrices of the graph $\mathbf{G}_e$ and the dual graph $\mathbf{G}_e^*$, respectively. $\mathbf{V}_e^{(l)}, \mathbf{E}_e^{(l)}$ are inputs to the l-th layer, $\mathbf{W}_{\mathbf{V}}^{(l)}, \mathbf{W}_{\mathbf{E}}^{(l)}$ are the vertices and edges weight matrices of graph convolution $\mathcal{E}_L^{(l)}$, and $\sigma(\cdot)$ is a non-linear activation.

Dynamic Diffusion Decoder The decoder $\mathcal{D}_L$ utilizes the conditional latent graph diffusion model. The noise level t serves as the forward sampling time during pre-training to generate $\mathbf{G}_d(t)$. Similar to the Transformer architecture (Vaswani et al. 2017), in the l-thdecode block, cross-attention $CA(\cdot,\cdot)$ utilizes the visible latent $\hat{\mathbf{G}}_e^{(l)}$ from the l-th encoder layers as the conditional control after adding the time embedding $\tau(t)$: $\hat{\mathbf{G}}_{e,\tau}^{(l)} = \hat{\mathbf{G}}_e^{(l)} + \tau(t)$, aiding in denoising $\hat{\mathbf{G}}_d^{(l)}(t)$ during decoding. And the graph convolution of decoder $C_d^{(l)}$ is to perform the message passing of the fused latent according to A_d next. The Feed Forward $\overline{FF}(\cdot)$ with layer normalized residual block is employed as the final layer within the Decoder Block to induce feature activation, resulting in $\hat{\mathbf{G}}_d^{(l-1)}(t)$. Then, it proceeds to the subsequent blocks to predict $\hat{\mathbf{G}}_d^{(0)}(t)$:

$$\begin{cases} \hat{\mathbf{G}}_d^{(l-1)}(t) = \overline{FF}(C_{dec}(CA(\hat{\mathbf{G}}_d^{(l)}(t), \hat{\mathbf{G}}_{e,\tau}^{(l)})) \\ \hat{\mathbf{G}}_d^{(0)}(t) = \mathcal{D}_L(\mathbf{G}_d(t), t, \{\hat{\mathbf{G}}_e^{(l)}\}_{l\in[0,L)}) \end{cases}. \tag{4}$$

The skip connection from $\hat{\mathbf{G}}_{e,\tau}^{(l)}$ forms a U-shaped configuration, typically advantageous for graph representation across different levels and noisy latent restoration.

Training Strategy

Objectives The optimization of the model in pre-taining has two stages: In order to avoid high-variance latent space in the first stage, the auto-encoder is optimized by a combination of a reconstruction loss and KL Divergence D_{KL} like VAE (Kingma and Welling 2013). The pixel-space reconstruction constraint $\mathcal{L}_{rec}$ enforces local realism avoids blurriness and ensures that the reconstructions are confined to the image manifold. Here we use the MSE form:

$$\mathcal{L}_{rec} = \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})}\|\boldsymbol{x} - \hat{\boldsymbol{x}}\|_2, \ \mathcal{L}_{VAE} = \mathcal{L}_{rec} + \lambda D_{KL}, \tag{5}$$

where $\mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})}$ represents the expectation of the distribution of the latent variable z, λ is the loss weight. And $q(\boldsymbol{z}|\boldsymbol{x})$ is the posterior distribution of the latent $\boldsymbol{z}$ given $\boldsymbol{x}$.

The second stage is guided by minimizing the following objective function to optimize parameters θ of the model. Notably, $\boldsymbol{x}_0$-mode is used dynamic denoising on graph masking of vertices and edges:

$$\mathcal{L}_{Simple} = \mathbb{E}_{t,\theta,\epsilon}[\|\mathbf{V}_d(0) - \hat{\mathbf{V}}_d^{(0)}(t)\| + \|\mathbf{E}_d(0) - \hat{\mathbf{E}}_d^{(0)}(t)\|]. \tag{6}$$

Here, ϵ represents sampled noise. Leveraging variational inference, the well-known Mean Square Error (MSE) objective, derived from the evidence lower bound, is utilized to predict the denoised $\hat{\mathbf{V}}_d^{(0)}$ and $\hat{\mathbf{E}}_d^{(0)}$ as reconstruction targets.

Cancer Subtype / Tissue Classification										
Strategies			**Datasets** / **Methods**	Komura et al.		PANDA		IBD	
Graph	Mask	Diffusion		ACC(%)	F1(%)	ACC(%)	F1(%)	ACC(%)	F1(%)
✗	✗	✗	SimCLR (Chen et al. 2020)	69.24±2.12	62.45±1.72	61.10±2.19	58.48±1.25	71.44±1.54	69.47±2.43
✗	✗	✗	KimiaNet* (Riasatian et al. 2021)	72.66±1.69	65.93±1.75	67.68±1.88	55.40±1.19	76.42±0.98	70.75±1.37
✗	✔	✗	Dino (Caron et al. 2021)	78.05±1.36	70.64±2.01	69.92±1.28	64.11±1.31	82.45±0.82	75.23±1.41
✗	✔	✗	MAE (He et al. 2022)	77.37±1.51	69.44±0.89	70.51±1.78	62.73±0.89	81.06±0.88	76.44±1.02
✔	✔	✗	GraphMAE (Hou et al. 2022b)	76.69±2.10	67.60±1.87	72.22±2.19	60.34±2.55	75.85±1.25	72.78±0.81
✔	✔	✗	GraphMAE2 (Hou et al. 2023)	78.86±3.05	65.97±1.05	72.56±1.80	63.49±2.64	78.12±1.42	72.32±0.79
✗	✗	✔	DiffAE (Preechakul et al. 2022)	79.11±1.92	68.18±1.74	71.54±2.12	61.62±1.11	81.24±1.48	74.51±2.08
✗	✔	✔	DiffMAE (Wei et al. 2023)	78.14±2.10	70.23±2.14	71.92±1.22	65.82±1.82	84.58±1.72	74.74±2.15
✔	✔	✔	**H-MGDM (Ours)**	**82.06±1.17**	**72.41±1.36**	**74.51±1.13**	**67.32±1.40**	**86.23±2.31**	**78.92±1.79**
Ablation study			w/o edge latents	80.91±1.92	70.85±1.58	72.39±1.29	65.89±1.82	84.11±1.72	77.41±1.54
			w/o skip connection	79.29±1.27	68.43±1.45	70.98±1.75	62.11±2.16	82.35±2.32	74.61±2.14
			noise intensity fixed	79.70±1.14	67.14±1.24	72.52±1.65	63.73±2.31	83.58±1.69	77.24±2.53

Table 1: Comparing classification performance across methods with ablations. The best results are marked in **bold**. "*" indicates label-supervised of the method. The results are reported in: *mean±std* (where *std* stands for standard deviation).

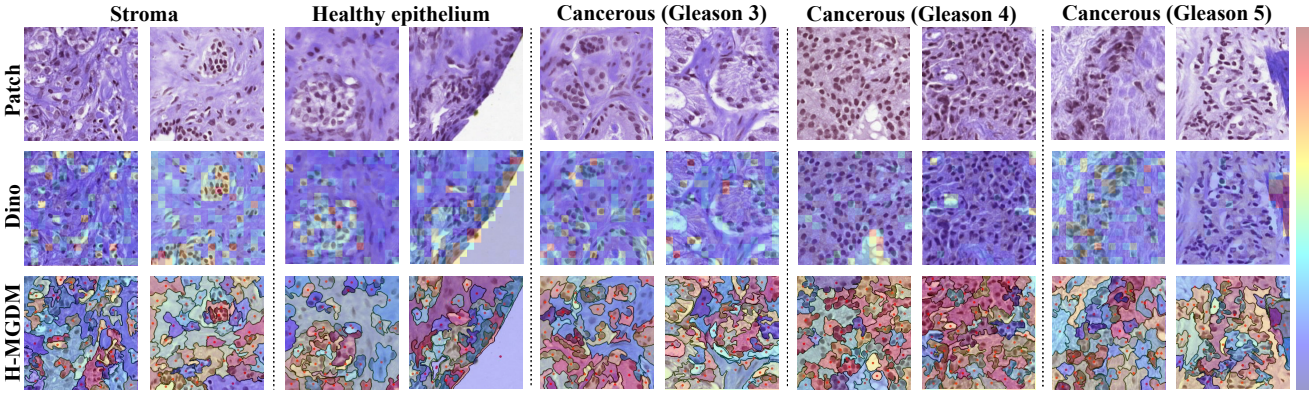

Figure 3: Original images and their attention heatmaps of five different categories of the PANDA dataset, showing the interpretability of our method under the pathological entity graph construction comparison with Dino.

Downstream Tasks For downstream tasks, we deploy two stages' encoders $\mathcal{E}_{\mathbf{P}}$ $\mathcal{E}_L$ to inference to obtain the global graph representation $\mathbf{O_G}$ by readout $r_{\mathbf{G}}$:

$$\mathbf{G}_o = \hat{\mathbf{G}}^{(L)} = \mathcal{E}_L(\mathcal{E}_{\mathbf{P}}(\mathbf{G_P})), \mathbf{O_G} = r_{\mathbf{G}}(\mathbf{G}_o). \quad (7)$$

- **Classification.** For the downstream tuning, the cross-entropy loss $\mathcal{L}_{CE}$ is adopted for patch classification tasks. The model is optimized through the classification layer MLP_c to obtain the predicted $\hat{Y}$ by global representation $\mathbf{O_G}$ as probabilities of the classes and to supervise it with the classification labels Y.
- **Regression.** We introduce the Cox proportional hazards model, a semi-parametric regression model. Using survival events and survival time as labels. Here, pathological image features are used to predict risks and analyze the impact on survival. Considering a problem involving two explanatory variables as predictors of survival time t_i and t_j of patients i, j, δ represents the termination event (1: death, recurrence, 0: not occurred) $R(t_i)$ represents the condition $t_j > t_i$, for neg log partial likelihood loss:

$$\mathcal{L}_{Cox} = -\sum_{i:\delta_i=1}[\hat{h}_i - \log \sum_{j \in R(t_i)} e^{\hat{h}_j}], \quad (8)$$

where we use MLP_h linear map the graph representation $\mathbf{O_G}_i$ to the final hazards prediction h_i.

Experiments

Experimental Settings

Datasets The proposed framework underwent pretraining and classification using three large extensive histopathology datasets: **Komura et al.** (Komura et al. 2022) (1.6M images, 32 cancer types), **Prostate ANnotation Data Archive (PANDA)** dataset (Bulten et al. 2022) (11,000 digitized H&E-stained WSIs, obtaining 12.5M images with 6 level Gleason region annotations), our in-house colorectal cancer data **IBD** (23M images, with 360K patches annotated into 9 common tissue types). For the survival analysis task, we compare the performances on two public cohorts: **TCGA-KIRC** (512 cases) and **TCGA-ESCA** (155 cases) and one privately collected primary-metastatic pathology colorectal cancer dataset **CRC-PM** (388 cases). We use different methods pre-trained on the pancancer dataset Komura et al. as feature extractors for patches in WSI.

Survival Analysis Regression									
Datasets / Methods	TCGA-KIRC			TCGA-ESCA			CRC-PM		
	DeepSurv	AB-MIL	PatchGCN	DeepSurv	AB-MIL	PatchGCN	DeepSurv	AB-MIL	PatchGCN
SimCLR (Chen et al. 2020)	61.91±4.71	62.09±4.33	62.46±4.26	59.26±4.35	58.06±4.60	62.59±4.82	56.30±2.47	58.76±6.51	59.87±4.06
KimiaNet* (Riasatian et al. 2021)	62.16±4.72	64.12±5.28	65.76±3.14	60.53±6.69	59.00±2.97	58.16±5.75	59.06±9.04	59.95±6.11	57.67±8.89
Dino (Caron et al. 2021)	67.92±5.61	66.94±4.42	66.64±3.01	57.88±3.91	58.85±4.14	56.58±5.11	58.02±6.72	61.01±9.91	63.20±8.50
MAE (He et al. 2022)	57.78±2.42	61.30±2.54	65.08±3.69	59.71±5.02	57.84±6.29	60.62±5.21	60.92±7.63	59.44±8.98	62.80±8.82
GraphMAE2 (Hou et al. 2023)	64.31±6.73	66.76±4.21	65.84±2.04	57.26±5.55	59.71±4.90	60.35±5.55	58.16±6.29	59.38±6.94	60.42±7.08
DiffAE (Preechakul et al. 2022)	63.26±3.88	67.31±1.98	68.29±3.84	60.74±5.54	60.23±5.74	63.61±5.49	60.81±6.30	60.61±5.80	62.87±9.52
DiffMAE (Wei et al. 2023)	62.41±3.24	65.92±5.04	67.32±4.18	60.60±2.57	59.49±4.83	63.89±6.32	61.16±5.27	60.64±6.05	61.29±7.12
H-MGDM (Ours)	**66.99±4.56**	**69.88±3.97**	**71.17±4.51**	**62.68±3.04**	**62.55±3.28**	**64.82±3.45**	62.27±5.09	**63.89±6.94**	**66.05±7.81**
w/o edge latent	65.91±5.74	67.88±4.73	70.53±5.01	59.23±4.29	60.62±4.29	64.10±4.83	**62.32±6.12**	61.39±7.12	65.12±7.28
w/o skip connection	63.64±5.32	67.00±3.10	69.07±5.62	58.66±3.85	60.34±3.24	59.40±2.36	59.78±2.91	60.09±4.57	60.97±8.36
noise intensity t fixed	63.16±3.62	69.44±4.50	69.99±4.45	59.17±6.49	58.97±3.97	61.17±3.80	60.34±7.38	62.05±8.69	62.33±5.29

Table 2: Survival Analysis performance across SOTAs features on external public validation data by conducting the 5-fold evaluation procedure with 5 runs. The experimental results of CI are reported by *mean±std*.

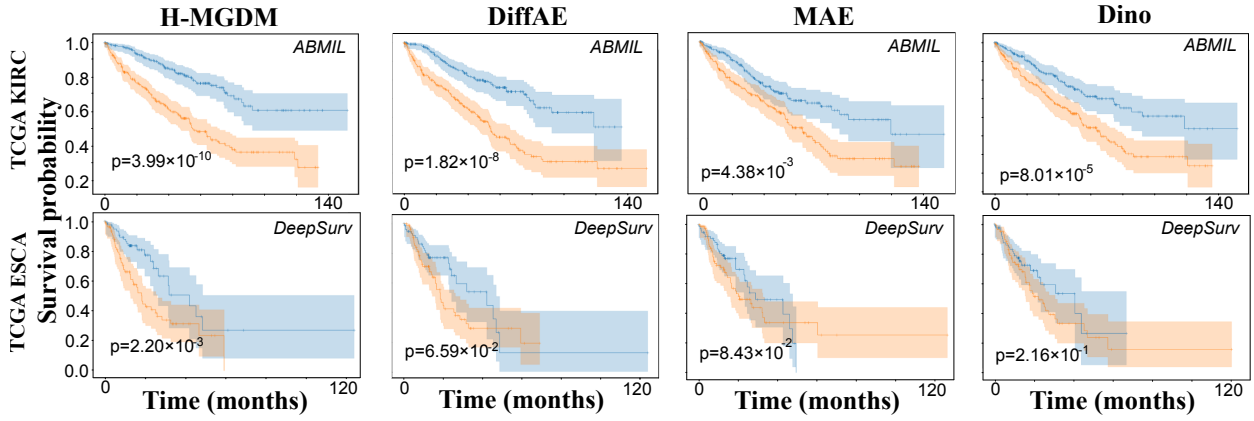

Figure 4: Kaplan-Meier Analysis of comparison methods and our framework. All patients from the five tests were pooled and analyzed. Each cohort is split into a high-risk (orange) and a low-risk group (blue) according to the median output of the cohort.

Implementation Details Experiments involving H-MGDM and comparative methods are conducted with a batch size set to 64 max. SLIC with the initial region number of 500, window size a is 64. Latent downsampling factors $f = 2$. Also, pre-training is optimized by Adam(Kingma and Ba 2014) with an initial learning rate of 3e-4 and the plateau scheduler with a minimum learning rate of 1e-5 for 250 epochs. The noise timesteps T is 1000, and the sigmoid schedule $\{\beta(t)\}_{[0,T]}$ from 1e-7 to 2e-3 is used.

Evaluation Metrics Accuracy (**ACC**) and Marco **F1** are used to evaluate the methods in classification comparison. To evaluate the effectiveness of entity latent diffusion, the Root Mean Square Error (**RMSE**) for graph entities is calculated as a qualitative evaluation metric. Harrell's concordance index (**C-index** or **CI**) measures the survival model's ability to correctly provide a reliable ranking of the survival times based on the individual risk scores. It ranges from 0 to 1, with higher values indicating better performance.

Comparison with Baseline Methods

Results We conduct comparative experiments among our H-MGDM and other baseline pre-training models: *i.e.* SimCLR (Chen et al. 2020), Dino, MAE, GraphMAE, GraphMAE2, DiffAE, DiffMAE. We also mark the use of strategies (graph construction, mask strategy, diffusion-guided) by the comparison methods in Table 1. For comparison of the features from various pre-training methods, we use three backbones: DeepSurv (Katzman et al. 2018), AB-MIL (Ilse, Tomczak, and Welling 2018) and PatchGCN (Chen et al. 2021) for the survival prediction task. To the best of our knowledge, H-MGDM is the first time these three mechanisms have been introduced simultaneously into histopathology pre-training. Our method achieved outstanding performance due to incorporating structural information of pathological entities and enhancing mask learning during the diffusion procedure. The results are presented in Table 1, 2. For all baselines, our method achieves an average improvement of 5.99%, 5.43%, and 4.146% on the ACC, F1, and CI.

Ablation Study To further explore the proposed components with the effectiveness of our model. We perform ablation experiments in Table 1, 2. The study investigates the influence of edges in graph data, skip connections from encoder representations as diffusion conditioning, and time-varying intensity noise on model performance. The results

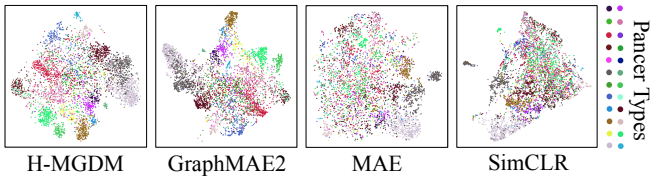

Figure 5: T-SNE plots of pan-cancer samples' readout representations learning with H-MGDM and baseline methods.

show these components can enhance the representational capacity of downstream tasks with improvement of 2.50%, 3.17%, and 2.46% on the ACC, F1, and CI metrics.

Kaplan-Meier Analysis and Significance Based on the median survival risk output by our model, each cancer cohort is divided into high-risk and low-risk groups. If the survival predictions are consistent, the Kaplan-Meier (KM) curves of these groups should show significant differences. As expected, Fig. 4 shows our method's log-rank p-values less than 0.05 for two different cancer cohorts, indicating the effectiveness of in predicting patient survival.

T-SNE Visualization of pan-cancer representations To further confirm the performances of pre-training models, we randomly select 6.4K samples from the trained pre-training model from the Komura et al. (Komura et al. 2022) data test set and visualize the distribution of t-SNE (Hinton and Roweis 2002) as shown in Fig. 5. Our H-MGDM model can better distinguish several pan-cancer types. They have high intra-class aggregation and inter-class separation.

Analysis of Our Framework

Interpretability The graphical representation of our approach provides scalable interpretability. Unlike posterior explanations of previous methods like Grad-CAM (Selvaraju et al. 2017) and Shap (Štrumbelj and Kononenko 2014), H-MGDM can directly generate high-attention regions during inference readout. We visualize the attention heatmap in Fig. 3 of tissue regions sampling from normal to pathological classes. We also visualize the Dino method's multi-head attention for comparison, highlighting its weaker resistance to interference (some Gleason 4 and 5 samples are focused on blank and outlined regions).

Masking ratio Investigation We examine rates of the dynamic masking strategy, which correspond to the ratio of the target subgraph division. As shown in Table 3, the entity graph masking rate r_m is about 50%-70% for reconstruction, and the rest is used as condition learning. Such pre-training features can improve the classification performance of the

Mask Ratio	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Komura et.al.	65.9	69.9	74.4	77.1	78.4	**82.1**	79.2	76.2	73.9
PANDA	63.1	63.2	67.9	65.4	70.5	72.5	**74.5**	70.5	66.2
IBD	70.5	75.2	79.5	84.5	**86.2**	82.7	79.2	74.2	64.3

Table 3: Hyperparameter Investigation of mask ratios

Strategy	NtoN	EtoE	NtoE	EtoN	NtoN & EtoE	NtoE & EtoN
ACC	73.74	71.59	71.98	68.59	**74.51**	73.82
RMSE	0.136	0.155	0.169	0.184	0.141	**0.127**

Table 4: Effectiveness of various decoder conditioning strategies on latent restoration (RMSE) and classification (ACC).

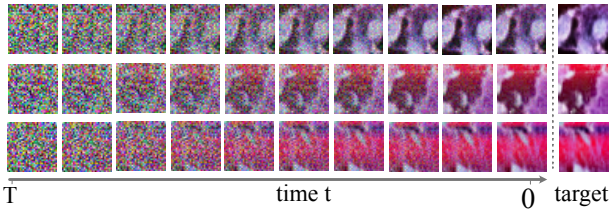

Figure 6: Visualization of diffusion process over time t.

three datasets. Lower masking rates may hinder the accurate learning of target entity semantics within the visible parts, while higher rates may result in excessive difficulty.

Decode Strategy Studies The decoder's conditioning strategies need empirical investigation. We test six cross-attention block strategies by aligning different graph attributes between the encoder and decoder (N for vertex, E for edge): NtoN, EtoE, NtoE, EtoN , NtoN & EtoE, and NtoE & EtoN. Graph vertices generally contain more information than edges. PANDA results in Table 4 show that graph attribute alignment enhances representation for classification, while attribute heterogeneity improves reconstruction.

Entity Latent Diffusion Process Fig. 6 represents the entity restoration as a pre-training proxy task in the latent space. We try to reconstruct fine-grained latent targets throughout the sampling process: as the sampling time t iterates, the latent similarity between diffuse entities and entities increases during the reverse of the process. It Demonstrates excellent representation for latent recovery by the conditioning encoder $\mathcal{E}_L$ in the diffusion decoder $\mathcal{D}_L$ of H-MGDM.

Conclusion

Our proposed novel framework, the dynamic entity mask on graph diffusion model for histopathology (H-MGDM), addresses the challenge in representation learning and enhances pre-training histopathology representation by incorporating tissue structural information and a suitable masking technique to guide diffusion models during reconstruction. The results of our experiments on six histopathology datasets covering common cancer types demonstrate the superior classification and regression performance of H-MGDM compared to existing methodologies. We anticipate the widespread applicability of H-MGDM in diverse downstream tasks such as prognosis and generation. Additionally, we are committed to exploring the potential of the newer entity extraction methods, the deeper interpretability, and the larger-scale experiments of our framework in our future research endeavors.

Acknowledgments

This work was supported by National Natural Science Foundation of China (Grant No. 62371409) and Fujian Provincial Natural Science Foundation of China (Grant No. 2023J01005).

References

Achanta, R.; Shaji, A.; Smith, K.; Lucchi, A.; Fua, P.; and Süsstrunk, S. 2012. SLIC superpixels compared to state-of-the-art superpixel methods. *IEEE transactions on pattern analysis and machine intelligence*, 34(11): 2274–2282.

Bulten, W.; Kartasalo, K.; Chen, P.-H. C.; Ström, P.; Pinckaers, H.; Nagpal, K.; Cai, Y.; Steiner, D. F.; van Boven, H.; Vink, R.; et al. 2022. Artificial intelligence for diagnosis and Gleason grading of prostate cancer: the PANDA challenge. *Nature medicine*, 28(1): 154–163.

Caron, M.; Touvron, H.; Misra, I.; Jégou, H.; Mairal, J.; Bojanowski, P.; and Joulin, A. 2021. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, 9650–9660.

Chen, R. J.; Lu, M. Y.; Shaban, M.; Chen, C.; Chen, T. Y.; Williamson, D. F.; and Mahmood, F. 2021. 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*, 339–349. Springer.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, 1597–1607. PMLR.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Esser, P.; Rombach, R.; and Ommer, B. 2021. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12873–12883.

Guan, H.; and Liu, M. 2021. Domain adaptation for medical image analysis: a survey. *IEEE Transactions on Biomedical Engineering*, 69(3): 1173–1185.

He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; and Girshick, R. 2022. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16000–16009.

Hinton, G. E.; and Roweis, S. 2002. Stochastic neighbor embedding. *Advances in neural information processing systems*, 15.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Hou, W.; Huang, H.; Peng, Q.; Yu, R.; Yu, L.; and Wang, L. 2022a. Spatial-hierarchical graph neural network with dynamic structure learning for histological image classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 181–191. Springer.

Hou, Z.; He, Y.; Cen, Y.; Liu, X.; Dong, Y.; Kharlamov, E.; and Tang, J. 2023. GraphMAE2: A Decoding-Enhanced Masked Self-Supervised Graph Learner. In *Proceedings of the ACM Web Conference 2023*, 737–746.

Hou, Z.; Liu, X.; Cen, Y.; Dong, Y.; Yang, H.; Wang, C.; and Tang, J. 2022b. GraphMAE: Self-Supervised Masked Graph Autoencoders. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 594–604.

Huang, H.; Sun, L.; Du, B.; and Lv, W. 2023. Conditional diffusion based on discrete graph structures for molecular graph generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 4302–4311.

Ilse, M.; Tomczak, J.; and Welling, M. 2018. Attention-based deep multiple instance learning. In *International conference on machine learning*, 2127–2136. PMLR.

Katzman, J. L.; Shaham, U.; Cloninger, A.; Bates, J.; Jiang, T.; and Kluger, Y. 2018. DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC medical research methodology*, 18: 1–12.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kingma, D. P.; and Welling, M. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Kipf, T. N.; and Welling, M. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Komura, D.; Kawabe, A.; Fukuta, K.; Sano, K.; Umezaki, T.; Koda, H.; Suzuki, R.; Tominaga, K.; Ochi, M.; Konishi, H.; et al. 2022. Universal encoding of pan-cancer histology by deep texture representations. *Cell Reports*, 38(9).

Li, X.; Cen, M.; Xu, J.; Zhang, H.; and Xu, X. S. 2022. Improving feature extraction from histopathological images through a fine-tuning ImageNet model. *Journal of Pathology Informatics*, 13: 100115.

Liu, J.; Huang, X.; Liu, Y.; and Li, H. 2022. Mixmim: Mixed and masked image modeling for efficient visual representation learning. *arXiv preprint arXiv:2205.13137*.

Panayides, A. S.; Amini, A.; Filipovic, N. D.; Sharma, A.; Tsaftaris, S. A.; Young, A.; Foran, D.; Do, N.; Golemati, S.; Kurc, T.; et al. 2020. AI in medical imaging informatics: current challenges and future directions. *IEEE journal of biomedical and health informatics*, 24(7): 1837–1857.

Pati, P.; Jaume, G.; Foncubierta-Rodriguez, A.; Feroce, F.; Anniciello, A. M.; Scognamiglio, G.; Brancati, N.; Fiche, M.; Dubruc, E.; Riccio, D.; et al. 2022. Hierarchical graph

representations in digital pathology. *Medical image analysis*, 75: 102264.

Preechakul, K.; Chatthee, N.; Wizadwongsa, S.; and Suwajanakorn, S. 2022. Diffusion Autoencoders: Toward a Meaningful and Decodable Representation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Purma, V.; Srinath, S.; Srirangarajan, S.; Kakkar, A.; et al. 2023. GenSelfDiff-HIS: Generative Self-Supervision Using Diffusion for Histopathological Image Segmentation. *arXiv preprint arXiv:2309.01487*.

Riasatian, A.; Babaie, M.; Maleki, D.; Kalra, S.; Valipour, M.; Hemati, S.; Zaveri, M.; Safarpoor, A.; Shafiei, S.; Afshari, M.; Rasoolijaberi, M.; Sikaroudi, M.; Adnan, M.; Shah, S.; Choi, C.; Damaskinos, S.; Campbell, C. J.; Diamandis, P.; Pantanowitz, L.; Kashani, H.; Ghodsi, A.; and Tizhoosh, H. 2021. Fine-Tuning and training of densenet for histopathology image representation using TCGA diagnostic slides. *Medical Image Analysis*, 70: 102032.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, 618–626.

Sharmay, Y.; Ehsany, L.; Syed, S.; and Brown, D. E. 2021. HistoTransfer: Understanding Transfer Learning for Histopathology. In *2021 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)*, 1–4.

Song, A. H.; Jaume, G.; Williamson, D. F.; Lu, M. Y.; Vaidya, A.; Miller, T. R.; and Mahmood, F. 2023. Artificial intelligence for digital and computational pathology. *Nature Reviews Bioengineering*, 1(12): 930–949.

Štrumbelj, E.; and Kononenko, I. 2014. Explaining prediction models and individual predictions with feature contributions. *Knowledge and information systems*, 41: 647–665.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wei, C.; Mangalam, K.; Huang, P.-Y.; Li, Y.; Fan, H.; Xu, H.; Wang, H.; Xie, C.; Yuille, A.; and Feichtenhofer, C. 2023. Diffusion Models as Masked Autoencoder. In *ICCV*.

Yang, X.; and Wang, X. 2023. Diffusion model as representation learner. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 18938–18949.

Zhou, Y.; Graham, S.; Koohbanani, N. A.; Shaban, M.; Heng, P.-A.; and Rajpoot, N. 2019. CGC-Net: Cell Graph Convolutional Network for Grading of Colorectal Cancer Histology Images. In *The IEEE International Conference on Computer Vision (ICCV) Workshops*.