Dual-Curriculum Contrastive Multi-Instance Learning for Cancer Prognosis Analysis with Whole Slide Images

Anonymous Author(s)
Affiliation
Address
email

Abstract

The multi-instance learning (MIL) has advanced cancer prognosis analysis with whole slide images (WSIs). However, current MIL methods for WSI analysis still confront unique challenges. Previous methods typically generate instance representations via a pre-trained model or a model trained by the instances with bag-level annotations, which, however, may not generalize well to the downstream task due to the introduction of excessive label noises and the lack of fine-grained information across multi-magnification WSIs. Additionally, existing methods generally aggregate instance representations as bag ones for prognosis prediction and have no consideration of intra-bag redundancy and inter-bag discrimination. To address these issues, we propose a dual-curriculum contrastive MIL method for cancer prognosis analysis with WSIs. The proposed method consists of two curriculums, i.e., saliency-guided weakly-supervised instance encoding with cross-scale tiles and contrastive-enhanced soft-bag prognosis inference. Extensive experiments on three public datasets demonstrate that our method outperforms state-of-the-art methods in this field. The code is avaliable.

1 Introduction

Cancer prognosis analysis is of great significance for risk-benefit assessment and clinical decision [1]. Computational whole slide image (WSI), which entails the quantitative profiling of spatial patterns and tumor microenvironments in tissue slides, has advanced the application of deep learning in cancer prognosis analysis [2, 3]. However, deep learning in WSIs is often hindered by the gigapixel size and the lack of pixel-level annotations [4]. Rencently, multi-instance learning (MIL) methods have attracted increasing attention to performing inference in a weakly-supervised manner and have been successfully applied to WSI analysis [5, 6, 7].

In general, the MIL consists of two stages, i.e., instance encoding and aggregation stages [8, 9]. In the encoding stage, several methods have utilized the pre-trained model (e.g., ResNet on ImageNet) to extract instance representations in an unsupervised way (i.e., without the reference of outcome labels) [10, 11, 12]. However, recent studies have demonstrated that the pre-trained model may not generalize well to the downstream tasks as it is task-agnostic and inclined to overfit the pretraining objective [13, 14, 15]. Alternatively, some studies directly assigned each instance the bag-level annotation (i.e., the survival time of the patient) for network training, which introduces task-specific supervision and improves model's generalisability [16, 17, 18]. Actually, the prognosis evaluation is primarily and comprehensively determined by certain representative regions, yet those regions might only occupy a small portion of WSI. As a result, the strong supervision on all instances, especially on prognosis-irrelevant tiles, will introduce excessive label noises and limit model's performance. For alleviating this issue, a feasible way is to supervise the instance encoding in a weak and easy-to-hard

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

manner. On the other hand, existing methods often generate mono-scale instance representations [11, 19] or typically concatenate multi-scale information from multi-magnification WSIs [20, 21], which may result in feature redundancy and ignore fine-grained details across multi-magnification images.

In the aggregation stage, most previous studies aggregate instance representations within a bag for prognosis estimation by certain fusion strategies (e.g. pooling operation and attention mechanism) [22, 23]. However, it may overwhelm prognosis-relevant information, cause intra-bag redundancy, and reduce inter-bag discrimination, if many instances from irrelevant regions are enrolled [12, 24]. To address the above-mentioned issues, we arm MIL with a dual-curriculum strategy and propose a dual-curriculum contrastive MIL method for cancer prognosis analysis with WSIs. The proposed method consists of two easy-to-hard curriculums, i.e., saliency-guided weakly-supervised instance encoding with cross-scale tiles and contrastive-enhanced soft-bag prognosis inference. The main contributions are summarized as follows:

1) We present a dual-curriculum contrastive MIL method which includes two easy-to-hard curriculums. We first conduct a preliminary task to learn instance representations by considering risk stratification status (degraded from survival time) as the annotation, followed by the prognosis inference with survival time as supervision.

2) We design the first curriculum of saliency-guided weakly-supervised instance encoding with cross-scale tiles. It is supervised by the relatively weak annotations so as to reduce label noises and maintain prognosis-related guidance. Additionally, to imitate the reviewing procedure of pathologists, we leverage the low-magnification saliency map to guide the encoding of high-magnification instances for exploring fine-grained information across multi-magnification WSIs.

3) We develop the second curriculum of contrastive-enhanced soft-bag prognosis inference. Instead of enrolling all instances, we adaptively identify and integrate representative instances within a bag (as the soft bag) for prognosis inference and leverage the constrained self-attention strategy to obtain extra sparseness for soft-bag representations, which can help reduce intra-bag redundancy in both instance and feature levels. Meanwhile, we improve the Cox loss with two-tier contrastive learning for enhancing intra-bag and inter-bag discrimination.

4) We evaluate the proposed method on three public cancer datasets and extensive experiments demonstrate that our method outperforms state-of-the-art methods in cancer prognosis analysis with WSIs.

2 Related Work

2.1 WSI-based Cancer Prognosis Model

In recent years, many methods, especially convolutional neural networks (CNNs), have been proposed for cancer prognosis evaluation with WSIs [25, 26, 27]. Due to the gigapixel size of WSI, some methods took small-size ROIs selectively sampled from WSIs as the input of the network for training and inference, in which all ROIs were assigned with patient-level outcome labels [28, 29]. However, these methods generally require prior knowledge from pathologists, which are experience-dependent and may suffer from high inter-observer variation. Therefore, current researches have proposed to crop the entire WSI into many tiles for the model's input so as to make full use of available information contained by WSI [10, 30, 31]. However, these methods often confront the lack of tile-level annotations and may result in a sub-optimal solution. To this end, MIL methods have been proposed to perform cancer prognosis analysis in a weakly-supervised manner and have shown promising performance in this field [5, 26].

2.2 Multi-instance Learning in WSI Analysis

The MIL methods in WSI analysis can be roughly divided into two categories according to the way instances are used in the model [12, 17]. The one infers bag-level ones from instance-level prediction via average pooling, maximum pooling, or other ensemble strategies [32, 33]. In contrast, the other one, referred to as the bag embedding method, aggregates instance representations as bag ones, upon which a model is constructed for bag-level prediction [34, 35]. The former is simple and straightforward but empirically proven to be inferior to the latter in terms of generalization

performance. Therefore, the bag embedding method has attracted extensive attention in WSI analysis [17]. For example, Courtiol et al. [6] developed a MIL method based on CNN, called MesoNet, to accurately predict the overall survival of mesothelioma patients with WSIs. Yao et al. [5] proposed deep attention multiple instance survival learning to efficiently learn features from the WSI and aggregated them for prognosis prediction. Our method essentially belongs to the bag embedding methods. Different from previous methods, the proposed method consists of two easy-to-hard curriculums, in which fine-grained information across multi-magnification WSIs, intra-bag redundancy, and inter-bag discrimination are comprehensively considered.

3 Method

3.1 Problem Formulation

Denote the triplet $\{\mathbf{X}_n, \mathbf{T}_n, \boldsymbol{\delta}_n\}_{n=1}^{N}$ as a histopathological dataset consisting of N patients, where $\mathbf{X}_n$ is the n-th patient with one or more WSIs, $\mathbf{T}_n$ and $\boldsymbol{\delta}_n$ are its corresponding observed survival time and event indicator (viz. is equal to 1 and 0 for uncensored and censored samples, respectively). Prognosis analysis is an ordinal regression task that models time-to-event distribution, which is formulated as

$$\hat{\mathbf{T}}_n = \mathbb{S}(\mathbf{X}_n), \tag{1}$$

where $\mathbb{S}$ and $\hat{\mathbf{T}}_n$ denote the prognosis model and the predicted survival time, respectively.

In practice, it is difficult to directly feed the whole WSI into the model due to its gigapixel size. Accordingly, many studies [10, 30, 31] have cropped WSIs into many tiles and generated bag $\mathbf{X}_n = \{\mathbf{x}_{n,i}\}_{i=1}^{N_n}$ which contains N_n instances (tiles) for the n-th patient. Then, MIL is leveraged to construct prognosis annalysis model as follows:

$$\hat{\mathbf{T}}_n = \mathbb{S}(\mathbb{A}\{\mathbb{E}(\mathbf{x}_{n,i}) : \mathbf{x}_{n,i} \in \mathbf{X}_n\}), \tag{2}$$

where $\mathbb{E}$ is an instance-level encoder that extracts feature representations for each instance, and $\mathbb{A}$ is a permutation-invariant aggregation network that pools the instance representations into the bag ones.

However, previous studies [10, 11, 12, 16, 17, 18] typically generate instance representations via a pre-trained model or a model trained by the instances with bag-level annotations, which may not generalize well to the downstream task. As a result, we decompose the prognosis analysis model into two easy-to-hard curriculums, including (1) Curriculum I (C-I): instance encoding $\mathbb{E}$ supervised by preliminary risk stratification $\mathbb{C}$ and (2) Curriculum II (C-II): prognosis inference $\mathbb{S}$ after instance aggregation $\mathbb{A}$. Formally, the optimization process can be formulated as

$$\begin{cases} \text{C-I} : \hat{\mathbb{E}}, \hat{\mathbb{C}} = \arg\min\limits_{\mathbb{E},\mathbb{C}} \sum\limits_{n=1}^{N} \dagger(\mathbf{Y}_n == 1 \| 0) \sum\limits_{i=1}^{N_n} [\mathbb{C}\{\mathbb{E}(\mathbf{x}_{n,i}) : \mathbf{x}_{n,i} \in \mathbf{X}_n\}, \mathbf{Y}_n]_{\mathcal{L}_{\text{I}}} \\ \text{C-II} : \hat{\mathbb{A}}, \hat{\mathbb{S}} = \arg\min\limits_{\mathbb{A},\mathbb{S}} \sum\limits_{n=1}^{N} \mathbb{S}(\mathbb{A}\{\hat{\mathbb{E}}(\mathbf{x}_{n,i}) : \mathbf{x}_{n,i} \in \mathbf{X}_n\}), \mathbf{T}_n, \boldsymbol{\delta}_n]_{\mathcal{L}_{\text{II}}} \end{cases}, \tag{3}$$

where $\dagger(\cdot)$ outputs 1 if ture, else 0, and $\mathcal{L}_{\text{I}}$ and $\mathcal{L}_{\text{II}}$ denote the loss functions for C-I and C-II, respectively. And the risk stratification status $\mathbf{Y}_n$ is determined by $\{\mathbf{T}_n, \boldsymbol{\delta}_n\}$ with a three-year time threshold (denoted as $\mathbf{T}_r$) as follows:

$$\mathbf{Y}_n = \begin{cases} 1 & , \mathbf{T}_n \leq \mathbf{T}_r \ \&\& \ \boldsymbol{\delta}_n == 1 \\ 0 & , \mathbf{T}_n > \mathbf{T}_r \\ \text{NAN} & , \text{others} \end{cases}. \tag{4}$$

3.2 Curriculum I: Saliency-guided Weakly-supervised Instance Encoding with Cross-scale Tiles

In this section, we introduce the first curriculum of saliency-guided weakly-supervised instance encoding with cross-scale tiles, as shown in Figure 1. Given the input instance with S magnifications $\{\mathbf{x}_{n,i}^{s}\}_{s=1}^{S}$, the proposed method contains S ($S = 3$ in this work) branches and each branch takes different-magnification WSI as input, which aims to explore multi-scale information from

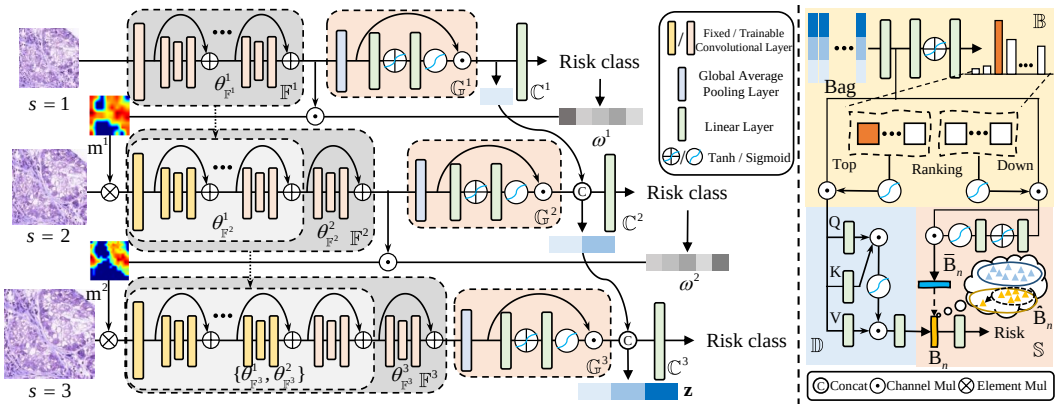

Figure 1: The pipeline of the proposed dual-curriculum contrastive MIL method for cancer prognosis analysis with WSIs. The left is saliency-guided weakly-supervised instance encoding with cross-scale tiles (Curriculum I), while the right one is contrastive-enhanced soft-bag prognosis inference (Curriculum II).

multi-magnification images. For the s-th branch, it mainly consists of a feature extractor $\mathbb{F}^s$, a feature aggregator $\mathbb{G}^s$ and a classifier $\mathbb{C}^s$, where $\mathbb{F}^s$ and $\mathbb{G}^s$ $(s = 1, \cdots, S)$ form $\mathbb{E}$, while $\mathbb{C}^s$ $(s = 1, \cdots, S)$ constitute $\mathbb{C}$ in Eq.(3). The feature extractor $\mathbb{F}^s$ contains s residual blocks (i.e., $\{\boldsymbol{\theta}^1_{\mathbb{F}^s}, \cdots, \boldsymbol{\theta}^s_{\mathbb{F}^s}\}$), in which each block primarily contains several convolutional layers, pooling layers and skip connections (detail architecture can refer to the supplemental material), and the first $s-1$ blocks (i.e., $\{\boldsymbol{\theta}^1_{\mathbb{F}^s}, \cdots, \boldsymbol{\theta}^{s-1}_{\mathbb{F}^s}\}$) share the same architecture with $\mathbb{F}^{s-1}$. Different from previous studies [20, 21] that separately feed input $\{\mathbf{x}^s_{n,i}\}_{s=1}^S$ into the network and ignore fine-grained details across multi-magnification images, we encourage the network to focus on salient regions by utilizing the prior knowledge of the $(s-1)$-th branch to highlight the input $\mathbf{x}^s_{n,i}$, which can be formulated as

$$\hat{\mathbf{x}}^s_{n,i} = \begin{cases} \mathbf{x}^s_{n,i} & , s = 1 \\ \mathbf{m}^{s-1}_{n,i} \otimes \mathbf{x}^s_{n,i} & , s > 1 \end{cases}, \tag{5}$$

where $\otimes$ indicates the element-wise multiplication operator, and $\mathbf{m}^{s-1}_{n,i}$ is the salient mask indicating high-risk regions and is computed by Class Activation Map [36] as follows:

$$\mathbf{m}^{s-1}_{n,i} = \boldsymbol{\omega}^{s-1}_{n,i} \odot \mathbb{F}^{s-1}(\hat{\mathbf{x}}^{s-1}_{n,i}), \tag{6}$$

where $\odot$ indicates the channel-wise multiplication operator and $\boldsymbol{\omega}^{s-1}_{n,i}$ is the importance of each channel of feature maps for risk stratification prediction in the last branch.

For convenience, we introduce how to acquire $\boldsymbol{\omega}^s_{n,i}$, which is easily generalized to $\boldsymbol{\omega}^{s-1}_{n,i}$. Specifically, we first obtain instance representations $\mathbf{z}^s_{n,i}$ by

$$\mathbf{z}^s_{n,i} = \mathbb{G}^s(\mathbb{F}^s(\hat{\mathbf{x}}^s_{n,i})), \tag{7}$$

where $\mathbb{G}^s$ consists of a global average pooling operation $(\rightarrow)$, two linear layers (denoted as f^s_1 and f^s_2) with tanh (ψ) and softmax (ϕ) activation functions, respectively, which is formulated as

$$\mathbb{G}^s(\mathbf{v}^s_{n,i}) = \phi(f^s_2(\psi(f^s_1(\overrightarrow{\mathbf{v}^s_{n,i}})))^T) \odot \overrightarrow{\mathbf{v}^s_{n,i}}, \tag{8}$$

where we have $\mathbf{v}^s_{n,i} := \mathbb{F}^s(\hat{\mathbf{x}}^s_{n,i})$. Incorporating comprehensive information from current and previous branches, we subsequently feed instance representations $\{\mathbf{z}^j_{n,i}\}^s_{j=1}$ into the classifier $\mathbb{C}^s$ for predicting high-risk probability $\mathbf{p}^s_{n,i}$, which is computed by

$$\mathbf{p}^s_{n,i} = \mathbb{C}^s(\{\mathbf{z}^j_{n,i}\}^s_{j=1}) = \phi(f^s_3(\{\sum_{j=1}^{s-1} \mathbf{z}^j_{n,i}, \mathbf{z}^s_{n,i}\})) = \phi(\overline{\boldsymbol{\omega}}^s \times \{\sum_{j=1}^{s-1} \mathbf{z}^j_{n,i}\}^T + \boldsymbol{\omega}^s \times \{\mathbf{z}^s_{n,i}\}^T), \tag{9}$$

where T denotes transposition operation and f^s_3 is a linear layer with its corresponding weight $[\overline{\boldsymbol{\omega}}^s, \boldsymbol{\omega}^s]$ that indicates the importance of instance representations for risk stratification prediction.

We introduce a hierarchical transfer strategy to leverage the prior knowledge from different-magnification WSIs, accelerate network convergence, and maintain the specificity of current magnification. Specifically, the first $s-2$ blocks $\{\boldsymbol{\theta}^1_{\mathbb{F}^s}, \cdots, \boldsymbol{\theta}^{s-2}_{\mathbb{F}^s}\}$ in $\mathbb{F}^s$ are initialized with the parameters of the blocks $\{\boldsymbol{\theta}^1_{\mathbb{F}^{s-1}}, \cdots, \boldsymbol{\theta}^{s-2}_{\mathbb{F}^{s-1}}\}$ in $\mathbb{F}^{s-1}$, and then are frozen. The block $\boldsymbol{\theta}^{s-1}_{\mathbb{F}^s}$ is learnable but constrained by the structural loss $\mathcal{L}_\Omega$ (see Eq.(12)) to approximate the block $\boldsymbol{\theta}^{s-1}_{\mathbb{F}^{s-1}}$, while the block $\boldsymbol{\theta}^{s}_{\mathbb{F}^s}$ does not have any constraints. Furthermore, we propose a hybrid loss function $\mathcal{L}_{\mathrm{I}}$ that contains two parts, i.e., empirical loss $\mathcal{L}_\ell$ and structural loss $\mathcal{L}_\Omega$, to train the network:

$$\mathcal{L}_{\mathrm{I}} = \mathcal{L}_\ell + \beta \mathcal{L}_\Omega, \tag{10}$$

where β is the weight coefficient, and the empirical loss $\mathcal{L}_\ell$ is defined as

$$\mathcal{L}_\ell = -\frac{1}{NN_n} \sum_{n=1}^{N} \sum_{i=1}^{N_n} \mathbf{y}^s_{n,i} \log(\mathbf{p}^s_{n,i}) + (1 - \mathbf{y}^s_{n,i}) \log(1 - \mathbf{p}^s_{n,i}), \tag{11}$$

where $\mathbf{y}^s_{n,i} \in \mathbf{Y}_n$ denotes the risk stratification status of the i-th instance in the n-th bag. And the structural loss $\mathcal{L}_\Omega$ is computed by

$$\mathcal{L}_\Omega = \left\| \boldsymbol{\theta}^{s-1}_{\mathbb{F}^s} - \boldsymbol{\theta}^{s-1}_{\mathbb{F}^{s-1}} \right\|_2. \tag{12}$$

Finally, we can obtain the multi-scale instance representations $\mathbf{z}_{n,i}$ for subsequent prognosis inference task by

$$\mathbf{z}_{n,i} = [\mathbf{z}^1_{n,i}, \cdots, \mathbf{z}^S_{n,i}]. \tag{13}$$

3.3 Curriculum II: Contrastive-enhanced Soft-bag Prognosis Inference

In this part, we develop the second curriculum of contrastive-enhanced soft-bag prognosis inference, as shown in Figure 1. Given a set of instance representations from n-th bag, i.e., $\mathbf{Z}_n = [\mathbf{z}_{n,1}; \mathbf{z}_{n,2}; \cdots; \mathbf{z}_{n,N_n}]^T \in \mathbb{R}^{N_n \times C}$, where C denotes the feature dimension of $\mathbf{Z}_n$, the proposed method mainly consists of a soft-bag learning module $\mathbb{B}$, a constrained self-attention module $\mathbb{D}$, a contrastive-enhanced prognosis inference function $\mathbb{S}$, where $\mathbb{B}$ and $\mathbb{D}$ form $\mathbb{A}$ in Eq.(3). For the soft-bag learning module $\mathbb{B}$, it further learns new bag representation $\mathbf{E}_n$ from $\mathbf{Z}_n$ via a linear layer f_4:

$$\mathbf{E}_n = f_4(\mathbf{Z}_n) \in \mathbb{R}^{N_n \times D}, \tag{14}$$

where D denotes the feature dimension of $\mathbf{E}_n$. Instead of enrolling all instances, we adaptively identify and integrate representative instances within a bag by introducing the function $\mathbb{H}$:

$$\hat{\mathbf{E}}_n = \mathbb{H}(\mathbf{E}_n, \hat{\mathbf{h}}_n) = \phi(diag\{\hat{\mathbf{h}}_n\} \times \mathbf{W}_n)^T \times (diag\{\hat{\mathbf{h}}_n\} \times \mathbf{E}_n), \tag{15}$$

where $\hat{\mathbf{h}}_n \in \mathbb{R}^{N_n \times 1}$ is an indicator vector that adaptively selects N_B representative instances and is computed by

$$\hat{\mathbf{h}}_n = \arg\max_{\mathbf{h}_n} \mathbb{S}\left(\mathbb{D}\left(\mathbb{H}(\mathbf{E}_n, \mathbf{h}_n)\right)\right) \quad s.t. \sum_{i=1}^{N_n} \mathbf{h}_{n,i} = N_B, \tag{16}$$

where N_B is the number of selected instances in each bag. And $\mathbf{W}_n \in \mathbb{R}^{N_n \times N_B}$ ($N_B \ll N_n$) denotes the weight matrix and has the following definition:

$$\mathbf{W}_n = f_6(\psi(f_5(\mathbf{E}_n))), \tag{17}$$

where f_5 and f_6 are two linear layers.

Subsequently, we leverage the constrained self-attention module $\mathbb{D}$ to obtain extra sparseness for soft-bag representations. Given the input $\hat{\mathbf{E}}_n \in \mathbb{R}^{N_B \times D}$, the module $\mathbb{D}$ first linearly projects $\hat{\mathbf{E}}_n$ into three feature spaces, that is $\mathbf{Q}_n = f_Q(\hat{\mathbf{E}}_n) \in \mathbb{R}^{N_B \times \hat{D}}$, $\mathbf{K}_n = f_K(\hat{\mathbf{E}}_n) \in \mathbb{R}^{N_B \times \hat{D}}$, and $\mathbf{V}_n = f_V(\hat{\mathbf{E}}_n) \in \mathbb{R}^{N_B \times \hat{D}}$, where f_Q, f_K, and f_V are corresponding linear layers and their parameters are constrained by

$$\mathcal{L}_s = (\|\mathbf{W}_Q\|_1 + \|\mathbf{W}_K\|_1 + \|\mathbf{W}_V\|_1). \tag{18}$$

Then, the sparse soft-bag representation $\mathbf{B}_n \in \mathbb{R}^{1 \times D_B}$ (where D_B is the feature dimension) can be obtained by the self-attention mechanism as follows:

$$\mathbf{B}_n = \mathbb{D}(\hat{\mathbf{E}}_n) = f_7(\phi(\mathbf{Q}_n \times {\mathbf{K}_n}^T) \times \mathbf{V}_n), \tag{19}$$

where f_7 denotes the linear layer. It is worth mentioning that the collaboration of $\mathbb{B}$ and $\mathbb{D}$ can significantly help reduce intra-bag redundancy in both instance and feature levels so as to improve the model's generalization.

In order to improve the ability of soft-bag inference, we perform two-tier contrastive learning (TCL) to enhance intra-bag and inter-bag discrimination. We assume that a bag $\mathbf{B}_n$ is exampled from the distribution $\mathcal{O} = \{\mathcal{O}^+, \mathcal{O}^-\}$, where the positive source $\mathcal{O}^+$ denotes the distribution of high-hazard samples (i.e. those with survival time $\leq \mathbf{T}_r$), while the negative source $\mathcal{O}^-$ presents the distribution of low-hazard samples (i.e. those with survival time $> \mathbf{T}_r$). We first focus on the inter-bag discrimination. Let $\mathcal{B}_n = \{\mathbf{B}_1, \cdots, \mathbf{B}_{N_{B_n}}\}/\mathbf{B}_n$ present the collection of bags (expect $\mathbf{B}_n$) from the same source $\mathcal{O}^+$ (or $\mathcal{O}^-$) with $\mathbf{B}_n$, where N_{B_n} denotes the number of such bags. We can obtain the representation of the centre (denoted as $\hat{\mathbf{B}}_n$) for the collection $\mathcal{B}_n$ by merging the representations of all bags in $\mathcal{B}_n$, i.e., $\hat{\mathbf{B}}_n = \sum_{i\in[N_{B_n}], i\neq n} \mathbf{B}_i$. Inspired by [37], the inter-bag discrimination can be enhanced by maximally preserving the mutual information between $\mathbf{B}_n$ and $\hat{\mathbf{B}}_n$ as follows:

$$\sum_{n\in[N]} \mathbb{I}(\mathbf{B}_n, \hat{\mathbf{B}}_n) = \sum_{n\in[N]} p(\mathbf{B}_n, \hat{\mathbf{B}}_n) \log \frac{p(\mathbf{B}_n|\hat{\mathbf{B}}_n)}{p(\mathbf{B}_n)}. \tag{20}$$

From the Eq.(15), partial instances of a bag will be adaptively discarded and the rest remained. However, all instances within the bag belong to homologous tissues sampled from the same patient, which means that it should show potential correlation among these instances. Therefore, similar to Eq.(20), we maximize the following mutual information to improve the ability of intra-bag discrimination:

$$\sum_{n\in[N]} \mathbb{I}(\mathbf{B}_n, \overline{\mathbf{B}}_n) = \sum_{n\in[N]} p(\mathbf{B}_n, \overline{\mathbf{B}}_n) \log \frac{p(\mathbf{B}_n|\overline{\mathbf{B}}_n)}{p(\mathbf{B}_n)}, \tag{21}$$

where $\overline{\mathbf{B}}_n = \phi(f_9(\psi(f_8(\overline{\mathbf{E}}_n)))) \times \overline{\mathbf{E}}_n$ (f_8 and f_9 are two linear layers) and $\overline{\mathbf{E}}_n = \mathbb{H}(\mathbf{E}_n, \mathbf{1} - \hat{\mathbf{h}}_n)$. To this end, we propose a two-tier contrastive learning method to jointly maximize the Eq.(20) and (21), i.e., $max(\sum_{n\in[N]} \mathbb{I}(\mathbf{B}_n, \hat{\mathbf{B}}_n) + \mathbb{I}(\mathbf{B}_n, \overline{\mathbf{B}}_n))$, which is formulated by:

$$max(\sum_{n\in[N]} \log \frac{p(\mathbf{B}_n|\hat{\mathbf{B}}_n)}{p(\mathbf{B}_n)} + \log \frac{p(\mathbf{B}_n|\overline{\mathbf{B}}_n)}{p(\mathbf{B}_n)}) = max(\sum_{n\in[N]} \log \frac{p(\mathbf{B}_n|\hat{\mathbf{B}}_n)p(\mathbf{B}_n|\overline{\mathbf{B}}_n)}{p(\mathbf{B}_n)p(\mathbf{B}_n)}). \tag{22}$$

Instead of constructing a generative model $p(\mathbf{B}_n|\hat{\mathbf{B}}_n)p(\mathbf{B}_n|\overline{\mathbf{B}}_n)$, we leverage a log-bilinear function to model the density ratio which preserves the mutual information as follows:

$$f(\mathbf{B}_n, \hat{\mathbf{B}}_n, \overline{\mathbf{B}}_n) = exp(\mathbf{B}_n^T \mathbf{W}_f \hat{\mathbf{B}}_n) exp(\mathbf{B}_n^T \mathbf{V}_f \overline{\mathbf{B}}_n) \tag{23}$$

$$= exp(\mathbf{B}_n^T \mathbf{W}_f \hat{\mathbf{B}}_n + \mathbf{B}_n^T \mathbf{V}_f \overline{\mathbf{B}}_n) \propto \frac{p(\mathbf{B}_n|\hat{\mathbf{B}}_n)p(\mathbf{B}_n|\overline{\mathbf{B}}_n)}{p(\mathbf{B}_n)p(\mathbf{B}_n)}, \tag{24}$$

where $\mathbf{W}_f$ and $\mathbf{V}_f$ are trainable parameters. Motivated by [38], we design the TCL loss based on the InfoNCE function, which has the following definition:

$$\mathcal{L}_{tcl} = -\underset{\mathcal{B}}{\boldsymbol{E}} \left[\log \frac{f(\mathbf{B}_n, \hat{\mathbf{B}}_n, \overline{\mathbf{B}}_n)}{\sum_{i\in[N], i\neq n} f(\mathbf{B}_n, \hat{\mathbf{B}}_i, \overline{\mathbf{B}}_i)} \right], \tag{25}$$

where $\mathcal{B}$ denotes the collection of all bags. Additionally, we utilize the negative log partial likelihood of Cox proportional hazard regression model [39] to train the network, which is computed by

$$\mathcal{L}_{cox} = \log(\prod_{n:\boldsymbol{\delta}_n=1} \frac{exp(\mathbf{B}_n \mathbf{W}_c)}{\sum_{j\in R(\mathbf{T}_n)} exp(\mathbf{B}_j \mathbf{W}_c)}) = \sum_{n:\boldsymbol{\delta}_n=1} (\mathbf{B}_n \mathbf{W}_c - \log(\sum_{j\in R(\mathbf{T}_n)} exp(\mathbf{B}_j \mathbf{W}_c))),$$

where $\mathbf{W}_c$ is learnable parameter, and $R(\mathbf{T}_n) = \{i : \mathbf{T}_i \geq \mathbf{T}_n\}$ is the collection of all risk samples (including censored and uncensored samples) with survival time being longer than $\mathbf{T}_n$. Finally, the loss function $\mathcal{L}_{\text{II}}$ for the second curriculum can be defined as

$$\mathcal{L}_{\text{II}} = \mathcal{L}_{cox} + \mathcal{L}_{tcl} + \beta_s \mathcal{L}_s, \tag{26}$$

where β_s is the weight coefficient.

Table 1: The comparison with all competing methods on three datasets (i.e., COAD, LIHC, and BLCA). And the boldface denotes the best results of each metric.

	COAD		LIHC		BLCA		Overall	
	CI	AUC	CI	AUC	CI	AUC	CI	AUC
WSISA[40]	0.641±0.027	0.673±0.037	0.661±0.022	0.700±0.071	0.639±0.018	0.651±0.043	0.647	0.675
DeepGraphSurv[41]	0.629±0.054	0.639±0.038	0.613±0.027	0.633±0.047	0.620±0.015	0.687±0.046	0.620	0.653
MesoNet[6]	0.643±0.039	0.684±0.052	0.629±0.022	0.667±0.042	0.630±0.021	0.648±0.052	0.634	0.666
DeepAttnMISL[5]	0.683±0.025	0.654±0.045	0.672±0.027	0.676±0.019	0.642±0.020	0.661±0.076	0.666	0.663
Patch-GCN[10]	0.650±0.013	0.562±0.010	0.565±0.024	0.630±0.061	0.571±0.041	0.542±0.027	0.595	0.578
DC_MIL(our)	**0.717±0.012**	**0.754±0.031**	**0.705±0.015**	**0.745±0.012**	**0.672±0.029**	**0.720±0.039**	**0.698**	**0.740**

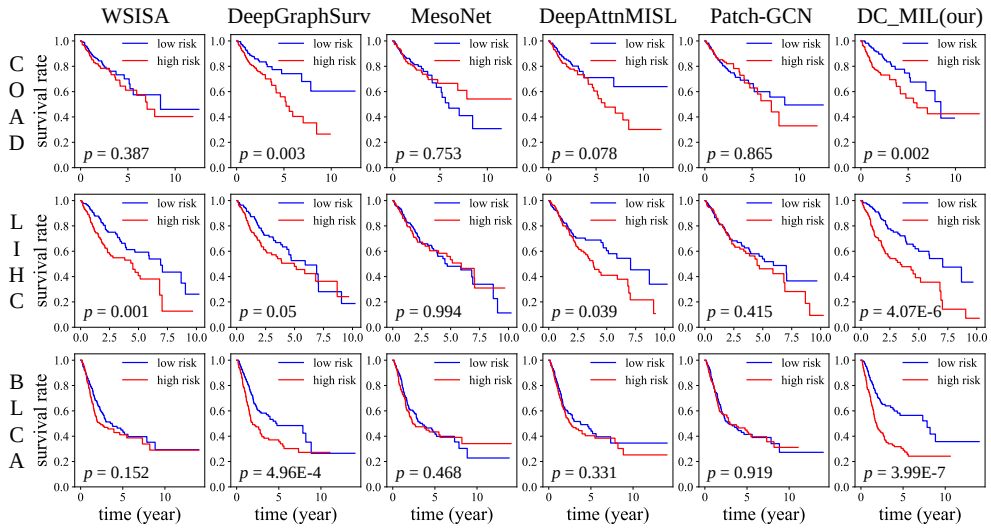

Figure 2: The KM curves with the p-value of all competing methods on three datasets. The farther apart the two curves are, the better the model performance.

4 Experiments and Results

4.1 Datasets and Preprocessing

To demonstrate the superiority of the proposed DC_MIL, various experiments were conducted over the studied datasets from the Cancer Genome Atlas databases (TCGA), including colon adenocarcinoma (COAD)($N = 365$), hepatocellular carcinoma (LIHC)($N = 334$), and bladder urothelial carcinoma (BLCA)($N = 396$). Each dataset contains WSIs stained with hematoxylin and eosin (H&E) and clinical information of survival time and event status. We first applied the Otsus method to the saturation channels of the WSI at each magnification after transforming it into the HSV color space for segmenting tissue areas. After segmenting tissue areas by the Otsus method, a set of tiles under three different magnifications ($20\times$, $10\times$, $5\times$) were non-overlappingly sampled from each segmented tissue area, and the corresponding window sizes were 512×512 pixels, 256×256 pixels, and 128×128 pixels, respectively. In this study, we constructed instances from 20 random sampling tiles and regarded a patient with 1$\sim$3 WSIs as a bag.

4.2 Implementation Details

The concordance index (CI), the receiver operating characteristic (ROC) curve (refer to supplemental material), and the area under the ROC curve (AUC) with the 3-year survival time as threshold were leveraged to measure the predictive performance of patient outcomes. We also utilized Kaplan-Meier (KM) curves to visualize the ability of stratifying patients with low- and high-risk distributions. In addition, we used Python3.6 to implement all experiments and Pytorch library to build the DC_MIL model. All computationally intensive calculations were offloaded to a 12 GB NVIDIA Pascal Titan X GPU.

Risk Stratification. For three different magnifications $s \in \{1, 2, 3\}$, we adopted Adam optimization with the default learning rate of $\{10^{-5}, 3 \times 10^{-6}, 3 \times 10^{-6}\}$, the weight decay of 0.2, and trained 50 epochs with the batch size $\{32, 16, 16\}$. The hyperparameter β was empirically set to 10^{-5}.

Table 2: The results of ablation experiments on three datasets (i.e., COAD, LIHC, and BLCA). And the boldface denotes the best results of each metric.

DC_MIL: C-I + C-II				COAD		LIHC		BLCA	
MM	SG	FT	C-II	CI	AUC	CI	AUC	CI	AUC
✗	✗	✗	✓	0.676±0.031	0.696±0.055	0.662±0.021	0.659±0.032	0.64±0.038	0.68±0.046
✓	✗	✓	✓	0.692±0.049	0.705±0.091	0.671±0.045	0.720±0.050	0.647±0.019	0.678±0.037
✓	✓	✗	✓	0.707±0.033	0.714±0.061	0.699±0.019	0.721±0.018	0.654±0.016	0.700±0.043
PT w/o MM			✓	0.642±0.030	0.643±0.022	0.644±0.030	0.640±0.018	0.618±0.013	0.635±0.028
PT w/ MM			✓	0.660±0.017	0.683±0.025	0.662±0.021	0.691±0.036	0.633±0.015	0.639±0.020
C-I	SB	CSA	TCL	CI	AUC	CI	AUC	CI	AUC
✓	✗	✗	✗	0.651±0.020	0.697±0.040	0.653±0.028	0.680±0.022	0.621±0.023	0.630±0.029
✓	✓	✗	✗	0.670±0.034	0.715±0.066	0.673±0.031	0.717±0.063	0.655±0.024	0.668±0.033
✓	✓	✗	✓	0.709±0.025	0.729±0.063	0.693±0.033	0.736±0.029	0.669±0.017	0.696±0.030
✓	✓	✓	✗	0.678±0.031	0.721±0.047	0.680±0.046	0.731±0.044	0.659±0.031	0.680±0.063
✓	✓	✓	✓	**0.717±0.012**	**0.754±0.031**	**0.705±0.015**	**0.745±0.012**	**0.672±0.029**	**0.720±0.039**

Prognosis Inference. For the task of prognosis inference, we adopted SGD optimization with the learning rate of 10^{-4} and the momentum of 0.9, and trained 1000 epochs with batch size 16. The soft-bag size was set to 3 and the hyperparameter β_s was empirically set to 10^{-3}.

4.3 Comparison With the State-of-the-art Methods

We compared our method with other weakly-supervised methods which aim to predict patient outcomes. Table 3 presents the 5-fold cross-validation results in terms of CI and AUC for methods based on clustering [40], graph network [10, 41], and multi-instance learning [5, 6]. We can observe the following important points. First, the proposed DC_MIL outperforms all previous methods on three cancer types in head-to-head comparisons, achieving an overall CI of 0.698 and AUC of 0.74. Three possible reasons are as follows: (1) In our method, instances are encoded in a task-specific weakly supervision, instead of leveraging the dimensionality reduction technique [40] or the pre-trained model on ImageNet [5, 6, 10, 41]; (2) We consider the fine-grained information of WSI and saliency guidance with cross-scale tiles which can lead to extract more discriminative features from low- to high-magnification instances.; (3) We improve the Cox loss with two-tier contrastive learning, which can enhance inter-bag and intra-bag discrimination and help achieve better prognosis inference. Moreover, WSISA [40], MesoNet [6], and DeepAttnMISL [5] outperform DeepGraphSurv [41] and Patch-GCN [10], which may owe to the former methods select discriminant tiles. More results can refer to supplemental material.

4.4 Ablation Study and Parameter Analysis

Abaltion on Curriculum I. First, we conduct experiments on crucial components on Curriculum I. Experimental results are shown in Table 2. Given max-magnification or multi-magnification (MM) inputs, our model is average CI gains of 0.025 and 0.046 and AUC gains of 0.039 and 0.069 ahead of the pre-trained Resnet34 model. Furthermore, compared with the pre-trained model without multi-magnification (MM) inputs, the pre-trained model with MM inputs obtains average CI gain of 0.017 and AUC gain of 0.032 and our model gets CI gain of 0.039 and AUC gain of 0.061. These demonstrate the superiority of our model in encoding multi-magnification instance and the importance of introducing task-specific supervision. In addition, Table 2 shows a substantial drop, average CI of 0.028 and AUC of 0.039 when salient-guided (SG) strategy is removed, and average CI of 0.011 and AUC of 0.028 when fine-turning (FT) strategy is removed. These highlight the importance of saliency-guided and fine-turning strategies.

Abaltion on Curriculum II. We also conducted ablation experiments on Curriculum II to verify the effectiveness of the soft-bag (SB) and constrained self-attention (CSA) modules and two-tier contrastive loss (TCL). In Table 2, we observe that (1) First, the baseline in this curriculum is an attention-based prognosis model with ordinary cox loss, which achieves average CI of 0.642 and AUC of 0.669. Compared with the baseline, the model with soft-bag strategy achieves average CI gain of 0.024 and AUC gain of 0.031 on three datasets, which might mainly profit by identifying and integrating the representative instances within a bag; (2) By introducing the constrained self-attention module, the two-tier contrastive learning module, and both them, the model leads to significantly

Table 3: Results (CI) achieved by the proposed model with different soft-bag sizes, respectively.

	N_B=1	N_B=2	N_B=3	N_B=4	N_B=5
COAD	0.660	0.679	0.717	0.671	0.705
LIHC	0.631	0.691	0.705	0.704	0.705
BLCA	0.648	0.657	0.672	0.686	0.667

Figure 3: Visualization of DC_MIL inputs: The first column is the original H&E stained tissue images, and the second and third columns are the heatmaps of 10× and 20× inputs for DC_MIL, respectively.

better performance with average CI gains of 0.006, 0.024, and 0.032 and AUC gains of 0.011, 0.02, and 0.04 on three datasets, respectively, which benefits from the reduction of intra-bag redundancy the enhancement of intra-bag and inter-bag discrimination. More results of abation experiments can refer to supplemental material.

Influence of the Soft-bag Size. To investigate the influence of the soft-bag size, we conducted a set of experiments via varying the soft-bag size N_B over the set of $\{1, 2, 3, 4, 5\}$ and record the corresponding CI in Table 1, respectively. As can be seen from Table 1, we can observe that: 1) For COAD, the model can achieve higher CI when N_B = 3, and that is N_B = 4 for BLAC; 2) The CI values on LIHC show low differences when N_B is bigger or equal to 3. It indicates that around 3-4 top instances are adequate to represente the bag and more would even lead to redundancy. Besides, the CI values are higher a lot when N_B is greater than 1, which reveals the effectiveness of our proposed soft-bag strategy on prognosis inference.

4.5 Visualization

To visualize and interpret the model predictions, we generated heatmaps by summarizing the saliency-guided maps (M1 and M2) of all instances. As shown in Fig. 3, 20× image mainly focuses on the fine-grained region that is wrapped by the region of interest in the 10× image. It demonstrates the effectiveness of multi-magnification input. Meanwhile, this also indicates that our proposed method can effectively guide the encoding of high-magnification instances, focusing the network's attention on more fine-grained information.

5 Conclusion

In this paper, we propose a dual-curriculum contrastive MIL method for cancer prognosis analysis with WSIs, which contains two easy-to-hard curriculums, i.e., saliency-guided weakly-supervised instance encoding with cross-scale tiles and contrastive-enhanced soft-bag prognosis inference. The proposed network is easy to train in each curriculum, and can be applied to gigapixel WSI analysis with good visualization and interpretability. Besides, the proposed method outperforms the state-of-the-art methods in cancer prognosis analysis over three public datasets.

References

[1] Ole-Johan Skrede, Sepp De Raedt, Andreas Kleppe, Tarjei S Hveem, Knut Liestøl, John Maddison, Hanne A Askautrud, Manohar Pradhan, John Arne Nesheim, Fritz Albregtsen, et al. Deep learning for prediction of colorectal cancer outcome: a discovery and validation study. *The Lancet*, 395(10221):350–360, 2020.

[2] Richard Colling, Helen Pitman, Karin Oien, Nasir Rajpoot, Philip Macklin, CM-Path AI in Histopathology Working Group, Velicia Bachtiar, Richard Booth, Alyson Bryant, Joshua Bull, et al. Artificial intelligence in digital pathology: a roadmap to routine use in clinical practice. *The Journal of pathology*, 249(2):143–150, 2019.

[3] Chetan L Srinidhi, Ozan Ciga, and Anne L Martel. Deep neural network models for computational histopathology: A survey. *Medical Image Analysis*, 67:101813, 2021.

[4] Jana Lipkova, Tiffany Y Chen, Ming Y Lu, Richard J Chen, Maha Shady, Mane Williams, Jingwen Wang, Zahra Noor, Richard N Mitchell, Mehmet Turan, et al. Deep learning-enabled assessment of cardiac allograft rejection from endomyocardial biopsies. *Nature medicine*, 28(3):575–582, 2022.

[5] Jiawen Yao, Xinliang Zhu, Jitendra Jonnagaddala, Nicholas Hawkins, and Junzhou Huang. Whole slide images based cancer survival prediction using attention guided deep multiple instance learning networks. *Medical Image Analysis*, 65:101789, 2020.

[6] Pierre Courtiol, Charles Maussion, Matahi Moarii, Elodie Pronier, Samuel Pilcer, Meriem Sefta, Pierre Manceron, Sylvain Toldo, Mikhail Zaslavskiy, Nolwenn Le Stang, et al. Deep learning-based classification of mesothelioma improves prediction of patient outcome. *Nature medicine*, 25(10):1519–1525, 2019.

[7] Hassan Muhammad, Chensu Xie, Carlie S Sigel, Michael Doukas, Lindsay Alpert, Amber Lea Simpson, and Thomas J Fuchs. Epic-survival: End-to-end part inferred clustering for survival analysis, with prognostic stratification boosting. In *Medical Imaging with Deep Learning*, 2021.

[8] Bin Li, Yin Li, and Kevin W Eliceiri. Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14318–14328, 2021.

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

[10] Richard J Chen, Ming Y Lu, Muhammad Shaban, Chengkuan Chen, Tiffany Y Chen, Drew FK Williamson, and Faisal Mahmood. Whole slide images are 2d point clouds: Context-aware survival prediction using patch-based graph convolutional networks. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 339–349. Springer, 2021.

[11] Ming Y Lu, Drew FK Williamson, Tiffany Y Chen, Richard J Chen, Matteo Barbieri, and Faisal Mahmood. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature biomedical engineering*, 5(6):555–570, 2021.

[12] Zhuchen Shao, Hao Bian, Yang Chen, Yifeng Wang, Jian Zhang, Xiangyang Ji, et al. Transmil: Transformer based correlated multiple instance learning for whole slide image classification. *Advances in Neural Information Processing Systems*, 34, 2021.

[13] Chetan L Srinidhi and Anne L Martel. Improving self-supervised learning with hardness-aware dynamic curriculum learning: An application to digital pathology. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 562–571, 2021.

[14] Chetan L Srinidhi, Seung Wook Kim, Fu-Der Chen, and Anne L Martel. Self-supervised driven consistency training for annotation efficient histopathology image analysis. *Medical Image Analysis*, 75:102256, 2022.

[15] Xueting Yan, Ishan Misra, Abhinav Gupta, Deepti Ghadiyaram, and Dhruv Mahajan. Clusterfit: Improving generalization of visual representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6509–6518, 2020.

[16] 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*, pages 3852–3861, 2020.

[17] Hongrun Zhang, Yanda Meng, Yitian Zhao, Yihong Qiao, Xiaoyun Yang, Sarah E Coupland, and Yalin Zheng. Dtfd-mil: Double-tier feature distillation multiple instance learning for histopathology whole slide image classification. *arXiv preprint arXiv:2203.12081*, 2022.

[18] Trung Vu, Phung Lai, Raviv Raich, Anh Pham, Xiaoli Z Fern, and UK Arvind Rao. A novel attribute-based symmetric multiple instance learning for histopathological image analysis. *IEEE transactions on medical imaging*, 39(10):3125–3136, 2020.

[19] Wei Shao, Tongxin Wang, Zhi Huang, Zhi Han, Jie Zhang, and Kun Huang. Weakly supervised deep ordinal cox model for survival prediction from whole-slide pathological images. *IEEE Transactions on Medical Imaging*, 40(12):3739–3747, 2021.

[20] Nima Hatami, Mohsin Bilal, and Nasir Rajpoot. Deep multi-resolution dictionary learning for histopathology image analysis. *arXiv preprint arXiv:2104.00669*, 2021.

[21] Hiroki Tokunaga, Yuki Teramoto, Akihiko Yoshizawa, and Ryoma Bise. Adaptive weighting multi-field-of-view cnn for semantic segmentation in pathology. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12597–12606, 2019.

[22] Naofumi Tomita, Behnaz Abdollahi, Jason Wei, Bing Ren, Arief Suriawinata, and Saeed Hassanpour. Attention-based deep neural networks for detection of cancerous and precancerous esophagus tissue on histopathological slides. *JAMA network open*, 2(11):e1914645–e1914645, 2019.

[23] Nikhil Naik, Ali Madani, Andre Esteva, Nitish Shirish Keskar, Michael F Press, Daniel Ruderman, David B Agus, and Richard Socher. Deep learning-enabled breast cancer hormonal receptor status determination from base-level h&e stains. *Nature communications*, 11(1):1–8, 2020.

[24] Shujun Wang, Yaxi Zhu, Lequan Yu, Hao Chen, Huangjing Lin, Xiangbo Wan, Xinjuan Fan, and Pheng-Ann Heng. Rmdl: Recalibrated multi-instance deep learning for whole slide gastric image classification. *Medical image analysis*, 58:101549, 2019.

[25] Dmitrii Bychkov, Nina Linder, Riku Turkki, Stig Nordling, Panu E Kovanen, Clare Verrill, Margarita Walliander, Mikael Lundin, Caj Haglund, and Johan Lundin. Deep learning based tissue analysis predicts outcome in colorectal cancer. *Scientific reports*, 8(1):1–11, 2018.

[26] Richard J Chen, Ming Y Lu, Wei-Hung Weng, Tiffany Y Chen, Drew FK Williamson, Trevor Manz, Maha Shady, and Faisal Mahmood. Multimodal co-attention transformer for survival prediction in gigapixel whole slide images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4025, 2021.

[27] Charlie Saillard, Benoit Schmauch, Oumeima Laifa, Matahi Moarii, Sylvain Toldo, Mikhail Zaslavskiy, Elodie Pronier, Alexis Laurent, Giuliana Amaddeo, Hélène Regnault, et al. Predicting survival after hepatocellular carcinoma resection using deep learning on histological slides. *Hepatology*, 72(6):2000–2013, 2020.

[28] Pooya Mobadersany, Safoora Yousefi, Mohamed Amgad, David A Gutman, Jill S Barnholtz-Sloan, José E Velázquez Vega, Daniel J Brat, and Lee AD Cooper. Predicting cancer outcomes from histology and genomics using convolutional networks. *Proceedings of the National Academy of Sciences*, 115(13):E2970–E2979, 2018.

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

[30] Shuai Jiang, George J Zanazzi, and Saeed Hassanpour. Predicting prognosis and idh mutation status for patients with lower-grade gliomas using whole slide images. *Scientific reports*, 11(1):1–9, 2021.

[31] David Tellez, Geert Litjens, Jeroen van der Laak, and Francesco Ciompi. Neural image compression for gigapixel histopathology image analysis. *IEEE transactions on pattern analysis and machine intelligence*, 43(2):567–578, 2019.

[32] Gang Xu, Zhigang Song, Zhuo Sun, Calvin Ku, Zhe Yang, Cancheng Liu, Shuhao Wang, Jianpeng Ma, and Wei Xu. Camel: A weakly supervised learning framework for histopathology image segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10682–10691, 2019.

[33] Fahdi Kanavati, Gouji Toyokawa, Seiya Momosaki, Michael Rambeau, Yuka Kozuma, Fumihiro Shoji, Koji Yamazaki, Sadanori Takeo, Osamu Iizuka, and Masayuki Tsuneki. Weakly-supervised learning for lung carcinoma classification using deep learning. *Scientific reports*, 10(1):1–11, 2020.

[34] Chensu Xie, Hassan Muhammad, Chad M Vanderbilt, Raul Caso, Dig Vijay Kumar Yarlagadda, Gabriele Campanella, and Thomas J Fuchs. Beyond classification: Whole slide tissue histopathology analysis by end-to-end part learning. In *Medical Imaging with Deep Learning*, pages 843–856. PMLR, 2020.

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

[36] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016.

[37] Aaron Van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv e-prints*, pages arXiv–1807, 2018.

[38] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[39] John Fox and Sanford Weisberg. Cox proportional-hazards regression for survival data. *An R and S-PLUS companion to applied regression*, 2002, 2002.

[40] Xinliang Zhu, Jiawen Yao, Feiyun Zhu, and Junzhou Huang. Wsisa: Making survival prediction from whole slide histopathological images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7234–7242, 2017.

[41] Ruoyu Li, Jiawen Yao, Xinliang Zhu, Yeqing Li, and Junzhou Huang. Graph cnn for survival analysis on whole slide pathological images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 174–182. Springer, 2018.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] See the claimed contributions in Introduction.
   (b) Did you describe the limitations of your work? [Yes] See our supplemental material.

(c) Did you discuss any potential negative societal impacts of your work? [Yes] Our work have no negative societal impacts.

(d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes] We have ensured that our paper conforms to them.

2. If you are including theoretical results...

(a) Did you state the full set of assumptions of all theoretical results? [N/A]

(b) Did you include complete proofs of all theoretical results? [N/A]

3. If you ran experiments...

(a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] See Section4.1 and the code is avaliable at GitHub.

(b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Section4.2

(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No] Error bars are not reported because it would be too computationally expensive. Besides, many previous works on medical report generation did not provide error bars as well.

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Section 4.2.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

(a) If your work uses existing assets, did you cite the creators? [Yes] See Section 4.1.

(b) Did you mention the license of the assets? [Yes]

(c) Did you include any new assets either in the supplemental material or as a URL? [No]

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes] See Section 4.1.The used datasets are publicly available. All necessary patient/participant consent has been obtained and the appropriate institutional forms have been archived.

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes]

5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]