DCMIL: A Progressive Representation Learning of Whole Slide Images for Cancer Prognosis Analysis

Chao Tu, Kun Huang, Jie Zhang, Qianjin Feng, Yu Zhang, and Zhenyuan Ning

***Abstract*— The burgeoning discipline of computational pathology shows promise in harnessing whole slide images (WSIs) to quantify morphological heterogeneity and develop objective prognostic modes for human cancers. However, progress is impeded by the computational bottleneck of gigapixel-size inputs and the scarcity of dense manual annotations. Current methods often overlook fine-grained information across multi-magnification WSIs and variations in tumor microenvironments. Here, we propose an easy-to-hard progressive representation learning, termed dual-curriculum contrastive multi-instance learning (DCMIL), to efficiently process WSIs for cancer prognosis. The model does not rely on dense annotations and enables the direct transformation of gigapixel-size WSIs into outcome predictions. Extensive experiments on twelve cancer types (5,954 patients, 12.54 million tiles) demonstrate that DCMIL outperforms standard WSI-based prognostic models. Additionally, DCMIL identifies fine-grained prognosis-salient regions, provides robust instance uncertainty estimation, and captures morphological differences between normal and tumor tissues, with the potential to generate new biological insights. All codes have been made publicly accessible at https://github.com/tuuuc/DCMIL.**

***Index Terms*— Computational Pathology, Whole Slide Images, Cancer, Prognostic analysis, Multi-instance Learning, Contrastive learning**

This work was supported in part by the National Natural Science Foundation of China under Grant U22A20350 and Grant 61971213, in part by the Basic and Applied Basic Research Foundation of Guangdong Province under Grant 2019A1515010417, in part by the Guangdong Provincial Key Laboratory of Medical Image Processing under Grant No.2020B1212060039, and in part by the Special Funds for the Cultivation of Guangdong College Students' Scientific and Technological Innovation under Grant No.pdjh2024a087 and Grant No.pdjh2023a0099. *(Corresponding authors: Zhenyuan Ning; Yu Zhang.)*

Chao Tu is with the School of Biomedical Engineering, Southern Medical University, Guangzhou 510515 and also with the Department of Pathology and Institute of Molecular Pathology, The First Affiliated Hospital, Jiangxi Medical College, Nanchang University, Nanchang 330006, China (e-mail: tuu23@foxmail.com)

Qianjin Feng, Yu Zhang, and Zhenyuan Ning is with the Guangdong Provincial Key Laboratory of Medical Image Processing, Guangdong Province Engineering Laboratory for Medical Imaging and Diagnostic Technology, School of Biomedical Engineering, Southern Medical University, Guangzhou, Guangdong 510515, China (e-mail: fengqj99@smu.edu.cn; yuzhang@smu.edu.cn; jonnyning@foxmail.com).

Kun Huang is with the Department of Biostatistics and Health Data Science, School of Medicine, Indiana University, Indianapolis, IN 46202, USA (e-mail: kunhuang@iu.edu).

Jie Zhang is with the Department of Medical and Molecular Genetics, School of Medicine, Indiana University, Indianapolis, IN 46202, USA (e-mail: jizhan@iu.edu).

I. Introduction

CANCER, as a highly complex disease, exhibits significant heterogeneity in the tumor microenvironment that dominates the variability in patient outcomes [1], [2]. Cancer prognosis analysis aims to stratify patients into different risk subgroups based on quantifying the heterogeneous characteristics of tumor [3]. Precise prognosis estimation can aid physicians in decision-making to optimize treatment strategies and improve patients' quality of life [4]. Clinically, a general paradigm involves manual assessment of certain criteria (e.g., tumor stage) to evaluate patient prognosis based on histopathological features (e.g., tumor invasion, necrosis, and mitosis) or biomarkers (e.g., the morphology and location of tumor cells) driven from the histopathological slides [5]. However, subjective evaluation may be subject to high inter-observer and intra-observer variability, and the identification of histopathological features or biomarkers typically requires labor-intensive efforts from pathologists. Moreover, substantial variation in prognosis persists among patients, regardless of their classification into the same grade or stage [6]. Therefore, developing an objective, accurate, and robust prognostic model in this field remains critical.

Recently, high-resolution whole slide image (WSI) using digital slide scanners has triggered tremendous excitement for the burgeoning field of computational pathology [7], [8]. Exploring the rich morphological information from WSIs, computational pathology has demonstrated promise in quantifying histopathological heterogeneity of the tumor microenvironment and is competent in many tasks towards precision medicine (e.g., cancer diagnosis [9], [10], biomarker discovery [11], [12], therapy and drug development [13], [14], and outcome prediction [6], [15]). The emergence of artificial intelligence technology, especially deep learning approach, has further advanced computational pathology, which revolutionized the routine WSI analysis paradigm [16], [17]. Current deep learning methods have been successfully applied to cancer prognosis analysis by learning latent prognosis-relevant morphological representations from WSIs [18], [19]. However, those methods have been impeded due to the computational bottleneck of gigapixel-size WSIs and the scarcity of manual dense annotations. Although some studies try to alleviate this issue by taking small-sized region of interest (ROI) selectively sampled from WSIs as model's input, such scheme is experience-dependent and requires prior knowledge from

pathologists [20]. Also, how to manually define prognosis-related ROIs is still an open issue, as such regions generally distribute throughout the tumor and its microenvironment.

Multi-instance learning (MIL) method provides a feasible solution to perform inference in a weakly-supervised manner, which has attracted increasing attention for WSI analysis in cancer prognosis estimation [21], [22]. MIL method adopts a divide-and-conquer strategy and typically consists of two stages, i.e., instance encoding stage and instance aggregation stage [9]. In the instance encoding stage, previous methods have utilized pre-trained or foundation models to extract instance representations in an unsupervised or self-supervision way (i.e., without the reference of outcome labels) [23], [24]. 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 [25]. Alternatively, some work has directly assigned each instance the bag-level annotation (i.e., the survival time of patient) for network training, which introduces task-specific supervision and improves model's generalizability [26]. Actually, the prognosis evaluation is primarily and comprehensively determined by certain representative regions, yet those regions might only occupy a small portion of WSI [27]. As a result, the strong supervision on all instances, especially on prognosis-irrelevant tiles, will introduce excessive label noises and limit model's performance. To alleviate this issue, a feasible way is to supervise the instance encoding in a weak and easy-to-hard manner. The second stage of instance aggregation generally involves aggregating instance representations within a bag for prognosis estimation by certain fusion strategies (e.g., pooling operation and attention mechanism) [28]. Although such strategies are simple and straightforward, they may overwhelm prognosis-relevant information, cause intra-bag redundancy, and reduce inter-bag discrimination, when many instances from irrelevant regions are enrolled [29]. Additionally, the above-mentioned MIL methods primarily focus on single-scale local instance representations [10], [30], such as the morphological changes or different growth patterns of tumor cells. However, pathologists typically identify salient regions at a low-magnification field of view, and then transition to a higher-magnification level to examine fine-grained histopathological features within the context and assess differences in tumor microenvironments to discern better the nature and extent of cancer cells [8], [31].

In this article, we extend our previous approach [32] and propose a progressive representation learning called Dual-Curriculum Contrastive Multi-Instance Learning (DCMIL) for efficient processing of whole slide images (WSIs) in cancer prognosis analysis. DCMIL incorporates two curriculums designed to learn robust representations in an easy-to-hard manner[1]. For the first curriculum, DCMIL introduces prognosis information by assigning risk stratification status (degraded from survival time) as instance labels to learn instance-level representations in a weakly-supervised manner, which helps reduce label noises and maintain prognosis-related guidance. Also, to imitate the reviewing procedure of pathologists, DCMIL leverages the low-magnification saliency map to guide the encoding of high-magnification instances for exploring fine-grained information across multi-magnification WSIs. For the second curriculum, instead of enrolling all instances, DCMIL adaptively identifies and integrates representative instances within a bag (as the soft-bag) for prognosis inference. It leverages the constrained self-attention strategy to obtain extra sparseness for soft-bag representations, which can help reduce intra-bag redundancy at both instance and feature levels. Furthermore, contrastive learning is a self-supervised technique that learns effective representations by contrasting positive and negative samples and has been proven effective in various machine learning tasks [33], [34]. Consequently, DCMIL arms with triple-tier contrastive learning strategy for enhancing intra-bag (i.e., high-salient and low-salient regions within a subject) and inter-bag (i.e., ① tumor and normal tissues ② high-risk and low-risk regions between different subjects) discrimination. Extensive experiments show that DCMIL outperforms standard WSI-based prognostic models on twelve cancer types. Furthermore, DCMIL provides fine-grained prognosis-salient regions on WSIs, presents robust instance uncertainty estimation, and captures the morphological differences between tumor and normal tissue microenvironments, with the potential to generate new biological insights.

[1]The main differences from the conference version are detailed in the Method section.

II. Materials and Methods

A. Study Population

We developed and evaluated our proposed method on 12 datasets collected from The Cancer Genome Atlas (TCGA) database, including Bladder Urothelial Carcinoma (BLCA, $N = 395$), Breast Invasive Carcinoma (BRCA, $N = 993$), Colon Adenocarcinoma (COAD, $N = 352$), Head and Neck Squamous Cell Carcinoma (HNSC, $N = 430$), Kidney Renal Clear Cell Carcinoma (KIRC, $N = 440$), Liver Hepatocellular Carcinoma (LIHC, $N = 333$), Lung Adenocarcinoma (LUAD, $N = 438$), Lung Squamous Cell Carcinoma (LUSC, $N = 425$), Ovarian Serous Cystadenocarcinoma (OV, $N = 528$), Stomach Adenocarcinoma (STAD, $N = 370$), Thyroid Carcinoma (THCA, $N = 501$), and Uterine Corpus Endometrial Carcinoma (UCEC, $N = 504$). Each dataset contained WSIs stained with hematoxylin and eosin (H&E) along with clinical information (i.e., survival time and event status). After segmenting tissue areas using Otsu's method, a set of non-overlapping tiles at different magnifications ($20\times$, $10\times$, $5\times$) were sampled from each segmented tissue area, with window sizes set to 512×512, 256×256, and 128×128, respectively.

B. Modeling Framework

1) Problem Formulation: The main notations used in this paper are summarized in Table S1 (in *Supplementary Materials*) for reference. Denote the triplet $\{\mathbf{X}_n, \mathbf{T}_n, \boldsymbol{\delta}_n\}_{n=1}^{N}$ as the dataset consisting of N patients, where $\mathbf{X}_n$ is the n-th patient with one or more WSIs, and $\mathbf{T}_n$ and $\boldsymbol{\delta}_n$ are respectively its observed survival time and event indicator (viz. equals to 1 and 0 for uncensored and censored patients, respectively).

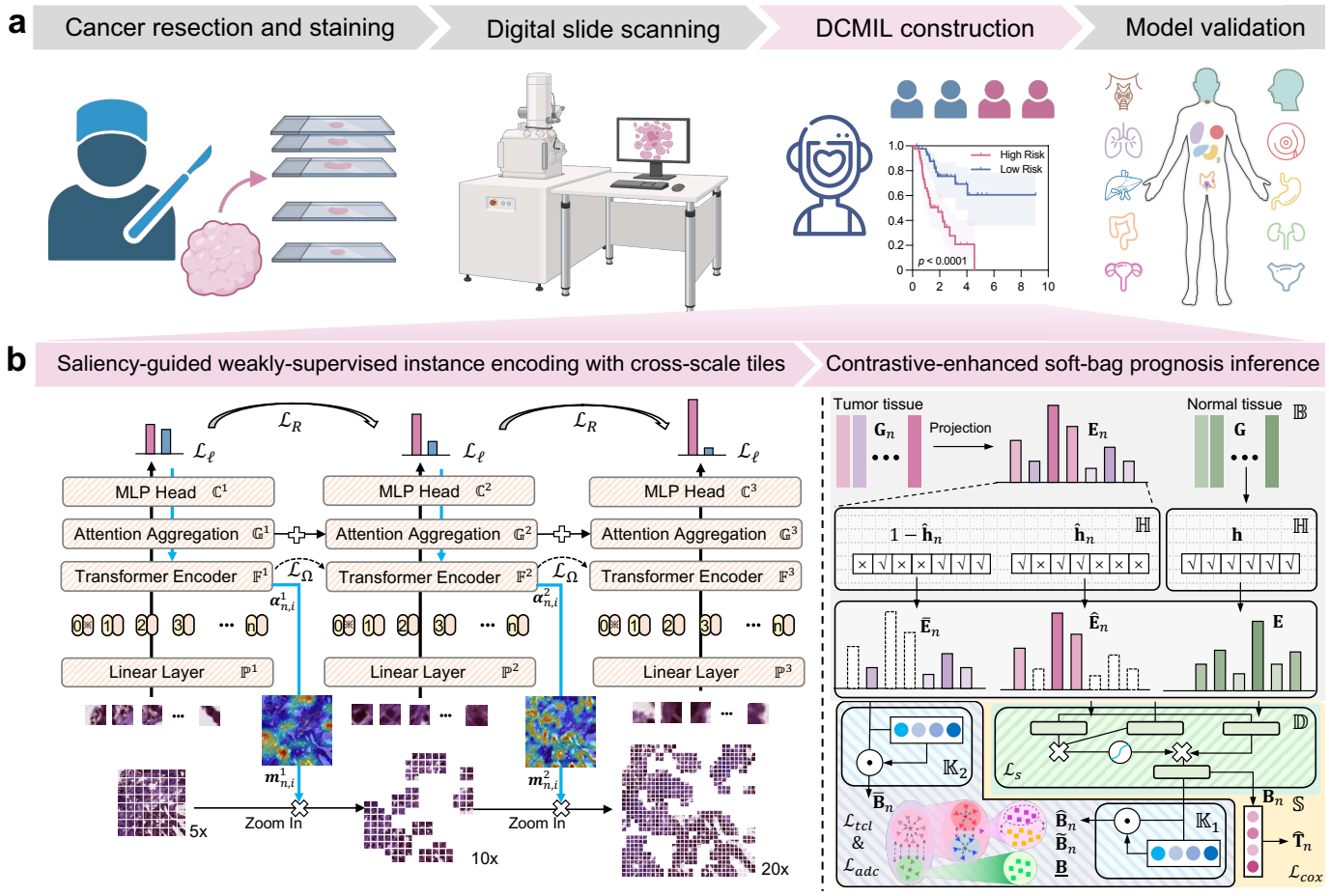

Fig. 1. **Overview of the framework. a**, The overall workflow, including cancer resection and staining, digital slide scanning, DCMIL construction, and model validation. **b**, The pipeline of DCMIL comprises two easy-to-hard curriculums, including 1) saliency-guided weakly-supervised instance encoding with cross-scale tiles, and 2) contrastive-enhanced soft-bag prognosis inference.

Prognosis analysis is an ordinal regression task that models time-to-event distribution, which is formulated as

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

where $\mathbb{S}$ and $\widehat{\mathbf{T}}_n$ denote the prognosis inference function and the estimated survival time, respectively. In practice, it is difficult to directly feed the WSI into the model due to its gigapixel size. Accordingly, many studies [23], [35] have decomposed the WSI into a large amount of tiles and generated a bag $\mathbf{X}_n = \{\mathbf{x}_{n,i}\}_{i=1}^{N_n}$ that contains N_n instances (tiles) for the n-th patient. Then, MIL is leveraged to construct the prognosis model as follows:

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

where $\mathbb{E}$ is an instance encoding function that extracts feature representation for each instance, and $\mathbb{A}$ is a permutation-invariant instance aggregation function that pools the instance representations into the bag ones. However, previous studies [26], [27] 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 model into two easy-to-hard curriculums, including (1) Curriculum I (C-I): instance encoding (via $\mathbb{E}$) during preliminary risk stratification (via $\mathbb{C}$, a risk stratification function), and (2) Curriculum II (C-II): prognosis inference (via $\mathbb{S}$) after instance aggregation (via $\mathbb{A}$). Formally, the optimization process can be formulated as

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

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

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

2) Curriculum I: Saliency-guided weakly-supervised instance encoding with cross-scale tiles: The pipeline of Curriculum I is shown in Figure 1b. Given the input instance with S magnifications $\{\mathbf{x}_{n,i}^s\}_{s=1}^{S}$ ($S = 3$ in this work), the proposed method contains S branches and each branch takes different-magnification tiles as input, which aims to explore multi-scale information from multi-magnification images. For the s-th branch, it mainly consists of a linear layer $\mathbb{P}^s$, a transformer

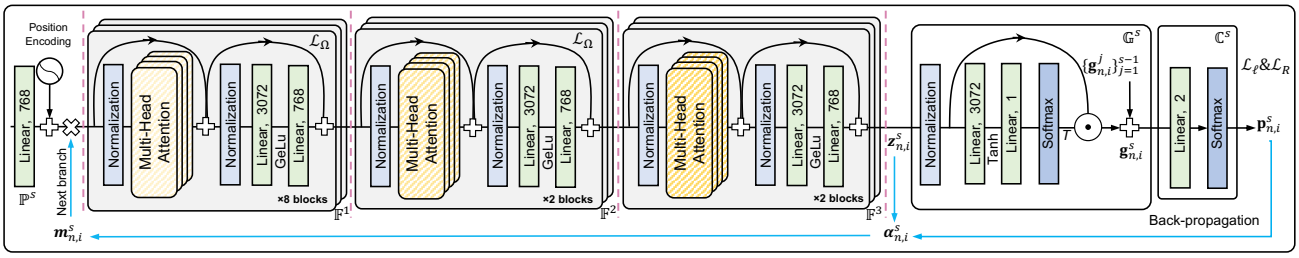

Fig. 2. The detailed architectures on Curriculum I. For the s-th branch, its architecture mainly consists of a linear layer $\mathbb{P}^s$, a transformer encoder $\mathbb{F}^s$, an attention aggregator $\mathbb{G}^s$ and a multilayer perceptron (MLP) head $\mathbb{C}^s$.

encoder $\mathbb{F}^s$, an attention aggregator $\mathbb{G}^s$ and a multilayer perceptron (MLP) head $\mathbb{C}^s$. $\{\mathbb{P}^s\}_{s=1}^S$, $\{\mathbb{F}^s\}_{s=1}^S$, and $\{\mathbb{G}^s\}_{s=1}^S$ form $\mathbb{E}$, while $\{\mathbb{C}^s\}_{s=1}^S$ constitutes $\mathbb{C}$ in Eq.(3). Specifically, the input instance (i.e., tile) is first divided into several 16×16 tokens. As exhibited in Figure 2, $\mathbb{P}^s$ is used to map the flattened tokens into 768-dimensional tokens. Following position encoding, $\mathbb{F}^s$ is placed to extract token representation, which is a ViT-like network architecture [36] with multiple blocks. Each block contains a multi-head attention layer, an MLP, and two normalization layers. The MLP comprises two linear layers with 3072 and 768 nodes, respectively, in which "GeLu" activation function is inserted between the two linear layers. Two normalization layers are respectively positioned before the multi-head attention layer and the MLP. Next, $\mathbb{G}$ processes all tokens through a normalization layer, an MLP, and a "Softmax" layer to obtain token attention. The MLP comprises two linear layers with 3072 and 1 nodes, respectively, in which "Tanh" activation function is inserted between the two linear layers. The tokens are then aggregated based on token attention, and the tile representation from the last scale is summarized through a skip connection, yielding the tile representation for this scale. Finally, $\mathbb{C}$ predicts the risk stratification status using a linear layer followed by a "Softmax" layer.

Different from previous studies [37] that separately feed $\{\mathbf{x}_{n,i}^s\}_{s=1}^S$ into the network and ignore fine-grained details across multi-magnification images, we encourage the network to focus on saliency regions by utilizing the prior knowledge of the $(s-1)$-th branch to highlight the input $\mathbf{x}_{n,i}^s$, which can be formulated as

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

where $\otimes$ denotes the element-wise multiplication operator, and $\mathbf{m}_{n,i}^{s-1}$ is the saliency mask indicating high-risk regions, computed by gradient class activation map (gradCAM) [38]. For convenience, we introduce how to acquire $\mathbf{m}_{n,i}^s$, which is easily generalized to $\mathbf{m}_{n,i}^{s-1}$. Specifically, for each branch, we first generate an instance token embedding $\mathbf{z}_{n,i}^s$ through $\mathbb{P}^s$ and $\mathbb{F}^s$:

$$\mathbf{z}_{n,i}^s = \mathbb{F}^s(\mathbb{P}^s(\widehat{\mathbf{x}}_{n,i}^s)). \tag{6}$$

Then, we aggregate $\mathbf{z}_{n,i}^s$ to obtain the instance representation $\mathbf{g}_{n,i}^s$ via $\mathbb{G}^s$, while assimilating comprehensive information from the current and preceding branches, which is computed by

$$\mathbf{g}_{n,i}^s = \mathbb{G}^s(\mathbf{z}_{n,i}^s, \mathbf{g}_{n,i}^{s-1}). \tag{7}$$

And $\mathbf{g}_{n,i}^s$ is used to predict the risk stratification status through $\mathbb{C}^s$:

$$\mathbf{p}_{n,i}^s = \mathbb{C}^s(\mathbf{g}_{n,i}^s), \tag{8}$$

where $\mathbf{p}_{n,i}^s$ denotes the probability of risk stratification status. After completing the forward propagation, we can obtain the gradient of each token through backward propagation, thereby generating the saliency mask $\mathbf{m}_{n,i}^s$:

$$\alpha_{n,i}^s = \frac{\partial \mathbf{p}_{n,i}^s}{\partial \mathbf{z}_{n,i}^s}, \tag{9}$$

$$\mathbf{m}_{n,i}^s = \mathbb{R}(\alpha_{n,i}^s \mathbf{z}_{n,i}^s) \geq \iota, \tag{10}$$

where $\alpha_{n,i}^s$ is the gradient of the score for $\mathbf{p}_{n,i}^s$, $\mathbb{R}$ represents the "ReLU" activation function and ι is a threshold set at 0.4.

The empirical loss $\mathcal{L}_\ell$ for Curriculum I is defined as

$$\begin{aligned} \mathcal{L}_\ell = -\frac{1}{NN_nS} \sum_{n=1}^{N} \sum_{i=1}^{N_n} \sum_{s=1}^{S} & \mathbf{y}_{n,i}^s \log(\mathbf{p}_{n,i}^s) \\ & + (1 - \mathbf{y}_{n,i}^s) \log(1 - \mathbf{p}_{n,i}^s), \end{aligned} \tag{11}$$

where $\mathbf{y}_{n,i}^s \in \mathbf{Y}_n$ denotes the risk stratification status of the i-th instance in the n-th bag. Furthermore, we introduce a hierarchical transfer strategy to leverage the prior from different-magnification images, accelerate network convergence, and maintain the specificity of the current branch. Specifically, as shown in Figure 11b, the parameter $\boldsymbol{\theta}_{\mathbb{F}^s}^s$ (for any s) of the s-th block in $\mathbb{F}^s$ is completely learnable and has not any constraints, while $\{\boldsymbol{\theta}_{\mathbb{F}^s}^i\}_{i=1}^{s-1}$ (for $s > 1$) is also learnable but constrained to approximate $\{\boldsymbol{\theta}_{\mathbb{F}^{s-1}}^i\}_{i=1}^{s-1}$ by the structural loss $\mathcal{L}_\Omega$ as follows:

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

To encourage the finer-scale branch to take more confident prediction from the coarse-scale branch as references, we further introduce a pairwise ranking loss $\mathcal{L}_R$ that has the following definition:

$$\begin{aligned} \mathcal{L}_R = \sum_{n=1}^{N} \sum_{i=1}^{N_n} \sum_{s=2}^{S} & \max(0, \eta - \mathbf{y}_{n,i}^s \log \frac{\mathbf{p}_{n,i}^s}{\mathbf{p}_{n,i}^{s-1}} \\ & - (1 - \mathbf{y}_{n,i}^s) \log(\frac{1 - \mathbf{p}_{n,i}^s}{1 - \mathbf{p}_{n,i}^{s-1}})), \end{aligned} \tag{13}$$

where η is a hyperparameter set to $1e^{-3}$ to make the prediction in the finer-scale branch more confident. Intuitively, when the instance label is high-risk, the high-risk prediction probability in the fine-scale branch is greater than that in the coarse-scale branch. The same applies to low-risk instances. Finally,

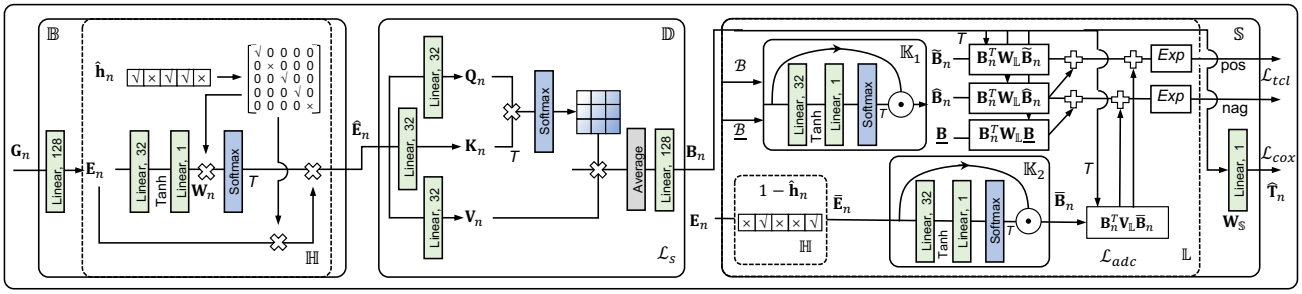

Fig. 3. The detailed architectures on Curriculum II, containing a soft-bag learning module $\mathbb{B}$, a constrained self-attention module $\mathbb{D}$, and a contrastive-enhanced prognosis inference function $\mathbb{S}$.

a hybrid loss function $\mathcal{L}_{\mathrm{I}}$ is designed to train the network, which contains three parts as follows:

$$\mathcal{L}_{\mathrm{I}} = \mathcal{L}_{\ell} + \beta_{\Omega}\mathcal{L}_{\Omega} + \beta_{R}\mathcal{L}_{R}, \tag{14}$$

where β_{Ω} and β_{R} are the weight coefficients. After network optimization, we can obtain $\mathbf{g}^{S}_{n,i}$ as multi-scale instance representation $\mathbf{g}_{n,i}$ (i.e., $\mathbf{g}_{n,i} = \mathbf{g}^{S}_{n,i}$, as $\mathbf{g}^{S}_{n,i}$ encompasses comprehensive information from the current and preceding branches by E.q.(7)) for subsequent prognosis inference task.

3) Curriculum II: Contrastive-enhanced soft-bag prognosis inference: In this section, we develop the second curriculum, and its pipeline is shown in Figure 3. Given a set of instance representations from the n-th bag, i.e., $\mathbf{G}_n = [\mathbf{g}_{n,1}; \mathbf{g}_{n,2}; \cdots; \mathbf{g}_{n,N_n}]^T \in \mathbb{R}^{N_n \times C}$, where C denotes the feature dimension, 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 $\mathbb{B}$ (as shown in Figure 11d), it first learns new bag representation $\mathbf{E}_n \in \mathbb{R}^{N_n \times D}$ (where D denotes the feature dimension) from $\mathbf{G}_n$ via a linear layer. Instead of enrolling all instances, we adaptively identify and integrate representative instances within a bag by introducing the function $\mathbb{H}$:

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

where $\widehat{\mathbf{E}}_n \in \mathbb{R}^{N_B \times D}$ (N_B is the number of selected instances in each bag, and $N_B \ll N_n$) is the feature representation of those selected instances, ϕ represents the "Softmax" activation function, and $\mathbf{W}_n \in \mathbb{R}^{N_n \times N_B}$ denotes the weight matrix that is generated by feeding $\mathbf{E}_n$ into two linear layers (as shown in Figure 11d). And $\widehat{\mathbf{h}}_n \in \mathbb{R}^{N_n \times 1}$ is an indicator that adaptively selects N_B representative instances and is computed by

$$\begin{aligned} \widehat{\mathbf{h}}_n =& \arg\max_{\mathbf{h}_n} \mathbb{S}\left(\mathbb{D}\left(\mathbb{H}(\mathbf{E}_n, \mathbf{h}_n)\right)\right), \\ & s.t. \sum_{i=1}^{N_n} \mathbf{h}_{n,i} = N_B \;\&\&\; \mathbf{h}_{n,i} \in \{0, 1\}. \end{aligned} \tag{16}$$

Since the top-N_B assignment operation via "$\arg\max$" is not differentiable, we instead use Gumbel-Softmax strategy [39] to compute $\widehat{\mathbf{h}}_n$.

Subsequently, we leverage $\mathbb{D}$ to obtain extra sparseness for soft-bag representations, as shown in Figure 11d. Given the input $\widehat{\mathbf{E}}_n$, it first linearly projects $\widehat{\mathbf{E}}_n$ into three feature spaces (i.e., $\mathbf{Q}_n \in \mathbb{R}^{N_B \times \widehat{D}}$, $\mathbf{K}_n \in \mathbb{R}^{N_B \times \widehat{D}}$, and $\mathbf{V}_n \in \mathbb{R}^{N_B \times \widehat{D}}$, and $\widehat{D}$ is the feature dimension) via three mapping matrices (i.e., $\mathbf{W}_Q \in \mathbb{R}^{D \times \widehat{D}}$, $\mathbf{W}_K \in \mathbb{R}^{D \times \widehat{D}}$, and $\mathbf{W}_V \in \mathbb{R}^{D \times \widehat{D}}$), in which the mapping matrices are constrained by

$$\mathcal{L}_s = (\|\mathbf{W}_Q\|_1 + \|\mathbf{W}_K\|_1 + \|\mathbf{W}_V\|_1). \tag{17}$$

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 correlation-based activation mechanism [40], as illustrated in Figure 11d. It is worth mentioning that the collaboration of $\mathbb{B}$ and $\mathbb{D}$ can significantly help reduce intra-bag redundancy at both instance and feature levels so as to improve model's generalization.

For $\mathbb{S}$, it takes $\mathbf{B}_n$ as input to infer the relative risk $\widehat{\mathbf{T}}_n$ via a linear projection matrix $\mathbf{W}_{\mathbb{S}} \in \mathbb{R}^{D_B \times 1}$, as shown in Figure 11d. In order to boost the ability of soft-bag inference, we introduce both tumor tissue and normal samples with $\mathcal{O}$ and $\underline{\mathcal{O}}$ distributions, respectively, and perform triple-tier contrastive learning to enhance intra-bag and inter-bag discrimination. We assume that a bag $\mathbf{B}_n$ is sampled from $\mathcal{O} = \{\mathcal{O}^+, \mathcal{O}^-\}$, where the positive source $\mathcal{O}^+$ denotes the distribution of high-risk samples (i.e., those with survival time $\leq \mathbf{T}_r$), while the negative source $\mathcal{O}^-$ presents the distribution of low-risk samples (i.e., those with survival time $> \mathbf{T}_r$). Additionally, we define $\mathcal{B}^{\mathcal{O}}_n$ as the collection of bags (excluding $\mathbf{B}_n$) from the source $\mathcal{O}$, and $\mathcal{B}^{\underline{\mathcal{O}}}$ as the one from the source $\underline{\mathcal{O}}$. Accordingly, $\mathcal{B}^{\mathcal{O}^+}_n$ (or $\mathcal{B}^{\mathcal{O}^-}_n$) is defined as the collection of bags from $\mathcal{O}^+$ (or $\mathcal{O}^-$) with the same source as $\mathbf{B}_n$. We first focus on inter-bag (i.e., ① tumor and normal tissues ② high-risk and low-risk subjects) discrimination. For the former, it is expected to distinguish the tumor tissue from normal tissue. Specifically, we obtain the integrated representations of $\mathcal{B}^{\mathcal{O}}_n$ and $\mathcal{B}^{\underline{\mathcal{O}}}$ (denoted as $\widetilde{\mathbf{B}}_n$ and $\underline{\mathbf{B}}$[2], respectively) by merging the representations of all bags in $\mathcal{B}^{\mathcal{O}}_n$ and $\mathcal{B}^{\underline{\mathcal{O}}}$ through a bag aggregation network $\mathbb{K}_1$ (as illustrated in Figure 11d). Inspired by [41], inter-bag discrimination (towards tumor and normal tissues) can be enhanced by maximally preserving the mutual information between $\mathbf{B}_n$ and $\widetilde{\mathbf{B}}_n$ so as to "pick up" $\widetilde{\mathbf{B}}_n$ from normal tissues:

$$\sum_{n \in [N]} \mathbb{I}(\mathbf{B}_n, \widetilde{\mathbf{B}}_n) = \sum_{n \in [N]} p(\mathbf{B}_n, \widetilde{\mathbf{B}}_n) \log \frac{p(\mathbf{B}_n | \widetilde{\mathbf{B}}_n)}{p(\mathbf{B}_n)}. \tag{18}$$

Similarly, we obtain the representation of $\mathcal{B}^{\mathcal{O}^+}_n$ (or $\mathcal{B}^{\mathcal{O}^-}_n$) (denoted as $\widehat{\mathbf{B}}_n$) via $\mathbb{K}_1$. And inter-bag discrimination (towards

[2]The normal tissue samples did not participate in the training of risk stratification status in Curriculum I but were directly used for inference.

high-risk and low-risk subjects) can be formulated as

$$\sum_{n\in[N]} \mathbb{I}(\mathbf{B}_n, \widehat{\mathbf{B}}_n) = \sum_{n\in[N]} p(\mathbf{B}_n, \widehat{\mathbf{B}}_n) \log \frac{p(\mathbf{B}_n|\widehat{\mathbf{B}}_n)}{p(\mathbf{B}_n)}. \quad (19)$$

Subsequently, we concentrate on intra-bag (i.e., high-salient and low-salient regions within a subject) discrimination. According to Eq.(15), partial instances within a bag are adaptively discarded, while the rest are retained. However, all instances within the bag belong to homologous tissues sampled from the same patient, which means that it should show a potential correlation among these instances. Therefore, we maximize the following mutual information to improve intra-bag discrimination:

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

where $\overline{\mathbf{B}}_n$ is obtained via the instance aggregation network $\mathbb{K}_2$ with the same architecture as $\mathbb{K}_1$, to merge the representations of those discarded instances as $\overline{\mathbf{E}}_n$ that has the definition of $\overline{\mathbf{E}}_n = \mathbb{H}(\mathbf{E}_n, \mathbf{1} - \widehat{\mathbf{h}}_n)$. To this end, we propose to jointly maximize Eqs.(18)-(20), i.e., $max(\sum_{n\in[N]} \mathbb{I}(\mathbf{B}_n, \widetilde{\mathbf{B}}_n) + \mathbb{I}(\mathbf{B}_n, \widehat{\mathbf{B}}_n) + \mathbb{I}(\mathbf{B}_n, \overline{\mathbf{B}}_n))$, which is formulated by

$$\begin{aligned} &\max\Big(\sum_{n\in[N]} \log \frac{p(\mathbf{B}_n|\widetilde{\mathbf{B}}_n)}{p(\mathbf{B}_n)} + \log \frac{p(\mathbf{B}_n|\widehat{\mathbf{B}}_n)}{p(\mathbf{B}_n)} + \log \frac{p(\mathbf{B}_n|\overline{\mathbf{B}}_n)}{p(\mathbf{B}_n)}\Big) \\ &= \max\Big(\sum_{n\in[N]} \log \frac{p(\mathbf{B}_n|\widetilde{\mathbf{B}}_n)p(\mathbf{B}_n|\widehat{\mathbf{B}}_n)p(\mathbf{B}_n|\overline{\mathbf{B}}_n)}{p(\mathbf{B}_n)p(\mathbf{B}_n)p(\mathbf{B}_n)}\Big). \end{aligned} \quad (21)$$

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

$$\begin{aligned} &\mathbb{L}(\mathbf{B}_n, \widetilde{\mathbf{B}}_n, \widehat{\mathbf{B}}_n, \overline{\mathbf{B}}_n) \\ &= exp(\mathbf{B}_n^T \mathbf{W}_{\mathbb{L}} \widetilde{\mathbf{B}}_n) exp(\mathbf{B}_n^T \mathbf{W}_{\mathbb{L}} \widehat{\mathbf{B}}_n) exp(\mathbf{B}_n^T \mathbf{V}_{\mathbb{L}} \overline{\mathbf{B}}_n) \\ &= exp(\mathbf{B}_n^T \mathbf{W}_{\mathbb{L}} \widetilde{\mathbf{B}}_n + \mathbf{B}_n^T \mathbf{W}_{\mathbb{L}} \widehat{\mathbf{B}}_n + \mathbf{B}_n^T \mathbf{V}_{\mathbb{L}} \overline{\mathbf{B}}_n) \\ &\propto \frac{p(\mathbf{B}_n|\widetilde{\mathbf{B}}_n)p(\mathbf{B}_n|\widehat{\mathbf{B}}_n)p(\mathbf{B}_n|\overline{\mathbf{B}}_n)}{p(\mathbf{B}_n)p(\mathbf{B}_n)p(\mathbf{B}_n)}, \end{aligned} \quad (22)$$

where $\mathbf{W}_{\mathbb{L}}$ and $\mathbf{V}_{\mathbb{L}}$ are trainable parameters. Motivated by [42], we design the triple-tier contrastive learning loss based on the InfoNCE function, which has the following definition:

$$\mathcal{L}_{tcl} = -\mathop{\boldsymbol{E}}_{\mathcal{B}} \left[\log \frac{\mathbb{L}(\mathbf{B}_n, \widetilde{\mathbf{B}}_n, \widehat{\mathbf{B}}_n, \overline{\mathbf{B}}_n)}{\mathbb{L}(\mathbf{B}_n, \widetilde{\mathbf{B}}_n, \widehat{\mathbf{B}}_n, \overline{\mathbf{B}}_n) + \sum_{i\in[N], i\neq n} \mathbb{L}(\mathbf{B}_i, \underline{\mathbf{B}}, \widehat{\mathbf{B}}_n, \overline{\mathbf{B}}_n)} \right], \quad (23)$$

where $\mathcal{B}$ denotes the collection of all bags. We provide proof that decreasing $\mathcal{L}_{tcl}$ is equivalent to increasing the lower bound on the Eq.(21)(in *Supplementary Materials*). Furthermore, we introduce the following absolute distance constraint to make bag representation more compact:

$$\begin{aligned} \mathcal{L}_{adc} = \sum_{i\in[N]} &\max(0, \mathrm{Cos}(\mathbf{B}_n, \underline{\mathbf{B}}) - \mathrm{Cos}(\mathbf{B}_n, \widetilde{\mathbf{B}}_\mathbf{n}) + \kappa) \\ &+ 1 - \mathrm{Cos}(\mathbf{B}_n, \widehat{\mathbf{B}}_n) + 1 - \mathrm{Cos}(\mathbf{B}_n, \overline{\mathbf{B}}_n), \end{aligned} \quad (24)$$

where $\kappa = 1$ refers to the margin that indicates the distance between the distribution of tumor tissue and normal tissue.

We utilize the negative log-partial likelihood of Cox proportional hazard regression model [43] to train the network, which is computed by

$$\begin{aligned} \mathcal{L}_{cox} &= -\log\Big(\prod_{n:\boldsymbol{\delta}_n=1} \frac{exp(\mathbf{B}_n \mathbf{W}_{\mathbb{S}})}{\sum_{j\in R(\mathbf{T}_n)} exp(\mathbf{B}_j \mathbf{W}_{\mathbb{S}})}\Big) \\ &= -\sum_{n:\boldsymbol{\delta}_n=1} (\mathbf{B}_n \mathbf{W}_{\mathbb{S}} - \log(\sum\nolimits_{j\in R(\mathbf{T}_n)} exp(\mathbf{B}_j \mathbf{W}_{\mathbb{S}}))), \end{aligned} \quad (25)$$

where $\mathbf{W}_{\mathbb{S}}$ is the linear projection matrix in $\mathbb{S}$, and $R(\mathbf{T}_n) = \{i : \mathbf{T}_i \geq \mathbf{T}_n\}$ is the collection of all patients (including censored and uncensored ones) 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

$$\mathcal{L}_{\text{II}} = \mathcal{L}_{cox} + \beta_{tcl}\mathcal{L}_{tcl} + \beta_{adc}\mathcal{L}_{adc} + \beta_s\mathcal{L}_s, \quad (26)$$

where β_{tcl}, β_{adc} and β_s are the weight coefficients.

III. Experiments

In this section, we first describe the data and code availability to ensure reproducibility. Subsequently, we perform a comprehensive performance evaluation and comparative analysis against multiple state-of-the-art models. To further verify the contribution of each module, we conduct ablation experiments under consistent experimental settings. Beyond quantitative evaluation, we further explore the interpretability of DCMIL by visualizing saliency maps across multi-magnification WSIs and analyzing representative soft-bag instance embeddings. Additional implementation details and visualization results are provided in the supplementary materials of the preprint version of this paper[3].

A. Data and code availability

All 12 cancer datasets from The Cancer Genome Atlas Program (TCGA), including WSIs and their corresponding clinical data, are publicly available at https://portal.gdc.cancer.gov/. All code was implemented in Python, using PyTorch as the primary deep-learning package.

B. Performance evaluation and comparative analysis

DCMIL enables direct transformation of gigapixel-size WSIs into outcome predictions, bypassing the need for manual region selection or patch-level annotations. We evaluated DCMIL on 12 cancer types using a 5-fold cross-validation strategy across 5,954 patients and approximately 12.54 million image tiles. Prognostic performance was assessed using the concordance index (CI) and Kaplan–Meier (KM) survival analysis. DCMIL was compared against other weakly supervised learning methods—including graph-based (PGCN [23]), transformer-based (HIPT [44]), and multi-instance learning models (MesoNet [19], AMiSL [21])—as well as recent whole slide foundation models such as TITAN [45], GigaPath [7], CHIEF [24], PRISM [46], and MADELEINE [47]. As summarized in Table I1, DCMIL consistently outperformed all

[3]Preprint available at https://doi.org/10.48550/arXiv.2510.14403.

TABLE I

PERFORMANCE COMPARISON ACROSS 12 DATASETS. VALUES ARE MEAN ± STD (STD IN SMALLER FONT).

Category	Method	KIRC	LUAD	LUSC	STAD	THCA	HNSC	LIHC	BRCA	BLCA	COAD	OV	UCEC
Weakly Supervised Learning	AMISL	0.659±0.021	0.624±0.025	0.614±0.035	0.625±0.044	0.812±0.093	0.624±0.053	0.653±0.033	0.650±0.043	0.625±0.044	0.639±0.094	0.591±0.027	0.688±0.068
	PGCN	0.565±0.089	0.532±0.043	0.578±0.048	0.590±0.060	0.809±0.084	0.605±0.064	0.605±0.071	0.599±0.040	0.594±0.030	0.639±0.047	0.540±0.030	0.613±0.077
	MesoNet	0.643±0.022	0.627±0.020	0.610±0.032	0.648±0.024	0.837±0.023	0.624±0.062	0.676±0.015	0.688±0.015	0.643±0.045	0.689±0.014	0.603±0.017	0.697±0.048
	HIPT	0.646±0.034	0.593±0.037	0.591±0.064	0.650±0.016	0.864±0.079	0.599±0.068	0.621±0.046	0.652±0.043	0.611±0.014	0.630±0.059	0.594±0.020	0.671±0.048
Whole Slide Foundation	TITAN	0.667±0.032	0.592±0.026	0.505±0.008	0.635±0.022	0.732±0.076	0.593±0.025	0.680±0.023	0.631±0.014	0.628±0.022	0.713±0.026	0.588±0.014	0.707±0.031
	GigaPath	0.675±0.032	0.537±0.008	0.554±0.015	0.619±0.013	0.858±0.038	0.631±0.003	0.643±0.039	0.612±0.007	0.643±0.028	0.694±0.030	0.571±0.014	0.708±0.026
	CHIEF	0.672±0.029	0.538±0.008	0.566±0.005	0.621±0.020	0.795±0.066	0.609±0.015	0.632±0.045	0.618±0.030	0.629±0.020	0.651±0.033	0.570±0.024	0.711±0.025
	PRISM	0.625±0.032	0.545±0.014	0.558±0.032	0.591±0.009	0.804±0.067	0.575±0.016	0.671±0.020	0.617±0.027	0.634±0.024	0.600±0.026	0.598±0.018	0.648±0.033
	MADELEINE	0.657±0.023	0.575±0.029	0.545±0.017	0.633±0.018	0.687±0.053	0.613±0.016	0.676±0.018	0.519±0.009	0.628±0.035	0.702±0.014	0.575±0.012	0.703±0.027
Proposed	DCMIL	**0.676**±0.018	**0.646**±0.020	**0.659**±0.024	**0.672**±0.017	**0.871**±0.052	**0.667**±0.016	**0.681**±0.022	**0.715**±0.017	**0.669**±0.047	**0.718**±0.030	**0.616**±0.022	**0.726**±0.035

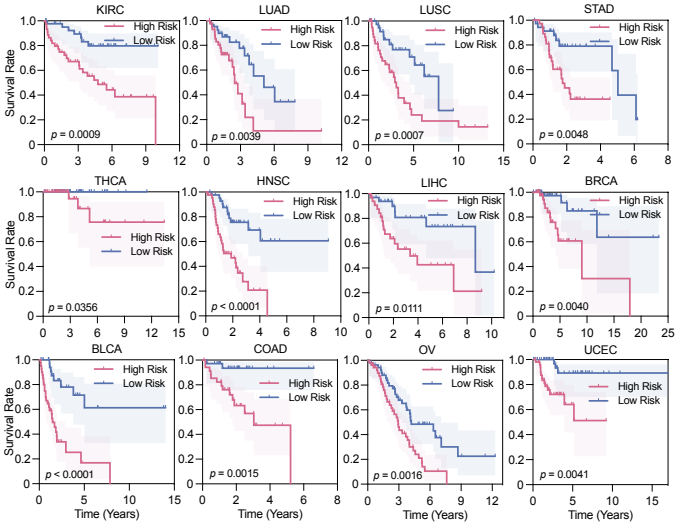

Fig. 4. The KM curves along with p-values on twelve datasets.

baseline methods, achieving the highest CI across all 12 cancer types with a mean CI of 0.693. Kaplan–Meier curves with corresponding *p*-values (Figure 4) further demonstrated its ability to significantly stratify patients into high- and low-risk groups. It may benefit from several potential advantages in DCMIL: 1) Unlike some methods that leverage pre-trained models [19], [21], [23], it encodes instances in a weakly-supervised manner to reduce label noise and maintain prognosis-related guidance. 2) It utilizes the low-magnification saliency map to guide the encoding of high-magnification instances for exploring fine-grained information across multi-magnification WSIs. 3) It introduces a soft-bag learning method and a constrained self-attention strategy to help reduce intra-bag redundancy at both the instance and feature levels. 4) It introduces a normal control and equips the Cox model with triple-tier contrastive learning to enhance both intra-bag and inter-bag discrimination.

C. Ablation experiments

We validated the efficacy of crucial components in Curriculum I and Curriculum II. From Table II, we can observe several key points: 1) the model with multi-magnification (MM) outperforms the one without MM, which benefits from the utilization of multi-magnification information and cross-scale view-aware guidance. 2) The model with triple-tier contrastive learning strategy (TCL) shows better performance, which benefits from the reduction of intra-bag redundancy and the enhancement of intra-bag and inter-bag discrimination. We validated the efficacy of crucial components in Curriculum I (C-I) and Curriculum II (C-II). From Table II, we can observe several key points: 1) the model with multi-magnification (MM) achieves superior performance by leveraging cross-scale information through view-aware guidance; 2) incorporating the saliency-guided (SG) method and hierarchical transfer (HT) strategy further enhances feature focus and cross-scale consistency; 3) the pairwise ranking loss (PR) facilitates more discriminative representation learning; 4) in C-II, the introduction of soft-bag learning (SB) and the constrained self-attention (CSA) module effectively mitigate label noise and improve intra-bag feature alignment; 5) moreover, the triple-tier contrastive learning (TCL) strategy reduces intra-bag redundancy and enhances both intra- and inter-bag discrimination; 6) finally, the absolute distance constraint (ADC) provides additional regularization to maintain a consistent and well-structured feature space.

TABLE II

THE RESULTS OF ABLATION EXPERIMENTS ON THREE DATASETS (I.E., BLCA, COAD, HNSC, LIHC, LUAD). THE BOLDFACE DENOTES THE BEST RESULT.

MM	SG	HT	PR	C-II	BLCA	COAD	HNSC	LIHC	LUAD
✗	✗	✗	✗	✓	0.653±0.024	0.684±0.036	0.647±0.035	0.649±0.046	0.632±0.016
✓	✗	✓	✓	✓	0.659±0.017	0.707±0.044	0.649±0.043	0.668±0.038	0.635±0.008
✓	✓	✗	✓	✓	0.663±0.033	0.709±0.031	0.660±0.035	0.678±0.035	0.642±0.033
✓	✓	✓	✗	✓	0.665±0.019	0.716±0.041	0.653±0.041	0.678±0.041	0.644±0.025
C-I	SB	CSA	TCL	ADC					
✓	✗	✗	✗	✗	0.654±0.033	0.691±0.033	0.647±0.016	0.659±0.036	0.634±0.016
✓	✓	✗	✓	✓	0.665±0.011	0.712±0.046	0.654±0.015	0.676±0.033	0.646±0.010
✓	✓	✓	✗	✗	0.660±0.011	0.704±0.036	0.651±0.018	0.663±0.044	0.637±0.004
✓	✓	✓	✓	✗	0.667±0.029	0.717±0.027	0.657±0.040	0.679±0.047	0.640±0.010
✓	✓	✓	✓	✓	**0.669**±0.047	**0.718**±0.030	**0.667**±0.016	**0.681**±0.022	**0.646**±0.020

Note: C-I, Curriculum I; MM, multi-magnification strategy; SG, saliency-guided method; HT, hierarchical transfer strategy; PR, pairwise ranking loss; C-II, Curriculum II; SB, soft-bag learning; CSA, constrained self-attention module; TCL, triple-tier contrastive learning; ADC, absolute distance constraint.

D. Saliency maps across multi-magnification WSIs

DCMIL can highlight fine-grained prognosis-salient regions on WSIs. Figure 5 illustrates two representative cases from the COAD dataset, including a high-risk case (Figure 5a, stage II, deceased at 14.7 months) and a low-risk case (Figure 5b, stage III, survived over 55.8 months). In Figure 5a/b, the first row shows original WSI, patch-level risk status estimation, and saliency maps of WSIs at different magnifications (i.e., 5×, 10×, and 20×). The second row exhibits derived from the low-magnification branch, aiding in learning fine-grained representations at higher magnifications. From Figure 5, we can observe that the lower-stage case exhibits a higher risk,

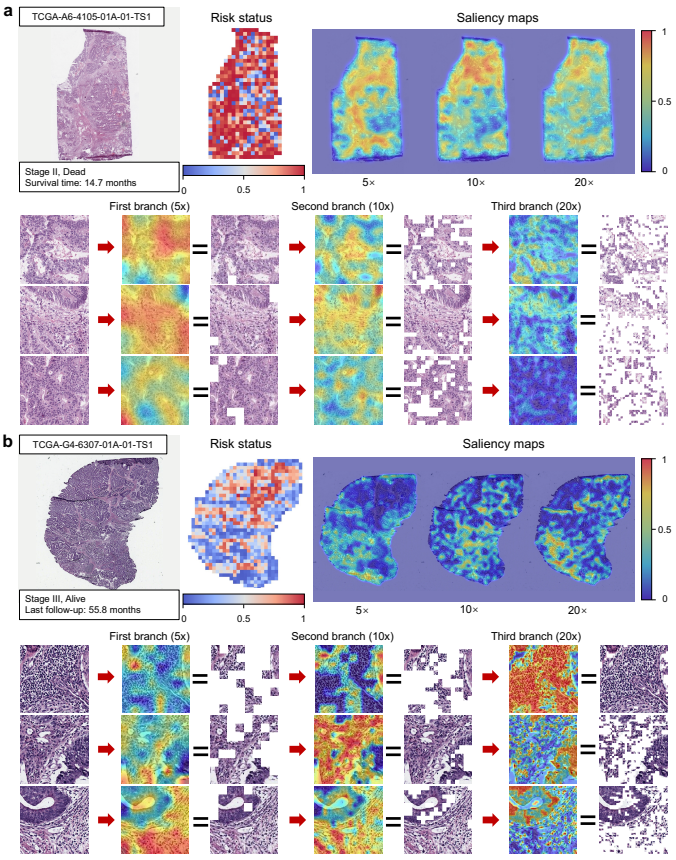

Fig. 5. Visualizations of saliency maps across multi-magnification WSIs on Curriculum I for COAD dataset.

indicating that the tumor stage is not a decisive factor for patient outcome. It also demonstrates that DCMIL can accurately predict risk stratification status, with the high-risk case exhibiting more extensive regions of high-risk probability than the low-risk case. The saliency maps of WSIs at different magnifications show that each branch highlights distinct yet complementary regions. From local tiles, we find that different branches emphasize unique spatial information: 5× tiles accentuate textural alterations in local microenvironment, 10× tiles highlight tissue and cell cluster patterns, and 20× tiles focus on nuclear morphology. This may stem from the token size in the ViT-like architecture, covering approximately 8 μm^2 (single cell) at 20× tiles, 32 μm^2 (tissue or cell cluster) at 10× tiles, and over 100 μm^2 (local microenvironment) at 5× tiles. From 5× tiles, we also notice inconsistent salient regions in homogenous tissues. This inconsistency likely arises from minimal inter-token variability within such tissues, posing a challenge for learning discriminative representations. Even so, the model appears to compensate by leveraging complementary cues from the tiles at other magnifications. Furthermore, we conjecture that two patterns occur in saliency map guidance: fine-grained representations of higher-magnification tiles are learned either from the salient regions at lower magnifications or from expanded regions that compensate for lower-magnification saliency map guidance. As exhibited in Figure 5a (second row), fine-grained representations are derived from smaller salient regions as magnification increases. As shown in Figure 5b (second row), fine-grained representations are extracted from those regions that may not be salient at lower magnifications. Actually, such phenomenon occurs as the selection of tiles is revisitable and not entirely determined by the low-magnification saliency map. This enables flexible learning of fine-grained yet complementary information, while saliency map guidance acts as a dynamic masking strategy to help alleviate overfitting.

E. Representative soft-bag instance embedding

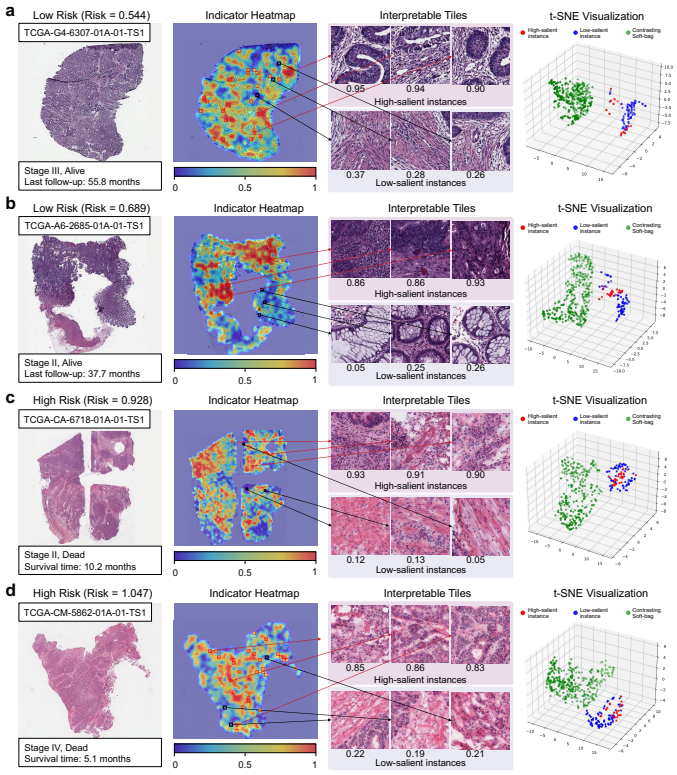

Fig. 6. Four representative cases from the COAD dataset are visualized on Curriculum II, illustrated with the original WSI, indicator heatmap, interpretable tiles, and t-SNE visualization.

Human-readable interpretability of the model can validate that its predictive basis aligns with the pathologists' knowledge, making it suitable for human-in-the-loop clinical decision-making. DCMIL model performs prognosis analysis by first identifying and aggregating representative (i.e., high-salient) regions within the WSI, while contrasting these with other WSIs. To visualize and interpret representative regions, we generate an indicator heatmap by normalizing the indicator scores and mapping them to their corresponding spatial locations in the WSI. Figure 6 presents four cases from the COAD dataset, illustrated with the original WSI, indicator heatmap, interpretable tiles, and t-SNE visualization. These cases include two low-risk patients (Figure 6a and 6b, stage III and II, survival times over 55.8 and 37.7 months) and two high-risk patients (Figure 6c and 6d, stage II and IV, decreased at 10.2 and 5.1 months). The indicator heatmap visualizes the soft-bag instance selection and its localization within the WSI, allowing us to observe that these selected representative instances are widely distributed throughout the tumor microenvironment. Interpretive tiles exhibit several high-salient and

low-salient instances, along with their corresponding indicator scores. The visual assessment of a board-certified pathologist across various cancer types indicates that the soft-bag captures fine-grained morphological embeddings, often focusing on regions with dense nuclei. The t-SNE visualization depicts the embedding distributions of high-salient instances, low-salient instances, and the contrasting soft-bags from other WSIs. We observe that the selected (high-salient) instances intermingle with the unselected (low-salient) instances within the same WSI and exhibit a consistent distribution, which suggests those high-salient instances can effectively represent the heterogeneous tumor microenvironment of the WSI. Additionally, the distribution of instances within the same WSI is significantly distinct from that of the contrasting soft-bags. Such phenomenon underscores the importance of maximizing the mutual information between selected and unselected instances by utilizing selected instances to identify unselected instances from the instance pool of soft-bag.

IV. DISCUSSION

This study shows DCMIL can quantify tumor heterogeneity, overcome computational bottlenecks by transforming gigapixel-size WSIs into risk stratification and outcome predictions without dense annotations, generalize to twelve cancer types, and locate fine-grained prognosis-salient regions.

In the first curriculum, DCMIL utilizes a low-magnification saliency map to guide the encoding of high-magnification instances, exploring fine-grained information across multi-magnification WSIs, and its quantitative performance improvement is validated in Table II. This mimics clinical practice where low-magnification cues guide doctors to significant regions at higher magnifications, supplemented by details observed at higher magnifications. We showcase saliency maps (i.e., Figure 5 on COAD) for entire WSIs and local tiles, which can be used as interpretability tools in biomarker mining to identify morphological features associated with survival outcomes or as visualization tools for secondary opinions in anatomic pathology. The saliency visualization reveals that multi-scale instance attention focuses on unique spatial survival-related information: $5\times$ tiles accentuate textural alterations in the local microenvironment, $10\times$ tiles highlight tissue and cell cluster patterns, and $20\times$ tiles focus on nuclear morphology, often forming a progressive and complementary relationship. This approach allows for flexible learning of fine-grained yet complementary information, with saliency map guidance acting as a dynamic masking strategy to reduce overfitting.

In the second curriculum, unlike general MIL-based approaches that use attention-based pooling and treat different slide locations as independent regions, DCMIL provides the model with the flexibility of selectively aggregating information from WSI, learning potential nonlinear interactions between instances, reducing intra-bag redundancy and enhancing context awareness. Additionally, it introduces a normal control and employs triple-tier contrastive learning to enhance both intra-bag and inter-bag discrimination, thereby improving the model's capability in prognosis inference. The quantitative performance improvement of this approach is validated in Table II. Indicator heatmaps and t-SNE visualizations (i.e., Figure 6 on COAD) underscore the importance of maximizing mutual information between selected and unselected instances by using selected instances to recognize unselected instances from the soft-bag instance pool. Importantly, while other models require post hoc analyses to understand their decision-making, DCMIL directly reveals which instances it considers important. This transparency is crucial as these measurements begin to be used in clinical settings and researchers increasingly turn to deep learning for analysis, ensuring models can explain their decisions in light of patients' right to understand treatment decisions.

V. CONCLUSION

In this paper, we propose Dual-Curriculum Contrastive Multi-Instance Learning (DCMIL), a novel framework designed to advance cancer prognosis analysis with WSIs. DCMIL introduces a two-stage curriculum that integrates saliency-guided weakly supervised instance encoding and contrastive-enhanced soft-bag prognosis inference, enabling efficient learning from gigapixel WSIs without dense annotations. Extensive experiments across twelve cancer types demonstrate that DCMIL achieves state-of-the-art performance in survival prediction while maintaining strong generalization capability. Moreover, DCMIL effectively quantifies tumor heterogeneity and identifies prognosis-salient regions. By jointly leveraging multi-magnification information, aliency-guided strategy, and triple-tier contrastive learning, DCMIL not only enhances predictive accuracy but also provides clinically interpretable visual evidence, bridging the gap between computational pathology and real-world clinical decision support.

REFERENCES

[1] A. Marusyk, V. Almendro, and K. Polyak, "Intra-tumour heterogeneity: a looking glass for cancer?" *Nature reviews cancer*, vol. 12, no. 5, pp. 323–334, 2012.
[2] I. Vitale, E. Shema, S. Loi *et al.*, "Intratumoral heterogeneity in cancer progression and response to immunotherapy," *Nature medicine*, vol. 27, no. 2, pp. 212–224, 2021.
[3] M. Kleppe and R. L. Levine, "Tumor heterogeneity confounds and illuminates: assessing the implications," *Nature medicine*, vol. 20, no. 4, pp. 342–344, 2014.
[4] S. A. Alowais, S. S. Alghamdi, N. Alsuhebany *et al.*, "Revolutionizing healthcare: the role of artificial intelligence in clinical practice," *BMC medical education*, vol. 23, no. 1, p. 689, 2023.
[5] M. B. Amin, F. L. Greene, S. B. Edge *et al.*, "The eighth edition ajcc cancer staging manual: continuing to build a bridge from a population-based to a more "personalized" approach to cancer staging," *CA: a cancer journal for clinicians*, vol. 67, no. 2, pp. 93–99, 2017.
[6] R. J. Chen, M. Y. Lu, D. F. Williamson *et al.*, "Pan-cancer integrative histology-genomic analysis via multimodal deep learning," *Cancer Cell*, vol. 40, no. 8, pp. 865–878, 2022.
[7] H. Xu, N. Usuyama, J. Bagga *et al.*, "A whole-slide foundation model for digital pathology from real-world data," *Nature*, pp. 1–8, 2024.
[8] A. H. Song, G. Jaume, D. F. Williamson *et al.*, "Artificial intelligence for digital and computational pathology," *Nature Reviews Bioengineering*, vol. 1, no. 12, pp. 930–949, 2023.
[9] G. Campanella *et al.*, "Clinical-grade computational pathology using weakly supervised deep learning on whole slide images," *Nature Medicine*, vol. 25, no. 8, pp. 1301–1309, 2019.
[10] M. Y. Lu, D. F. Williamson, T. Y. Chen *et al.*, "Data-efficient and weakly supervised computational pathology on whole-slide images," *Nature Biomedical Engineering*, vol. 5, no. 6, pp. 555–570, 2021.

[11] Y. Lee, J. H. Park, S. Oh *et al.*, "Derivation of prognostic contextual histopathological features from whole-slide images of tumours via graph deep learning," *Nature Biomedical Engineering*, pp. 1–15, 2022.

[12] P. Tarantino, L. Mazzarella, A. Marra *et al.*, "The evolving paradigm of biomarker actionability: histology-agnosticism as a spectrum, rather than a binary quality," *Cancer Treatment Reviews*, vol. 94, p. 102169, 2021.

[13] S.-J. Sammut, M. Crispin-Ortuzar, S.-F. Chin *et al.*, "Multi-omic machine learning predictor of breast cancer therapy response," *Nature*, vol. 601, no. 7894, pp. 623–629, 2022.

[14] J. Vamathevan, D. Clark, P. Czodrowski *et al.*, "Applications of machine learning in drug discovery and development," *Nature reviews Drug discovery*, vol. 18, no. 6, pp. 463–477, 2019.

[15] J. N. Kather, A. T. Pearson, N. Halama *et al.*, "Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer," *Nature medicine*, vol. 25, no. 7, pp. 1054–1056, 2019.

[16] A. Shmatko, N. Ghaffari Laleh, M. Gerstung *et al.*, "Artificial intelligence in histopathology: enhancing cancer research and clinical oncology," *Nature cancer*, vol. 3, no. 9, pp. 1026–1038, 2022.

[17] J. Lipkova, R. J. Chen, B. Chen *et al.*, "Artificial intelligence for multimodal data integration in oncology," *Cancer cell*, vol. 40, no. 10, pp. 1095–1110, 2022.

[18] O.-J. Skrede *et al.*, "Deep learning for prediction of colorectal cancer outcome: a discovery and validation study," *The Lancet*, vol. 395, no. 10221, pp. 350–360, 2020.

[19] P. Courtiol *et al.*, "Deep learning-based classification of mesothelioma improves prediction of patient outcome," *Nature Medicine*, vol. 25, no. 10, pp. 1519–1525, 2019.

[20] P. Mobadersany *et al.*, "Predicting cancer outcomes from histology and genomics using convolutional networks," *Proceedings of the National Academy of Sciences*, vol. 115, no. 13, pp. E2970–E2979, 2018.

[21] J. Yao, X. Zhu, J. Jonnagaddala *et al.*, "Whole slide images based cancer survival prediction using attention guided deep multiple instance learning networks," *Medical Image Analysis*, vol. 65, p. 101789, 2020.

[22] W. Shao, H. Shi, J. Liu *et al.*, "Multi-instance multi-task learning for joint clinical outcome and genomic profile predictions from the histopathological images," *IEEE Transactions on Medical Imaging*, 2024.

[23] R. J. Chen *et al.*, "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*, 2021, pp. 339–349.

[24] X. Wang, J. Zhao, E. Marostica *et al.*, "A pathology foundation model for cancer diagnosis and prognosis prediction," *Nature*, vol. 634, no. 8035, pp. 970–978, 2024.

[25] C. L. Srinidhi, S. W. Kim, F.-D. Chen *et al.*, "Self-supervised driven consistency training for annotation efficient histopathology image analysis," *Medical Image Analysis*, vol. 75, p. 102256, 2022.

[26] T. Vu, P. Lai, R. Raich *et al.*, "A novel attribute-based symmetric multiple instance learning for histopathological image analysis," *IEEE Transactions on Medical Imaging*, vol. 39, no. 10, pp. 3125–3136, 2020.

[27] H. Zhang *et al.*, "Dtfd-mil: double-tier feature distillation multiple instance learning for histopathology whole slide image classification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18 802–18 812.

[28] N. Naik *et al.*, "Deep learning-enabled breast cancer hormonal receptor status determination from base-level h&e stains," *Nature Communications*, vol. 11, no. 1, pp. 1–8, 2020.

[29] S. Wang *et al.*, "Rmdl: recalibrated multi-instance deep learning for whole slide gastric image classification," *Medical Image Analysis*, vol. 58, p. 101549, 2019.

[30] W. Shao, T. Wang, Z. Huang *et al.*, "Weakly supervised deep ordinal cox model for survival prediction from whole-slide pathological images," *IEEE Transactions on Medical Imaging*, vol. 40, no. 12, pp. 3739–3747, 2021.

[31] N. Coudray, P. S. Ocampo, T. Sakellaropoulos *et al.*, "Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning," *Nature medicine*, vol. 24, no. 10, pp. 1559–1567, 2018.

[32] C. Tu, Y. Zhang, and Z. Ning, "Dual-curriculum contrastive multi-instance learning for cancer prognosis analysis with whole slide images," *Advances in Neural Information Processing Systems*, vol. 35, pp. 29 484–29 497, 2022.

[33] M. Y. Lu, B. Chen, D. F. Williamson *et al.*, "A visual-language foundation model for computational pathology," *Nature Medicine*, vol. 30, no. 3, pp. 863–874, 2024.

[34] Z. Huang, F. Bianchi, M. Yuksekgonul *et al.*, "A visual–language foundation model for pathology image analysis using medical twitter," *Nature medicine*, vol. 29, no. 9, pp. 2307–2316, 2023.

[35] D. Tellez, G. Litjens, J. van der Laak *et al.*, "Neural image compression for gigapixel histopathology image analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 2, pp. 567–578, 2019.

[36] A. Dosovitskiy *et al.*, "An image is worth 16x16 words: transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[37] H. Tokunaga, Y. Teramoto, A. Yoshizawa *et al.*, "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*, 2019, pp. 12 597–12 606.

[38] B. Zhou, A. Khosla, A. Lapedriza *et al.*, "Learning deep features for discriminative localization," in *in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2921–2929.

[39] A. Van Den Oord, O. Vinyals *et al.*, "Neural discrete representation learning," *Advances in neural information processing systems*, vol. 30, 2017.

[40] A. Vaswani *et al.*, "Attention is all you need," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[41] A. Van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[42] T. Chen, S. Kornblith, M. Norouzi *et al.*, "A simple framework for contrastive learning of visual representations," in *International Conference on Machine Learning*, 2020, pp. 1597–1607.

[43] J. Fox and S. Weisberg, "Cox proportional-hazards regression for survival data," *An R and S-PLUS Companion to Applied Regression*, vol. 2002, 2002.

[44] R. J. Chen *et al.*, "Scaling vision transformers to gigapixel images via hierarchical self-supervised learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 16 144–16 155.

[45] T. Ding, S. J. Wagner, A. H. Song *et al.*, "Multimodal whole slide foundation model for pathology," 2024. [Online]. Available: https://arxiv.org/abs/2411.19666

[46] G. Shaikovski, A. Casson, K. Severson *et al.*, "Prism: A multimodal generative foundation model for slide-level histopathology," *arXiv preprint arXiv:2405.10254*, 2024.

[47] G. Jaume, A. Vaidya, A. Zhang *et al.*, "Multistain pretraining for slide representation learning in pathology," in *European Conference on Computer Vision*. Springer, 2024, pp. 19–37.

Supplementary Materials for "DCMIL: A Progressive Representation Learning of Whole Slide Images for Cancer Prognosis Analysis"

A. Main notations used in the work

In this part, we list the main notations in Table S1 for clear reference

B. Pseudocode of the Proposed Method

The detailed procedure of the proposed method is summarized in **Algorithm 1**.

C. Implementation Details.

We implemented all competing methods using the Pytorch2.0 library on Python3.9. All intensive calculations were offloaded to a 32 GB NVIDIA Tesla V100 GPU. For Curriculum I, we adopted Adam optimizer for network training. The learning rate and batch size were set to 10^{-5} and 12, respectively. The weight coefficients β_Ω and β_R were set to 10^{-5} and 1, respectively.

To better guide fine-grained representation learning with coarse-grained saliency map, we first pre-trained the *1*-st (i.e., $5\times$) branch for 200 epochs. Subsequently, its parameters were shared with the *2*-nd (i.e., $10\times$) and *3*-rd (i.e., $20\times$) branches, and all of them were trained together for 50 epochs with the early stopping rate of 5. We selected a three-year survival time as the threshold for six datasets (BLCA, COAD, HNSC, LIHC, OV, and STAD) and a five-year survival time for the remaining datasets (BRCA, KIRC, LUAD, LUSC, THCA, and UCEC). The reasons for these choices are twofold: 1) Three-year and five-year survival times are significant clinical prognosis assessment timepoints. 2) Selecting three-year or five-year survival time as the threshold ensures a relatively balanced distribution of positive and negative samples, which benefits effective model training.

For Curriculum II, we adopted SGD optimizer with the momentum of 0.9 for network training. The learning rate, batch size, and epoch number were set to 10^{-5}, 64 and 1000, respectively. The soft-bag size N_B was set to 30 and the weight coefficients β_{tcl}, β_{adc}, and β_s were set to 1, 0.1, and 10^{-4}, respectively. Additionally, we introduced a self-paced learning strategy [1] to train Curriculum I, employing a learnable indicator DCMIL to select instances from easy to hard for network training. This process is essentially a biconvex optimization problem, for which alternative convex search (ACS) [2] offers a feasible solution. Specifically, with the network parameters fixed, those instances with the loss smaller than a certain threshold λ were considered easy instances and selected for training, while others were excluded. The parameter λ controlled the pace at which the network learned new instances. When λ is small, only easy instances with small losses are considered. As λ increases, more instances with larger losses are gradually included, allowing the network to mature.

D. The key advancements and contributions of this work.

The key advancements and contributions of this work, compared to our previous NeurIPS paper [**?**], include several significant improvements. First, the backbone network has been upgraded with a ViT-like network architecture [**?**], enabling the incorporation of long-range dependencies and enhancing the model's ability to understand complex data patterns. Second, saliency guidance has been refined across multi-magnification WSIs to facilitate the exploration of fine-grained information, thereby providing a more detailed analysis. Third, the introduction of normal control and the integration of the Cox model with triple-tier contrastive learning add robustness to the analysis framework. Additionally, we have provisioned fine-grained prognosis-salient regions on WSIs, allowing for more precise prognosis predictions. Robust instance uncertainty estimation has also been implemented to increase the reliability of the model's outputs. Furthermore, we capture morphological differences between normal and tumor tissues, potentially generating new biological insights. The validation process has been enhanced through the utilization of more datasets, ensuring the model's robustness and generalizability. Finally, we have significantly improved the model's overall speed, precision, and interpretability, making it more efficient and effective for cancer prognosis analysis.

E. Estimating the Mutual Information with the Triple Loss

As already shown in Section Curriculum II, the optimal value for $\mathbb{L}(\mathbf{B}_n, \widetilde{\mathbf{B}}_n, \widehat{\mathbf{B}}_n, \overline{\mathbf{B}}_n)$ is given by

$$\mathbb{L}(\mathbf{B}_n, \widetilde{\mathbf{B}}_n, \widehat{\mathbf{B}}_n, \overline{\mathbf{B}}_n) = \frac{p(\mathbf{B}_n|\widetilde{\mathbf{B}}_n)p(\mathbf{B}_n|\widehat{\mathbf{B}}_n)p(\mathbf{B}_n|\overline{\mathbf{B}}_n)}{p(\mathbf{B}_n)p(\mathbf{B}_n)p(\mathbf{B}_n)}. \tag{1}$$

Inserting this back in to $\mathcal{L}_{tcl}$ and splitting $\mathcal{B}$ into the positive bag and negative bags results in:

$$\mathcal{L}_{tcl} = -\underset{\mathcal{B}}{\boldsymbol{E}}\left[\log \frac{\mathbb{L}(\mathbf{B}_n, \widetilde{\mathbf{B}}_n, \widehat{\mathbf{B}}_n, \overline{\mathbf{B}}_n)}{\mathbb{L}(\mathbf{B}_n, \widetilde{\mathbf{B}}_n, \widehat{\mathbf{B}}_n, \overline{\mathbf{B}}_n) + \sum_{i\in[N], i\neq n}\mathbb{L}(\mathbf{B}_i, \underline{\mathbf{B}}, \widehat{\mathbf{B}}_n, \overline{\mathbf{B}}_n)}\right] \tag{2}$$

$$= \underset{\mathcal{B}}{\boldsymbol{E}}\log\left[1 + \frac{1}{\mathbb{L}(\mathbf{B}_n, \widetilde{\mathbf{B}}_n, \widehat{\mathbf{B}}_n, \overline{\mathbf{B}}_n)}\sum_{i\in[N], i\neq n}\mathbb{L}(\mathbf{B}_i, \underline{\mathbf{B}}, \widehat{\mathbf{B}}_n, \overline{\mathbf{B}}_n)\right] \tag{3}$$

$$\approx \underset{\mathcal{B}}{\boldsymbol{E}}\log\left[1 + \frac{1}{\mathbb{L}(\mathbf{B}_n, \widetilde{\mathbf{B}}_n, \widehat{\mathbf{B}}_n, \overline{\mathbf{B}}_n)}(N-1)\boldsymbol{E}_{i\in[N], i\neq n}\mathbb{L}(\mathbf{B}_i, \underline{\mathbf{B}}, \widehat{\mathbf{B}}_n, \overline{\mathbf{B}}_n)\right] \tag{4}$$

$$= \underset{\mathcal{B}}{\boldsymbol{E}}\log\left[1 + \frac{(N-1)p(\mathbf{B}_n)^3}{p(\mathbf{B}_n|\widetilde{\mathbf{B}}_n)p(\mathbf{B}_n|\widehat{\mathbf{B}}_n)p(\mathbf{B}_n|\overline{\mathbf{B}}_n)}\boldsymbol{E}_{i\in[N], i\neq n}\frac{p(\mathbf{B}_i|\underline{\mathbf{B}})p(\mathbf{B}_i|\widehat{\mathbf{B}}_n)p(\mathbf{B}_i|\overline{\mathbf{B}}_n)}{p(\mathbf{B}_i)^3}\right] \tag{5}$$

$$= \underset{\mathcal{B}}{\boldsymbol{E}}\log\left[1 + \frac{(N-1)p(\mathbf{B}_n)^3}{p(\mathbf{B}_n|\widetilde{\mathbf{B}}_n)p(\mathbf{B}_n|\widehat{\mathbf{B}}_n)p(\mathbf{B}_n|\overline{\mathbf{B}}_n)}\right] \tag{6}$$

$$\geq \underset{\mathcal{B}}{\boldsymbol{E}}\log\left[N\frac{p(\mathbf{B}_n)^3}{p(\mathbf{B}_n|\widetilde{\mathbf{B}}_n)p(\mathbf{B}_n|\widehat{\mathbf{B}}_n)p(\mathbf{B}_n|\overline{\mathbf{B}}_n)}\right] \tag{7}$$

$$= -\underset{\mathcal{B}}{\boldsymbol{E}}\log\frac{p(\mathbf{B}_n|\widetilde{\mathbf{B}}_n)}{p(\mathbf{B}_n)} - \underset{\mathcal{B}}{\boldsymbol{E}}\log\frac{p(\mathbf{B}_n|\overline{\mathbf{B}}_n)}{p(\mathbf{B}_n)} - \underset{\mathcal{B}}{\boldsymbol{E}}\log\frac{p(\mathbf{B}_n|\widehat{\mathbf{B}}_n)}{p(\mathbf{B}_n)} + \log N \tag{8}$$

$$= -\mathcal{I}(\mathbf{B}_n, \widetilde{\mathbf{B}}_n) - \mathcal{I}(\mathbf{B}_n, \overline{\mathbf{B}}_n) - \mathcal{I}(\mathbf{B}_n, \widehat{\mathbf{B}}_n) + \log N \tag{9}$$

Therefore, $\mathcal{I}(\mathbf{B}_n, \widetilde{\mathbf{B}}_n) + \mathcal{I}(\mathbf{B}_n, \overline{\mathbf{B}}_n) + \mathcal{I}(\mathbf{B}_n, \widehat{\mathbf{B}}_n) \geq -\mathcal{L}_{tcl} + \log N$. This relationship also holds for other functions f that result in a higher loss value $\mathcal{L}_{tcl}$. Equation 4 becomes more accurate as N increases. Meanwhile, the value of $log(N) - \mathcal{L}_{tcl}$ also increases with larger N. Therefore, it is beneficial to use large values of N.

F. Robust instance uncertainty estimation towards risk stratification.

We performed instance uncertainty estimation towards risk stratification by applying a dropout operator to the network, enabling multiple stochastic forward passes (set to 30) during the prediction stage. The standard deviation of network output from multiple stochastic forward passes provides an intuitive indication of the uncertainty. Given the distribution of instance uncertainty, we determined the uncertainty threshold which optimally separates correct and incorrect predictions by maximizing the Youden index [3], [4]. Figure S1 shows two representative cases from the COAD dataset, including a high-risk case (Figure S1a-S1g, stage IV, deceased at 14.1 months) and a low-risk case (Figure S1h-S1n, stage III, survived over 47.3 months). Figure S1a and S1h show the original WSIs, and Figure S1b and S1i present the patch-level risk prediction maps. The instance uncertainty maps are exhibited in Figure S1c and S1j. We applied the uncertainty threshold to filter out those regions with high uncertainty, masking them on the patch-level risk prediction maps to retain the instances with only high-confidence predictions. As illustrated in Figure S1d and S1k, nearly all instances with high-confidence predictions correspond to those with correct predictions. In other words, for low-risk cases, high-confidence predictions cluster in low-risk regions, while for high-risk cases, they cluster in high-risk regions. Figures S1e and S1l display the probability density distributions of correct (red curve) and incorrect (blue curve) predictions, along with Youden index curve (purple, dotted). It is evident that the uncertainty threshold (black dotted line) effectively separates correct from incorrect predictions. Figures S1f and S1m illustrate the distribution of instances with varying levels of uncertainty. Intuitively, there are fewer high-confidence incorrect predictions (red fork) compared to high-confidence correct predictions (red dot), whereas low-confidence incorrect predictions (blue fork) constitute a large proportion of low-confidence predictions. Additionally, Figures S1g and S1n illustrate the distribution of uncertainty concerning correct and incorrect predictions, which further substantiates that correct predictions are associated with lower uncertainties and vice versa. These results suggest that the model has the capability to perform robust instance uncertainty estimation for risk stratification, thereby laying the foundation for Curriculum II.

G. Comparative analysis between tumor and normal tissues.

Tumor-adjacent normal tissue/microenvironment may convey significant prognostic information and reveal certain prognostic indicators [5], including pH levels, stromal behavior, and transcriptomic and epigenetic aberrations. Comparative analysis against

normal tissue/microenvironment facilitates the identification of key morphological malignancy hallmarks, expediting definitive prognosis evaluation [6]. To support these advances, we introduced normal tissue slides as control samples for contrastive learning to enhance intra-bag discrimination, as illustrated in Figure S2. Figure S2a presents t-SNE visualization showing the distribution of tumor and normal tissues before and after training. We observe that the distribution of normal tissues overlaps with that of tumor tissues before training, and the well-trained DCMIL can effectively distinguish between normal and tumor samples. For further analysis, we randomly selected a normal tissue as a reference, as shown in Figure S2b. Then, we utilized the Euclidean distance to quantify and analyze the similarity/difference between instance-level representations of a normal sample, a low-risk tumor sample, and a high-risk tumor sample compared to the reference one. The first column of Figure S2c-e exhibits the original WSIs of these samples, the second column illustrates the distance heatmaps along with two representative tiles, and the third column shows the histogram statistics of the distance heatmaps. The distance heatmaps reveal that the low-risk sample is more dissimilar to the reference sample than the normal sample, but less dissimilar than the high-risk sample. The histogram statistics show that the distance (or dissimilarity) distribution for the normal sample predominantly falls within 10 units. In contrast, for the low-risk sample, the distance is distributed across the entire range, decreasing with increasing distance. For the high-risk sample, the distances primarily cluster within the ranges of 15-20 units and 5-10 units. Therefore, using normal samples as a reference allows DCMIL to effectively differentiate normal and tumor tissues, and also help distinguish samples with different risk levels.

H. Additional Visualization Results

To further demonstrate the generalization and interpretability of the proposed model, we present additional visualization results on the LUAD (Figure S3–S6) and HNSC (Figure S7–S10) datasets.

I. Limitations of the study

Our method has several limitations that future studies may address. While the model provides interpretable visualizations, it cannot always explain the underlying reasons for discovered features, which require further quantification and human knowledge. For example, incorporating genomic data may provide molecular biology understanding of the characterization and visualization of histopathological images. Although morphological biomarkers from WSIs can potentially predict patient outcomes, cancer prognosis is inherently a multi-modality problem driven by markers in pathophysiology, genomics, and transcriptomics. Therefore, integrating multimodal data to develop joint image-omic prognostic models can leverage complementary information from different modalities, leading to more accurate and robust cancer prognosis estimations and comprehensive assessments of cancer progression and patient outcome.

Additionally, while DCMIL can model the dependency among local tiles and capture fine-grained information, their ability to recognize visual concepts at the region level (e.g., mitosis detection, cell counting, and nuclei quantification) remains unexplored. Another challenge for future research is to develop data-efficient methods to handle noisy labels, poorly differentiated cases, mixed cancer subtypes, and extremely limited labeled data, as well as incorporating human-in-the-loop decision-making. Overall, DCMIL demonstrates the potential and effectiveness of the multi-instance curriculum learning paradigm in WSI-based cancer prognosis analysis. This approach lays the foundation for future studies to harness the potential of larger datasets at scale, contributing to more accurate patient outcome prediction and driving innovation in computational pathology.

REFERENCES

[1] M. Kumar, B. Packer, and D. Koller, "Self-paced learning for latent variable models," *Advances in neural information processing systems*, vol. 23, 2010.
[2] M. S. Bazaraa, H. D. Sherali, and C. M. Shetty, *Nonlinear programming: theory and algorithms*. John wiley & sons, 2006.
[3] R. Fluss, D. Faraggi, and B. Reiser, "Estimation of the youden index and its associated cutoff point," *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, vol. 47, no. 4, pp. 458–472, 2005.
[4] J. M. Dolezal, A. Srisuwananukorn, D. Karpeyev *et al.*, "Uncertainty-informed deep learning models enable high-confidence predictions for digital histopathology," *Nature communications*, vol. 13, no. 1, p. 6572, 2022.
[5] D. Aran, R. Camarda, J. Odegaard *et al.*, "Comprehensive analysis of normal adjacent to tumor transcriptomes," *Nature communications*, vol. 8, no. 1, p. 1077, 2017.
[6] E. Oh and H. Lee, "Transcriptomic data in tumor-adjacent normal tissues harbor prognostic information on multiple cancer types," *Cancer Medicine*, vol. 12, no. 10, pp. 11 960–11 970, 2023.

Algorithm 1 Pseudocode of the Proposed Method.

Input: Dataset $\{\mathbf{X}_n, \mathbf{T}_n, \boldsymbol{\delta}_n\}_{n=1}^{N}$, risk stratification status $\mathbf{Y}_n = \{\mathbf{y}_{n,i}^s\}_{i=1}^{N_n}$, and the weight coefficients $\beta_\Omega, \beta_R, \beta_s, \beta_{tcl}, \beta_{adc}$.
Output: Prognosis inference $\widehat{\mathbf{T}}_n \leftarrow \mathbb{S}(\mathbb{A}\{\mathbb{E}(\mathbf{x}_{n,i} : \mathbf{x}_{n,i} \in \mathbf{X}_n\})$, where $\{\mathbb{P}^s\}_{s=1}^{S}$, $\{\mathbb{F}^s\}_{s=1}^{S}$, and $\{\mathbb{G}^s\}_{s=1}^{S}$ form $\mathbb{E}$ while $\mathbb{B}$ and $\mathbb{D}$ constitute $\mathbb{A}$.
Curriculum I (C-I): Saliency-guided weakly-supervised instance encoding with cross-scale tiles.

$s \leftarrow 1$
while $s \leq S$ **do**
  for $[n, i] = [1, 1] \rightarrow [N, N_n]$ **do**
    if $s == 1$ **then**
      $\widehat{\mathbf{x}}_{n,i}^s \leftarrow \mathbf{x}_{n,i}^s$
    else
      $\alpha_{n,i}^{s-1} \leftarrow \frac{\partial \mathbf{p}_{n,i}^{s-1}}{\partial \mathbf{z}_{n,i}^{s-1}}$ ▷ Obtain token gradient through backward propagation
      $\mathbf{m}_{n,i}^{s-1} \leftarrow \mathbb{R}(\alpha_{n,i}^s \mathbf{z}_{n,i}^s) \geq \iota$ ▷ Generate salient mask
      $\widehat{\mathbf{x}}_{n,i}^s \leftarrow \mathbf{m}_{n,i}^{s-1} \otimes \mathbf{x}_{n,i}^s$ ▷ Utilize salient regions to highlight the input
    $\mathbf{z}_{n,i}^s \leftarrow \mathbb{F}^s(\mathbb{P}^s(\widehat{\mathbf{x}}_{n,i}^s))$ ▷ Generate instance token embedding
    $\mathbf{g}_{n,i}^s \leftarrow \mathbb{G}^s(\mathbf{z}_{n,i}^s, \mathbf{g}_{n,i}^{s-1})$ ▷ Obtain instance representation
    $\mathbf{p}_{n,i}^s \leftarrow \mathbb{C}^s(\mathbf{g}_{n,i}^s)$ ▷ Predict risk stratification status
    $\mathcal{L}_\ell \leftarrow$ the empirical loss calculated with $\{\mathbf{p}_{n,i}^s, \mathbf{y}_{n,i}^s\}$
    $\mathcal{L}_R \leftarrow$ the pairwise ranking loss calculated with $\{\mathbf{p}_{n,i}^{s-1}, \mathbf{p}_{n,i}^s, \mathbf{y}_{n,i}^s\}$
  $\mathcal{L}_\Omega \leftarrow$ the structural loss calculated with $\{\boldsymbol{\theta}_{\mathbb{F}^s}^i, \boldsymbol{\theta}_{\mathbb{F}^{s-1}}^i\}_{i=1}^{s-1}$
  $\mathcal{L}_{\text{I}} = \mathcal{L}_\ell + \beta_\Omega \mathcal{L}_\Omega + \beta_R \mathcal{L}_R$ ▷ Hybrid loss of Curriculum I
  Update $\{\mathbb{P}^s, \mathbb{F}^s, \mathbb{G}^s, \mathbb{C}^s\}$ by gradient descent
  $s \leftarrow s + 1$
$\mathbf{g}_{n,i} \leftarrow \mathbf{g}_{n,i}^S$. ▷ Obtain multi-scale instance representation

Curriculum II (C-II): Contrastive-enhanced Soft-bag Prognosis Inference.

for $n = 1 \rightarrow N$ **do**
  $\mathbf{G}_n \leftarrow [\mathbf{g}_{n,1}; \mathbf{g}_{n,2}; \cdots; \mathbf{g}_{n,N_n}]^T$ ▷ Initialize bag representation
  $\mathbf{E}_n \leftarrow$ the new bag representation by projecting $\mathbf{G}_n$ via a linear layer
  $\widehat{\mathbf{h}}_n \leftarrow \arg\max_{\mathbf{h}_n} \mathbb{S}(\mathbb{D}(\mathbb{H}(\mathbf{E}_n, \mathbf{h}_n)))$ ▷ Generate indicator vector
  $\widehat{\mathbf{E}}_n \leftarrow \mathbb{H}(\mathbf{E}_n, \widehat{\mathbf{h}}_n)$ ▷ Adaptively select representative instances within a bag
  $\mathbf{B}_n \leftarrow \mathbb{D}(\widehat{\mathbf{E}}_n)$ ▷ Obtain sparse soft-bag representation
  $\widehat{\mathbf{T}}_n \leftarrow \mathbb{S}(\mathbf{B}_n)$ ▷ Prognosis inference
  $\underline{\mathbf{B}}$, $\widetilde{\mathbf{B}}_n \leftarrow \mathbb{K}_1(\mathcal{B}_{\underline{n}}^{\mathcal{O}}), \mathbb{K}_1(\mathcal{B}_n^{\mathcal{O}})$ ▷ Merge the representations of normal tissue and tumor tissue through $\mathbb{K}_1$
  $\widehat{\overline{\mathbf{B}}}_n \leftarrow \mathbb{K}_1(\mathcal{B}_n^{\mathcal{O}^+} | \mathcal{B}_n^{\mathcal{O}^-})$ ▷ Merge the bag representations (expect $\mathbf{B}_n$) from the same source with $\mathbf{B}_n$
  $\overline{\mathbf{B}}_n \leftarrow \mathbb{K}_2(\mathbb{H}(\mathbf{E}_n, \mathbf{1} - \widehat{\mathbf{h}}_n))$ ▷ Merge the representation of the discarded instances
  $\mathcal{L}_{tcl} \leftarrow$ the triple-tier contrastive learning loss calculated with $\{\mathbf{B}_{\underline{n}}, \underline{\mathbf{B}}, \widetilde{\mathbf{B}}_n, \widehat{\mathbf{B}}_n, \overline{\mathbf{B}}_n\}$
  $\mathcal{L}_{adc} \leftarrow$ the absolute distance constraint calculated with $\{\mathbf{B}_n, \underline{\mathbf{B}}, \widetilde{\mathbf{B}}_n, \widehat{\mathbf{B}}_n, \overline{\mathbf{B}}_n\}$
  $\mathcal{L}_{cox} \leftarrow$ the Cox loss calculated with $\{\widehat{\mathbf{T}}_n, \mathbf{T}_n, \boldsymbol{\delta}_n\}$
$\mathcal{L}_s \leftarrow$ the sparseness loss calculated with $\{\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V\}$
$\mathcal{L}_{\text{II}} \leftarrow \mathcal{L}_{cox} + \beta_{tcl}\mathcal{L}_{tcl} + \beta_{adc}\mathcal{L}_{adc} + \beta_s \mathcal{L}_s$ ▷ Hybrid loss of Curriculum II
Update $\{\mathbb{B}, \mathbb{D}, \mathbb{S}\}$ by gradient descent.

TABLE S1
MAIN NOTATIONS USED IN THE WORK.

	Symbol	Description
Indices	S	Number of magnifications (branches) ($s \in \{1, \ldots, S\}$)
	N	Number of patients ($n \in \{1, \ldots, N\}$)
	N_n	Number of instances in the n-th bag
	N_B	Number of selected instances in each bag
	$C, D, D_B, \widehat{D}$	Feature dimension of $\mathbf{G}_n$, $\mathbf{E}_n$, $\mathbf{B}_n$, or $\mathbf{Q}_n/\mathbf{K}_n/\mathbf{V}_n$
Input	$\mathbf{X}_n$	The n-th bag (patient)
	$\mathbf{x}_{n,i}$	The i-th instance of the n-th bag
	$\mathbf{T}_n$	Observed survival time of the n-th patient
	$\mathbf{T}_r$	Time threshold
	$\boldsymbol{\delta}_n$	Event indicator of the n-th patient
	$\mathbf{Y}_n$	Risk stratification status of the n-th patient
	$\beta_\Omega, \beta_R, \beta_s, \beta_{tcl}, \beta_{adc}$	Weight coefficients in Curriculum I and II
Output	$\mathbf{p}_{n,i}^s$	Predicted probability of the i-th instance of the n-th bag in the s-th branch
	$\widehat{\mathbf{T}}_n$	Estimated survival time of the n-th patient
	$\mathcal{L}_{\text{I}}, \mathcal{L}_{\text{II}}$	Loss functions for Curriculum I and II
	$\mathcal{L}_\ell$	Empirical loss
	$\mathcal{L}_\Omega$	Structural loss
	$\mathcal{L}_R$	Pairwise ranking loss
	$\mathcal{L}_{cox}$	Cox proportional hazard regression loss
	$\mathcal{L}_{tcl}$	Triple-tier contrastive learning loss
	$\mathcal{L}_{adc}$	Absolute distance constraint
	$\mathcal{L}_s$	Sparseness loss
Feature map	$\alpha_{n,i}^s$	Gradient of the score for $\mathbf{p}_{n,i}^s$
	$\mathbf{m}_{n,i}^s$	Salient mask of the i-th instance of the n-th bag in the s-th branch
	$\widehat{\mathbf{x}}_{n,i}^s$	Highlighted input of the i-th instance of the n-th bag in the s-th branch
	$\mathbf{z}_{n,i}$	Instance token embedding
	$\mathbf{g}_{n,i}$	Multi-scale instance representation of the i-th instance of the n-th bag
	$\widehat{\mathbf{h}}_n$	Indicator that adaptively selects representative instances of the n-th bag
	$\mathbf{G}_n$	A set of instance representations of the n-th bag
	$\mathbf{B}_n$	Sparse soft-bag representation of the n-th bag
	$\mathcal{B}_n^{\mathcal{O}^+}/\mathcal{B}_n^{\mathcal{O}^-}$	Collection of bags from $\mathcal{O}^+$ (or $\mathcal{O}^-$) with the same source as $\mathbf{B}_n$
	$\mathcal{B}$	Collection of all bags
	$\mathcal{B}^{\underline{\mathcal{O}}}, \mathcal{B}_n^{\mathcal{O}}$	Collection of bags (expect $\mathbf{B}_n$) from the source $\underline{\mathcal{O}}$ or $\mathcal{O}$
	$\overline{\mathbf{B}}_n$	Integrated representation of the discarded instances
	$\widehat{\mathbf{B}}_n$	Integrated representation of $B_n^{\mathcal{O}^+}$ (or $B_n^{\mathcal{O}^-}$)
	$\underline{\mathbf{B}}, \widetilde{\mathbf{B}}_n$	Integrated representations of $\mathcal{B}^{\underline{\mathcal{O}}}$ or $\mathcal{B}_n^{\mathcal{O}}$
	$\mathbf{E}_n, \widehat{\mathbf{E}}_n, \overline{\mathbf{E}}_n$	Representation of the (all, selected, or discarded) instances of the n-th bag
	$\mathbf{Q}_n, \mathbf{K}_n, \mathbf{V}_n$	Feature spaces in $\mathbb{D}$
Network components	$\mathbb{A}$	Instance aggregation function
	$\mathbb{B}$	Soft-bag learning module
	$\mathbb{C}$	Risk stratification function (i.e, multilayer perceptron (MLP) heads)
	$\mathbb{D}$	Constrained self-attention module
	$\mathbb{E}$	Instance encoding function
	$\mathbb{F}$	Transformer extractor
	$\mathbb{G}$	Attention aggregator
	$\mathbb{H}$	Indicator function
	$\mathbb{K}_1$, $\mathbb{K}_2$	Bag or instance aggregation networks
	$\mathbb{L}$	Log-bilinear function
	$\mathbb{P}$	Linear layer
	$\mathbb{R}$	"ReLU" activation function
	$\mathbb{S}$	Prognosis inference function
Others	ϕ	"Softmax" activation function
	$\boldsymbol{\theta}_{\mathbb{F}^s}^i$	Parameter of the i-th module in $\mathbb{F}^s$
	$\mathbf{W}_n, \mathbf{W}_Q/\mathbf{W}_K/\mathbf{W}_V, \mathbf{W}_{\mathbb{L}}/\mathbf{V}_{\mathbb{L}}, \mathbf{W}_{\mathbb{S}}$	Weight matrices in $\mathbb{H}, \mathbb{D}, \mathbb{L}, \mathbb{S}$
	$\mathcal{O} = \{\mathcal{O}^+, \mathcal{O}^-\}$	Distribution of (high-risk or low-risk) tumor tissue slides
	$\underline{\mathcal{O}}$	Distribution of normal tissue slides
	ι,η,κ	Hyperparameters in Curriculum I and II

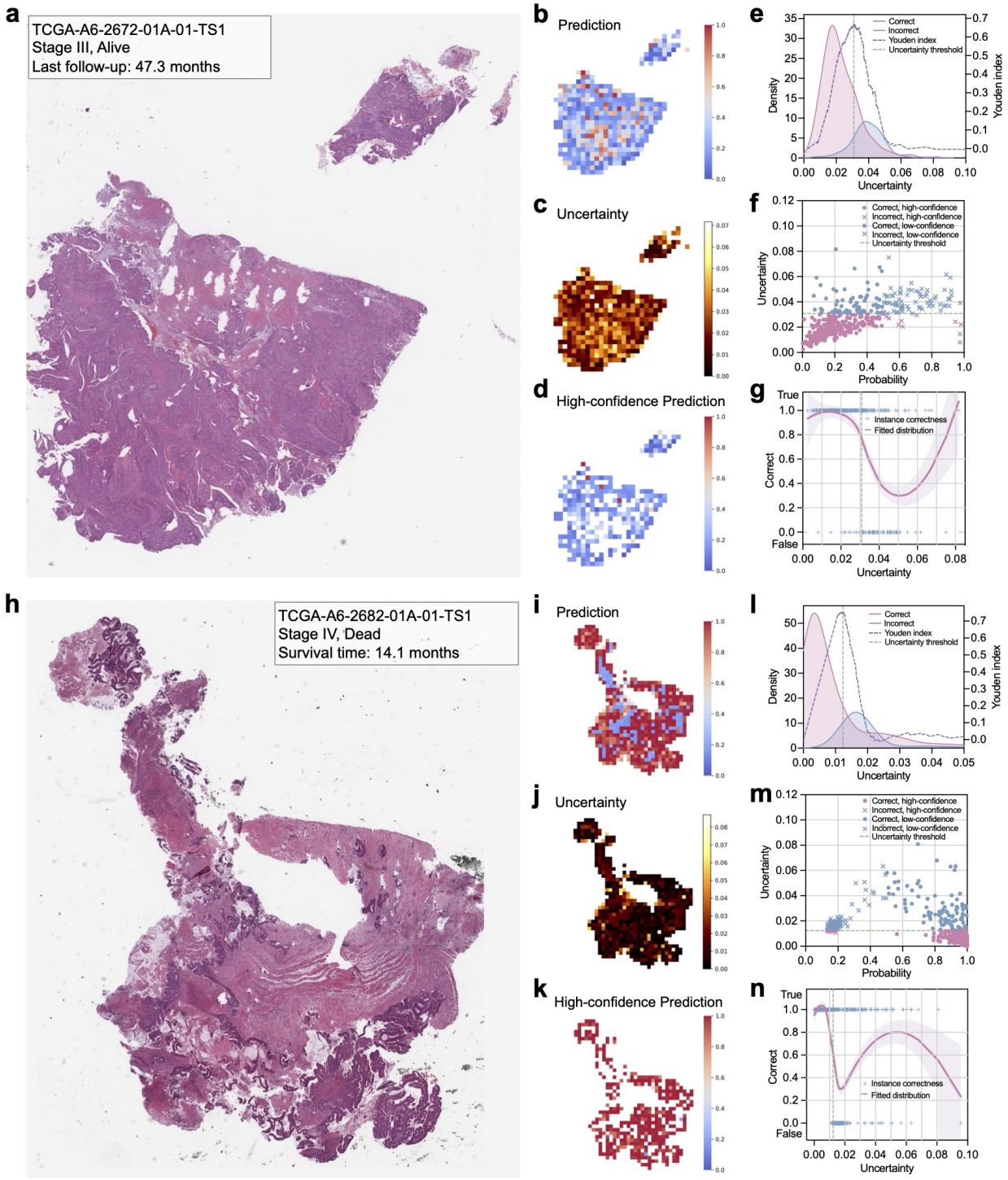

Fig. S1. **Instance uncertainty estimation towards risk stratification on Curriculum I for COAD dataset.** This includes a high-risk case (stage IV, deceased at 14.1 months) and a low-risk case (stage III, survived over 47.3 months). **a** and **h** show the original WSIs. **b** and **i** present the patch-level risk prediction maps. **c** and **j** exhibit the instance uncertainty maps. **d** and **k** present nearly all instances with high-confidence predictions, corresponding to those with correct predictions. **e** and **l** display the probability density distributions of correct (red curve) and incorrect (blue curve) predictions, along with Youden index curve (purple, dotted). It is evident that the uncertainty threshold (black dotted line) effectively separates correct from incorrect predictions. **f** and **m** illustrate the distribution of instances with varying levels of uncertainty. Intuitively, there are fewer high-confidence incorrect predictions (red fork) compared to high-confidence correct predictions (red dot). In contrast, low-confidence incorrect predictions (blue fork) constitute a large proportion of low-confidence predictions. **g** and **n** illustrate the distribution of uncertainty concerning correct and incorrect predictions, which further substantiates that correct predictions are associated with lower uncertainties and vice versa.

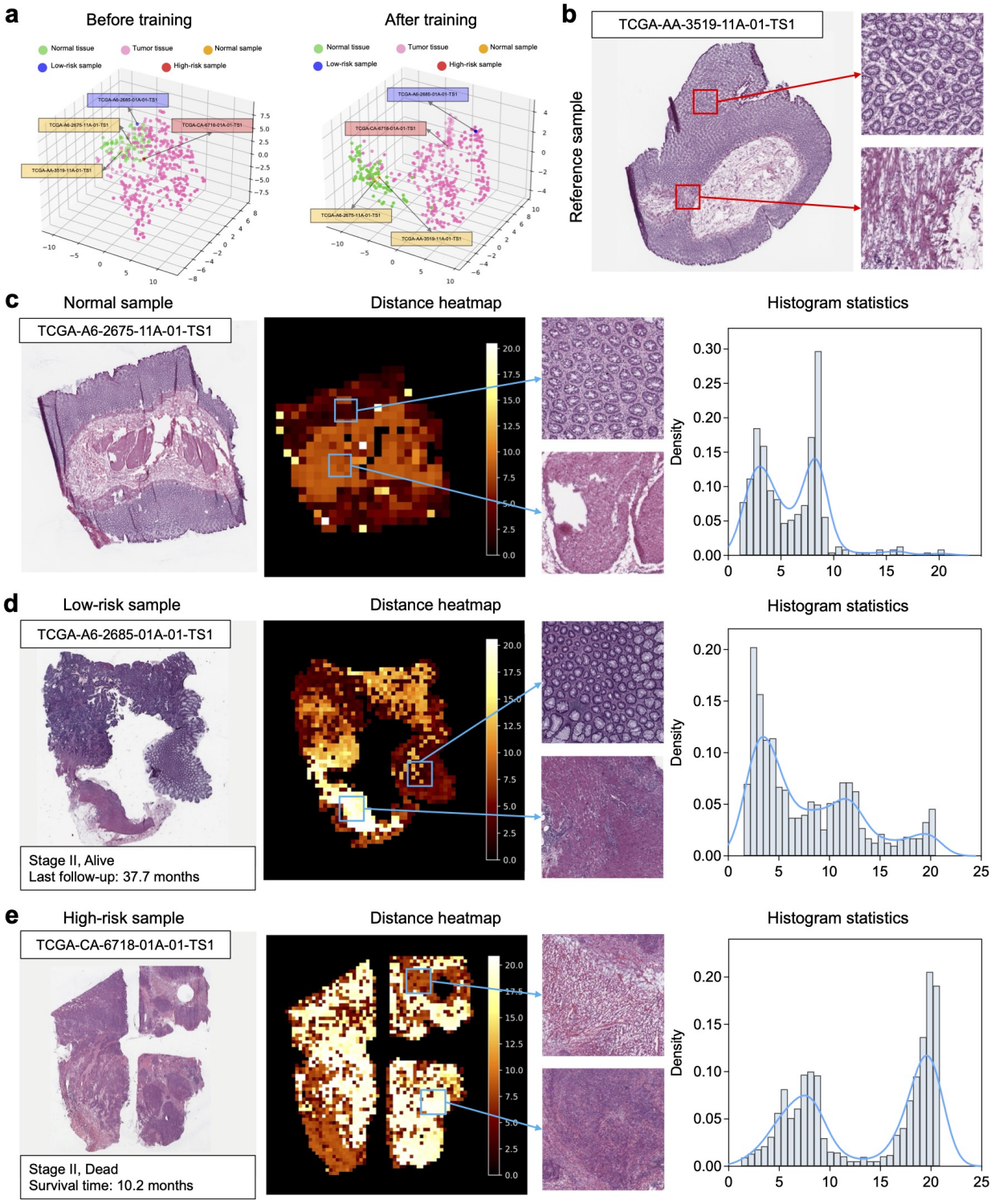

Fig. S2. **Visualization of comparative analysis between tumor and normal tissues on Curriculum II for COAD dataset. a**, t-SNE visualization shows the distribution of tumor and normal tissues before and after training. **b**, a normal sample as reference. **c**, **d**, **e**, Three cases of comparisons are visualized, including a normal sample (**c**), a low-risk sample (**d**, stage II, survival times over 37.7 months), and a high-risk sample (**e**, stage II, decreased at 10.2 months). The first column exhibits the original WSIs of these samples, the second column illustrates the distance heatmaps along with two representative tiles, and the third column shows the histogram statistics of the distance heatmaps.

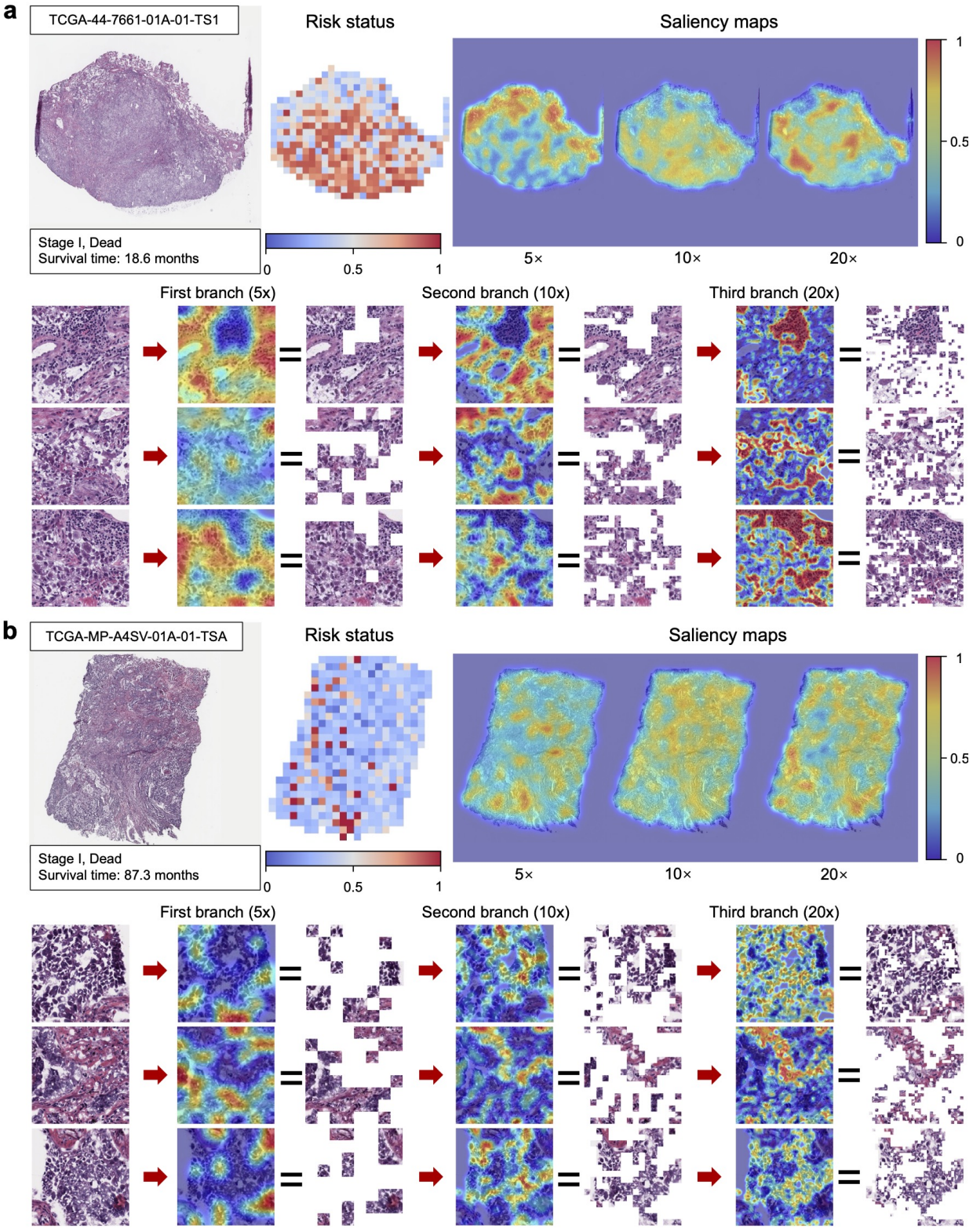

Fig. S3. **Visualizations of saliency maps across multi-magnification WSIs on Curriculum I for LUAD dataset. Related to Figure 5.** This includes a representative high-risk case (**a**) which is stage I and deceased at 18.6 months, a representative low-risk case(**b**) which is stage I and deceased at 87.3 months. **a**, **b**, The first row shows original WSI, patch-level risk status estimation, and saliency maps of WSIs at different magnifications (i.e., 5×, 10×, and 20×). The second row exhibits derived from the low-magnification branch, aiding in learning fine-grained representations at higher magnifications.

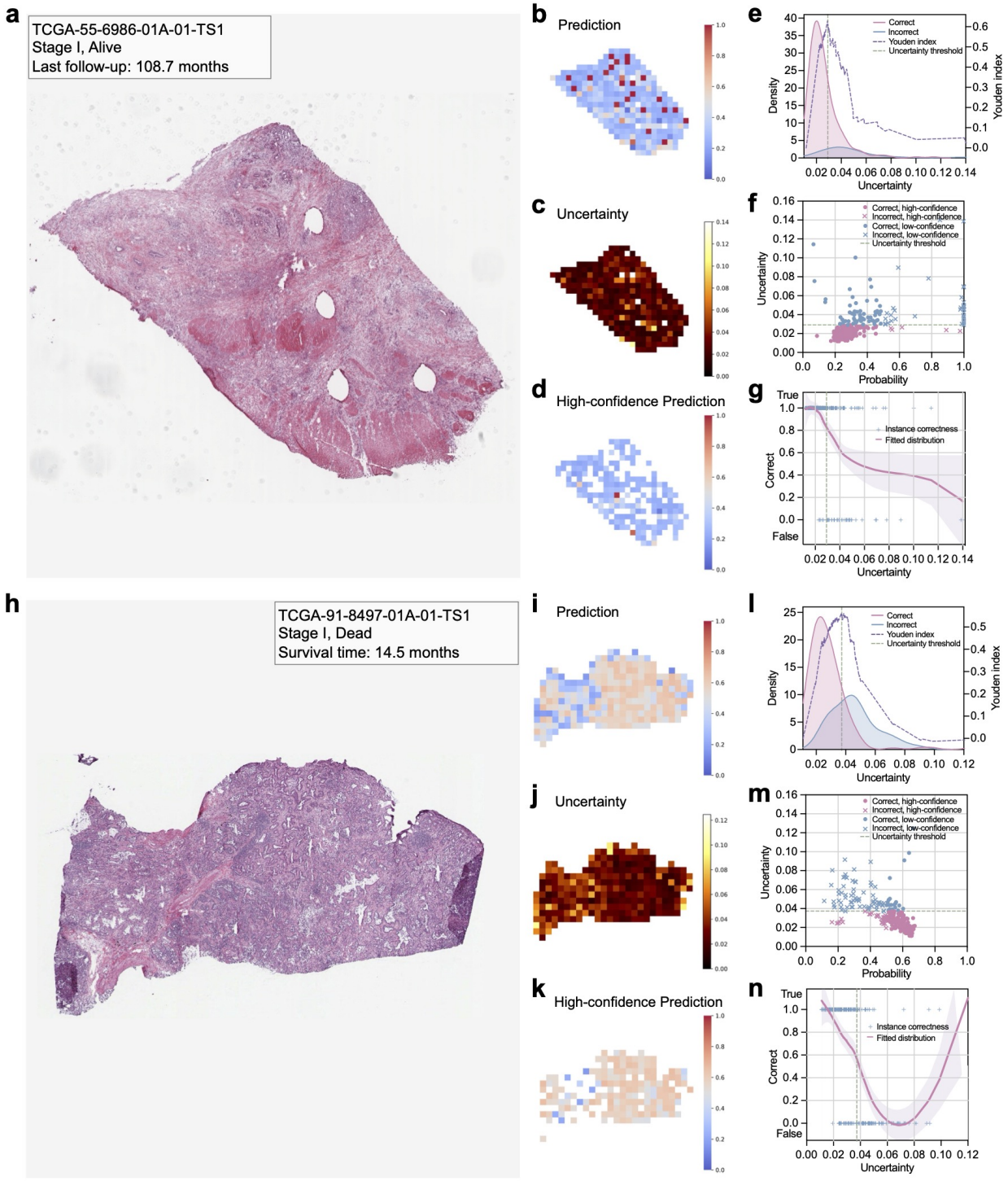

Fig. S4. **Instance uncertainty estimation towards risk stratification on Curriculum I for LUAD dataset. Related to Figure S1.** It includes a high-risk case (stage I, deceased at 14.5 months) and a low-risk case (stage I, survived over 108.7 months). **a** and **h** show the original WSIs. **b** and **i** present the patch-level risk prediction maps. **c** and **j** exhibit the instance uncertainty maps. **d** and **k** present nearly all instances with high-confidence predictions, corresponding to those with correct predictions. **e** and **l** display the probability density distributions of correct (red curve) and incorrect (blue curve) predictions, along with Youden index curve (purple, dotted). It is evident that the uncertainty threshold (black dotted line) effectively separates correct from incorrect predictions. **f** and **m** illustrate the distribution of instances with varying levels of uncertainty. Intuitively, there are fewer high-confidence incorrect predictions (red fork) compared to high-confidence correct predictions (red dot). In contrast, low-confidence incorrect predictions (blue fork) constitute a large proportion of low-confidence predictions. **g** and **n** illustrate the distribution of uncertainty concerning correct and incorrect predictions, which further substantiates that correct predictions are associated with lower uncertainties and vice versa.

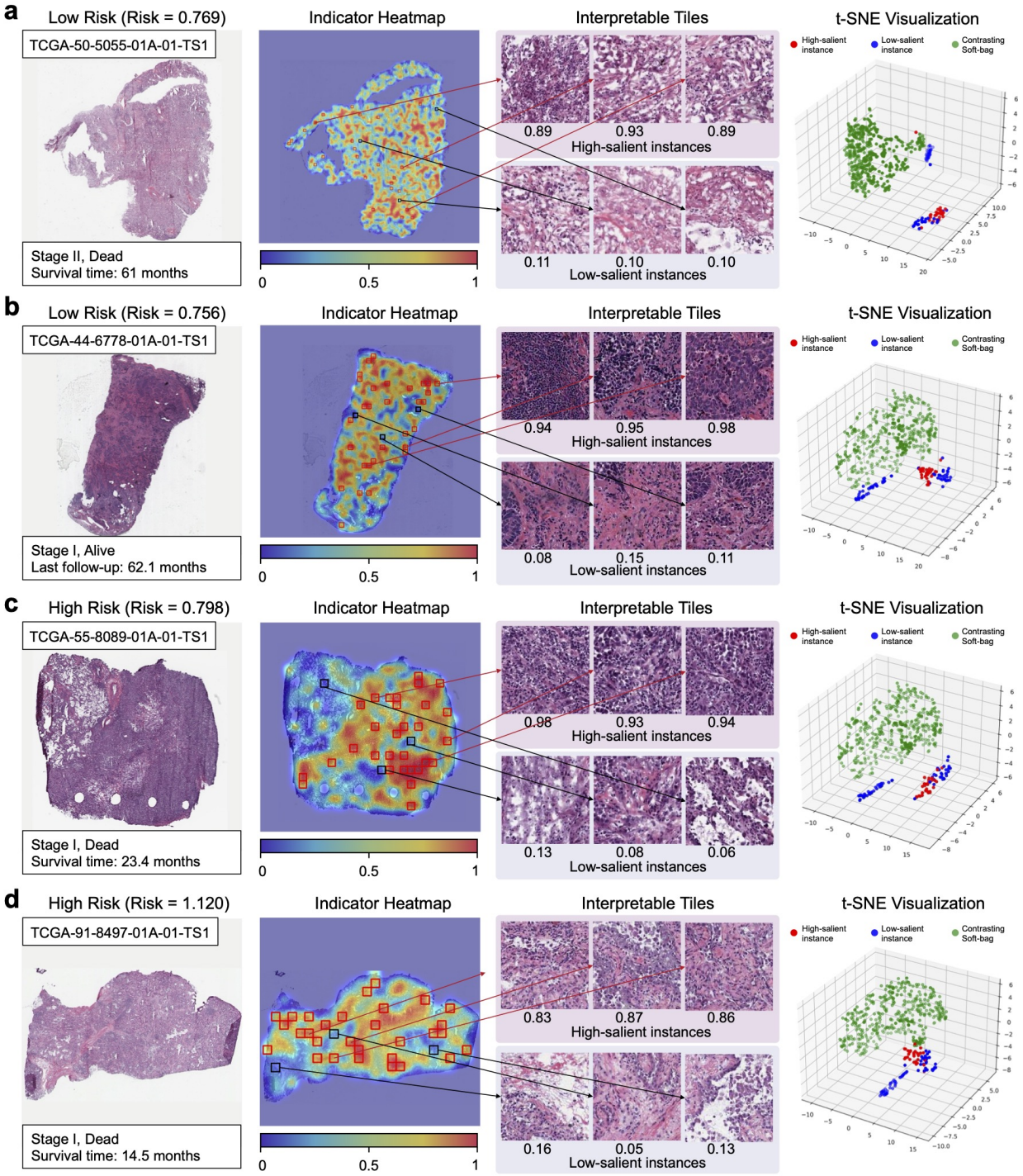

Fig. S5. **Four representative cases from the COAD dataset are visualized on Curriculum II, illustrated with the original WSI, indicator heatmap, interpretable tiles, and t-SNE visualization. Related to Figure 6.** It includes two low-risk patients (**a**, stage II, decreased at 61 months; **b**, and stage I, survival times over 62.1 months) and two high-risk patients (**c**, stage I, decreased at 23.4 months; **d**, stage I, decreased at 14.5 months). **a**, **b**, **c**, **d**, The indicator heatmap visualizes the soft-bag instance selection and its localization within the WSI. Interpretive tiles exhibit several high-salient and low-salient instances, along with their corresponding indicator scores. The t-SNE visualization depicts the embedding distributions of high-salient instances, low-salient instances, and the contrasting soft-bags from other WSIs.

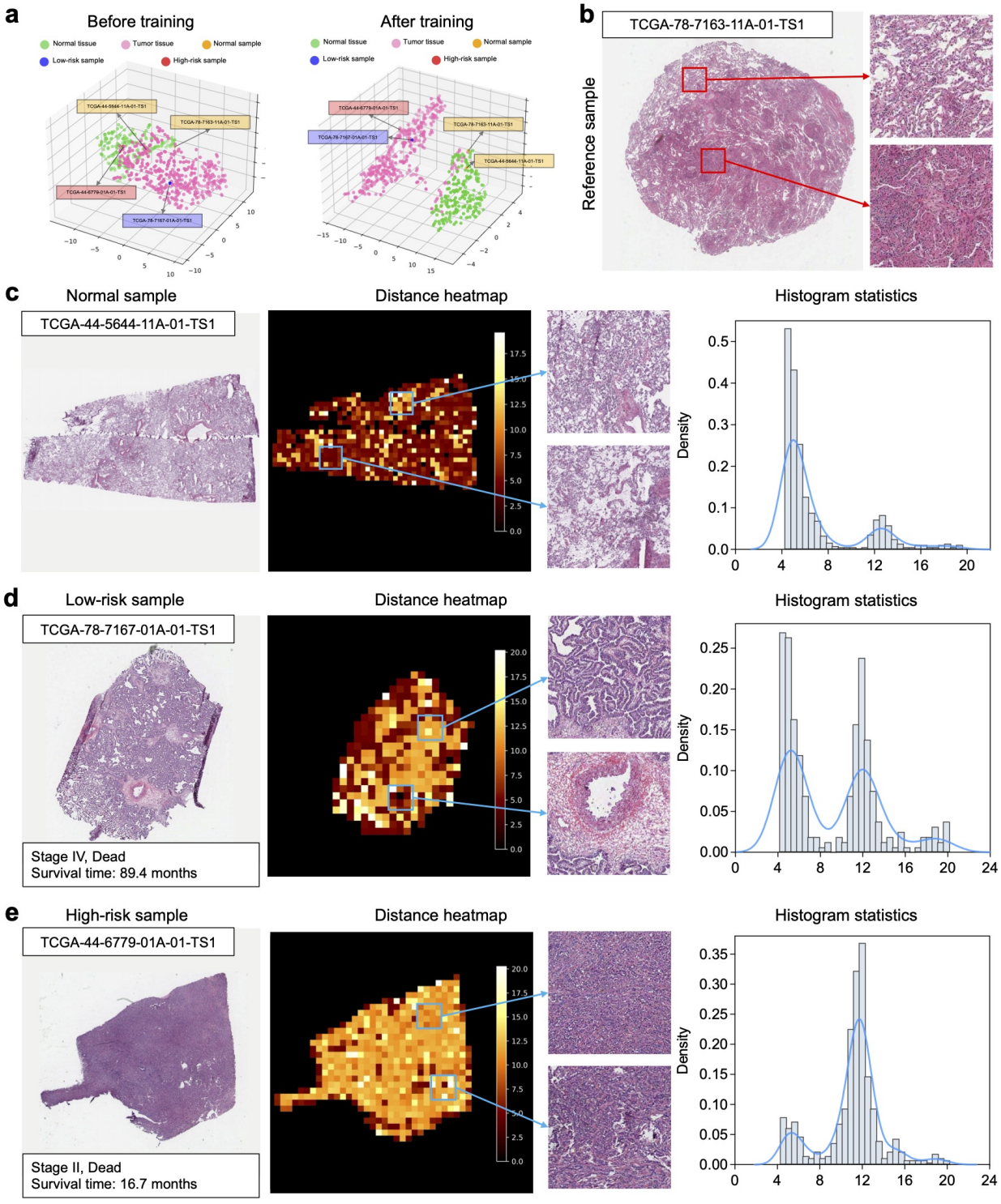

Fig. S6. **Visualization of comparative analysis between tumor and normal tissues on Curriculum II for LUAD dataset. Related to Figure S2. a**, t-SNE visualization shows the distribution of tumor and normal tissues before and after training. **b**, a normal sample as reference. **c**, **d**, **e**, Three cases of comparisons are visualized, including a normal sample (**c**), a low-risk sample (**d**, stage IV, decreased at 89.4 months), and a high-risk sample (**e**, stage II, decreased at 16.7 months). The first column exhibits the original WSIs of these samples, the second column illustrates the distance heatmaps along with two representative tiles, and the third column shows the histogram statistics of the distance heatmaps.

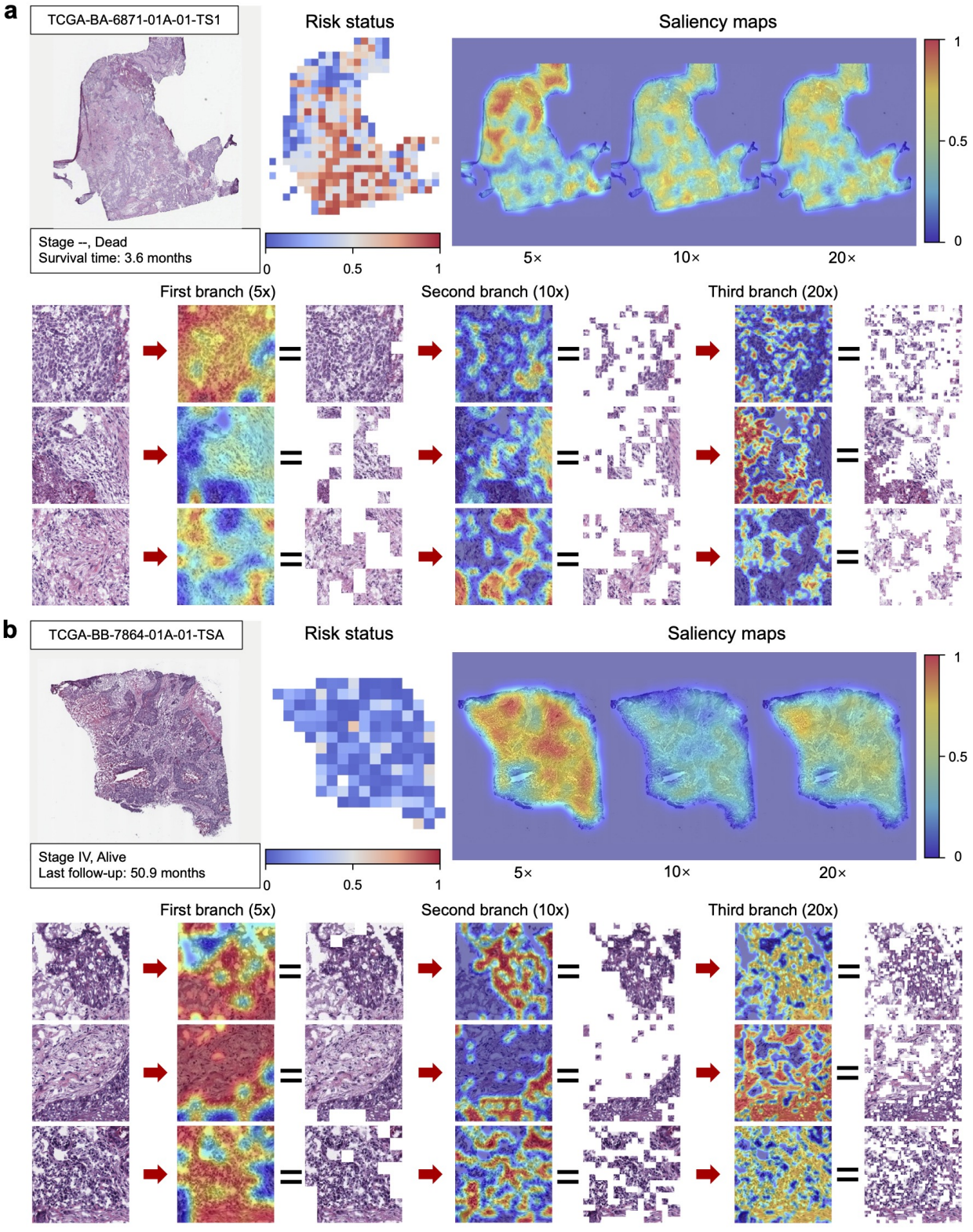

Fig. S7. **Visualizations of saliency maps across multi-magnification WSIs on Curriculum I for HNSC dataset. Related to Figure 5.** This includes a representative high-risk case (**a**) which is stage – and deceased at 3.6 months, a representative low-risk case(**b**) which is stage IV and survived over 50.9 months. **a**, **b**, The first row shows original WSI, patch-level risk status estimation, and saliency maps of WSIs at different magnifications (i.e., 5×, 10×, and 20×). The second row exhibits derived from the low-magnification branch, aiding in learning fine-grained representations at higher magnifications.

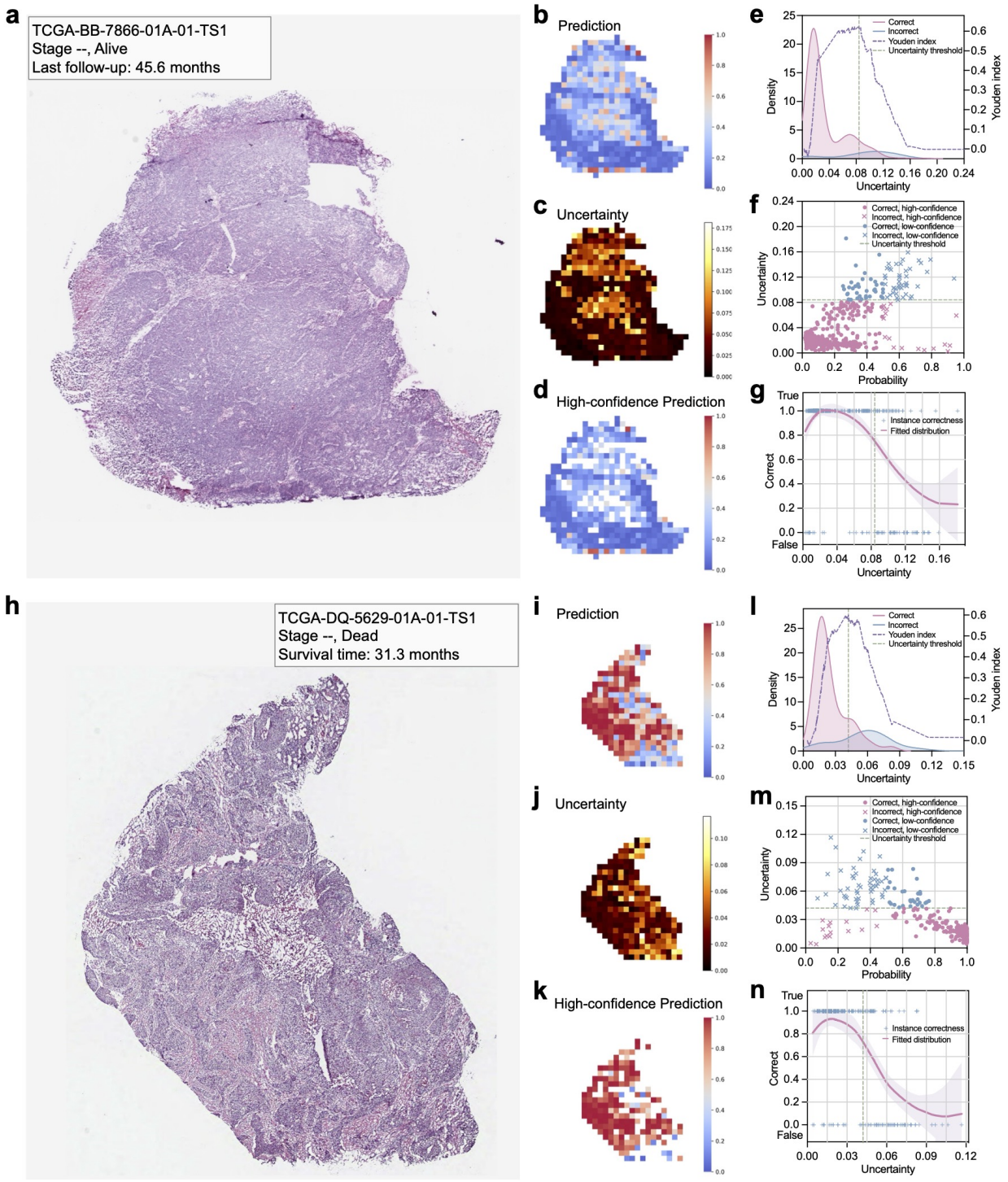

Fig. S8. **Instance uncertainty estimation towards risk stratification on Curriculum I for HNSC dataset. Related to Figure S1.** It includes a high-risk case (stage –, deceased at 31.3 months) and a low-risk case (stage –, survived over 45.6 months). **a** and **h** show the original WSIs. **b** and **i** present the patch-level risk prediction maps. **c** and **j** exhibit the instance uncertainty maps. **d** and **k** present nearly all instances with high-confidence predictions, corresponding to those with correct predictions. **e** and **l** display the probability density distributions of correct (red curve) and incorrect (blue curve) predictions, along with Youden index curve (purple, dotted). It is evident that the uncertainty threshold (black dotted line) effectively separates correct from incorrect predictions. **f** and **m** illustrate the distribution of instances with varying levels of uncertainty. Intuitively, there are fewer high-confidence incorrect predictions (red fork) compared to high-confidence correct predictions (red dot). In contrast, low-confidence incorrect predictions (blue fork) constitute a large proportion of low-confidence predictions. **g** and **n** illustrate the distribution of uncertainty concerning correct and incorrect predictions, which further substantiates that correct predictions are associated with lower uncertainties and vice versa.

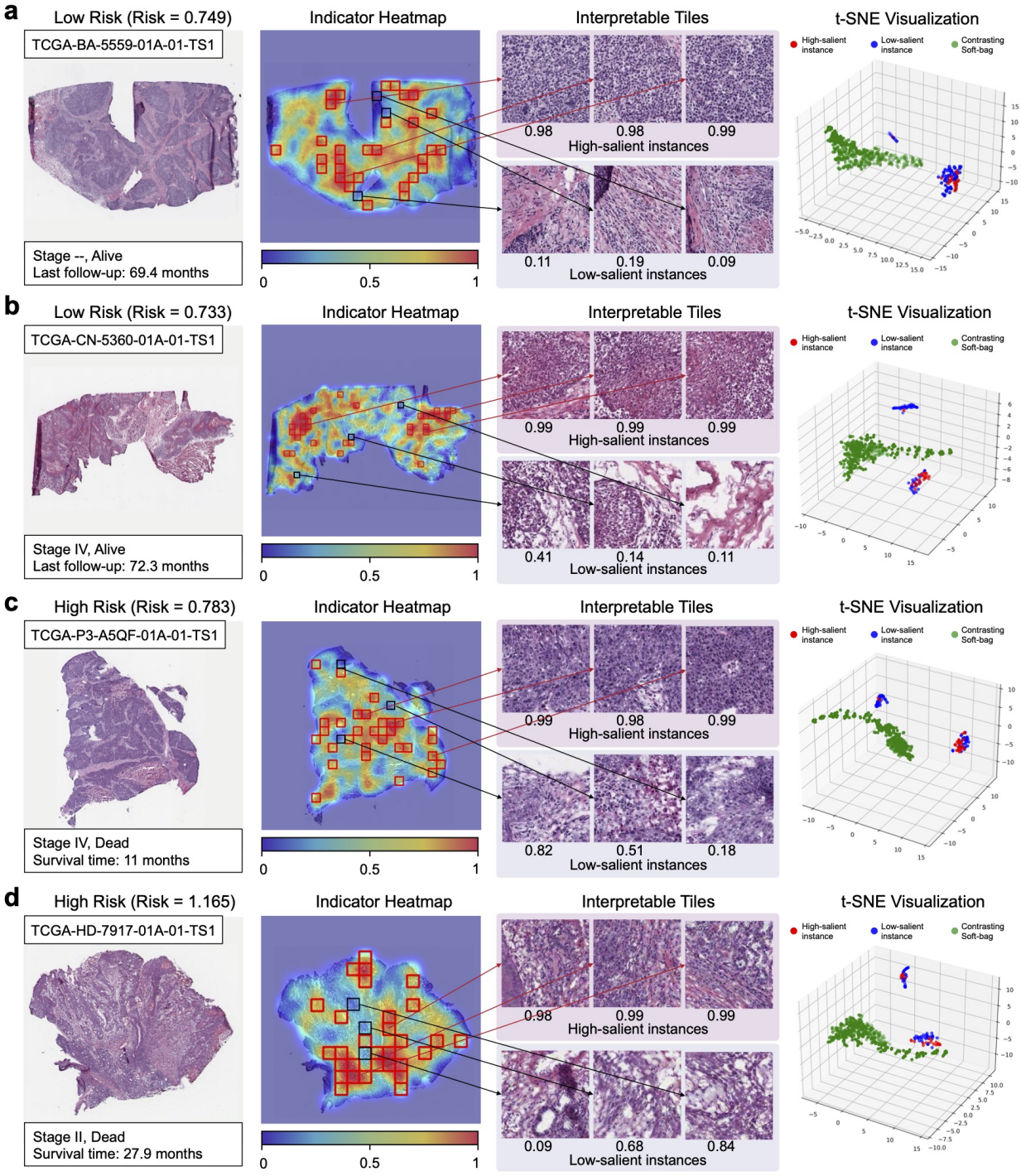

Fig. S9. **Four representative cases from the HNSC dataset are visualized on Curriculum II, illustrated with the original WSI, indicator heatmap, interpretable tiles, and t-SNE visualization. Related to Figure 6.** It includes two low-risk patients (**a**, stage –, survival times over 69.4 months; **b**, and stage IV, survival times over 72.3 months) and two high-risk patients (**c**, stage IV, decreased at 11 months; **d**, stage II, decreased at 27.9 months). **a**, **b**, **c**, **d**, The indicator heatmap visualizes the soft-bag instance selection and its localization within the WSI. Interpretive tiles exhibit several high-salient and low-salient instances, along with their corresponding indicator scores. The t-SNE visualization depicts the embedding distributions of high-salient instances, low-salient instances, and the contrasting soft-bags from other WSIs.

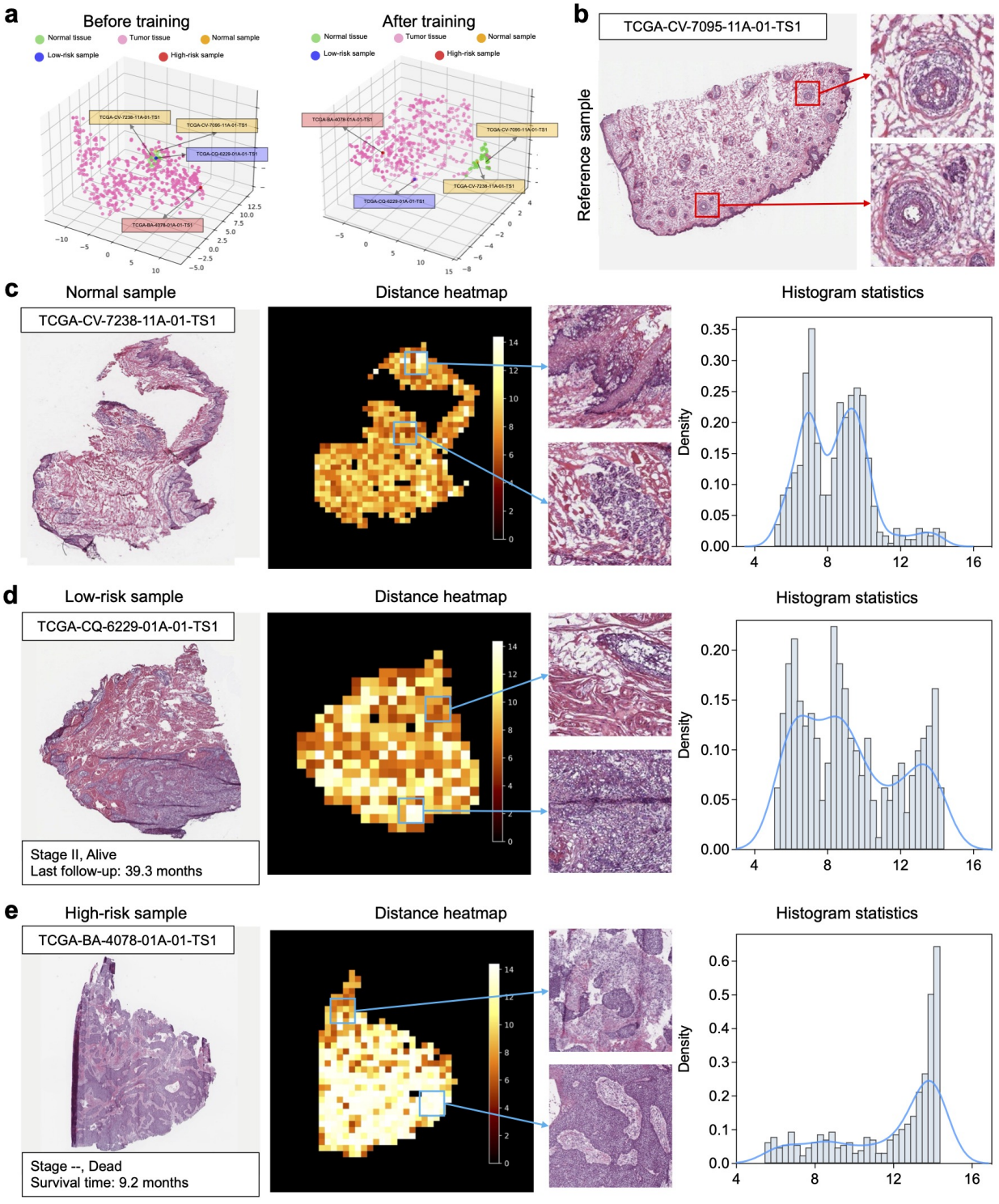

Fig. S10. **Visualization of comparative analysis between tumor and normal tissues on Curriculum II for HNSC dataset. Related to Figure S2. a**, t-SNE visualization shows the distribution of tumor and normal tissues before and after training. **b**, a normal sample as reference. **c**, **d**, **e**, Three cases of comparisons are visualized, including a normal sample (**c**), a low-risk sample (**d**, stage II, survival times over 39.3 months), and a high-risk sample (**e**, stage –, decreased at 9.2 months). The first column exhibits the original WSIs of these samples, the second column illustrates the distance heatmaps along with two representative tiles, and the third column shows the histogram statistics of the distance heatmaps.