2021 IEEE/CVF International Conference on Computer Vision (ICCV) | 978-1-6654-2812-5/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICCV48922.2021.00398

Multimodal Co-Attention Transformer for Survival Prediction in Gigapixel Whole Slide Images

Richard J. Chen[1,2,3,4], Ming Y. Lu[1,3,4], Wei-Hung Weng[5], Tiffany Y. Chen[1,3,4], Drew FK. Williamson[1,3,4], Trevor Manz[1,2], Maha Shady[1,2,3,4], Faisal Mahmood[1,3,4]

[1]Department of Pathology, Brigham and Women's Hospital
[2]Department of Biomedical Informatics, Harvard Medical School
[3]Cancer Program, Broad Institute of Harvard and MIT
[4]Cancer Data Science Program, Dana-Farber Cancer Institute
[5]Computer Science and Artificial Intelligence Laboratory, MIT
richardchen@g.harvard.edu, faisalmahmood@bwh.harvard.edu

Abstract

Survival outcome prediction is a challenging weakly-supervised and ordinal regression task in computational pathology that involves modeling complex interactions within the tumor microenvironment in gigapixel whole slide images (WSIs). Despite recent progress in formulating WSIs as bags for multiple instance learning (MIL), representation learning of entire WSIs remains an open and challenging problem, especially in overcoming: 1) the computational complexity of feature aggregation in large bags, and 2) the data heterogeneity gap in incorporating biological priors such as genomic measurements. In this work, we present a Multimodal Co-Attention Transformer (MCAT) framework that learns an interpretable, dense co-attention mapping between WSIs and genomic features formulated in an embedding space. Inspired by approaches in Visual Question Answering (VQA) that can attribute how word embeddings attend to salient objects in an image when answering a question, MCAT learns how histology patches attend to genes when predicting patient survival. In addition to visualizing multimodal interactions, our co-attention transformation also reduces the space complexity of WSI bags, which enables the adaptation of Transformer layers as a general encoder backbone in MIL. We apply our proposed method on five different cancer datasets (4,730 WSIs, 67 million patches). Our experimental results demonstrate that the proposed method consistently achieves superior performance compared to the state-of-the-art methods.

1. Introduction

Though deep learning has revolutionized computer vision in many disciplines, gigapixel whole-slide imaging (WSI) in computational pathology remains a complex computer vision domain with barriers that render current approaches infeasible for supervised learning tasks such as cancer prognosis. In image classification of natural images, the goal is usually to assign an image-level label to an image with approximate size 256 $\times$ 256 pixels, with the label clearly visible and well-represented in the image. In pathology, WSIs break these assumptions as images exhibit enormous heterogeneity and can be as large as 150,000 $\times$ 150,000 pixels. Depending on the problem, labels for slide-level classification may be: 1) localized in a small pixel region that occupies a tiny proportion of the total image (*i.e.* - a needle-in-a-haystack problem such as differentiating normal tissue vs micro-metastases) [4, 3, 5, 45], or 2) spanning the entire composition of a WSI and dependent on the interactions of its components (*i.e.* - a fine-grained visual recognition problem such as one that involves understanding the complex milieu of stroma, tumor aggregates, immune cells and other visual concepts) [68, 7, 47, 22, 40, 21, 6, 39, 15].

Due to the enormous gigapixel resolutions of WSIs, many approaches adopt a two-stage multiple instance learning-based (MIL) approach for tractable representation learning of WSIs, in which: 1) instance-level feature representations are extracted from randomly sampled image patches in the WSI, and then 2) global aggregation schemes are applied to the bag of instances to obtain a WSI-level representation for subsequent supervision [23, 59, 11, 38, 69]. Though unable to model complex interactions between instances, MIL is able to solve many needle-in-a-haystack problems in pathology, as the classification of normal tissue vs micro-metastases depends on discriminating only binary instance-level visual concepts [3, 35]. Survival outcome prediction, however, is a challenging ordinal regression task

978-1-6654-2812-5/21/$31.00 ©2021 IEEE
DOI 10.1109/ICCV48922.2021.00398

that aims to predict relative risk of cancer death, and fits into the latter class of fine-grained visual recognition problems [70]. In contrast to needle-in-a-haystack problems, survival outcome prediction requires modeling a heterogeneous spectrum of visual concepts in the tumor microenvironment that are indiscernable by conventional MIL approaches, *e.g.* - the co-localization of tumor cells with lymphocyte infiltrates that is associated with favorable prognosis, which would require modeling mid-to-long range interactions between instances in the WSI [48, 25, 1].

Though often approached as a weakly-supervised task using only gigapixel WSIs, survival outcome prediction is traditionally framed as a multimodal learning task in which genomic information can be used as an additional modality for supervision or integration. In the current state-of-the-art, the manual assessment of histology and genomics by pathologists is the gold standard for patient triage, risk assessment, stratification into treatment groups [33]. In further extending weakly-supervised learning with multimodal fusion mechanisms, survival prediction faces an additional challenge due to the large data heterogeneity gap between WSIs and genomics: WSIs represented as bags containing tens of thousands of image patches as instances, while genomic features are often represented as 1×1 tabular attributes. As a result, many approaches use late fusion mechanisms for feature integration, which prevents learning important multimodal interactions [42, 9, 10]. Overall, cancer prognostication using WSIs is both a difficult weakly-supervised learning and multimodal learning problem, and is a grand challenge in the characterization of disease progression of many cancer subtypes.

To address these challenges, we propose an interpretable, weakly-supervised, multimodal learning framework called MCAT (**M**ultimodal **C**o-**A**ttention **T**ransformer) that learns a dense co-attention mapping between WSIs and genomics for interpretable survival outcome prediction. Inspired by deep learning approaches in Visual Question Answering (VQA) that learn relationships attributing how word embeddings attend to salient objects in an image when answering a question [34, 29, 64, 28, 44, 63], in our framework, we learn how instance-level histology features attend to genes when predicting patient survival. One of the key contributions of our work is that we use a cross-modality attention (or co-attention) called genomic-guided co-attention (GCA) as an early fusion strategy for identifying informative instances from a large permutation-invariant set / MIL bag using genomic features as queries (formulated in an embedding space). This yields two advantages for survival outcome prediction:

1. In comparison to late fusion-based architectures that concatenate the WSI-level bag representation with genomic features, our GCA layer captures multimodal interactions that relate histology-based visual concepts to gene embeddings similar to VQA, visualized as a WSI-level attention heatmap for each genomic embedding.

2. We demonstrate how the GCA layer also reduces the effective "sequence length" of WSI bags from M instance-level patch features to N gene-guided visual concepts, where N is the effective sequence length of set of gene embeddings (and $M \gg N$). This allows us to develop more sophisticated feature aggregation strategies using self-attention and Transformers for supervision using entire WSIs, which has previously been impossible. In § 3.3, we draw connections between MIL that operate on set-based data structures (bags), and Transformers.

Results in Table 1 shows that MCAT outperforms existing state-of-the-art weakly-supervised methods for survival outcome prediction using gigapixel WSIs, as well as multimodal networks (augmented from existing methods) that would conventionally integrate WSI with genomics via late fusion. We conduct our ablation study on five large publicly-available cancer datasets, and demonstrate that MCAT consistently improves on all prior approaches by $3.0\% - 6.87\%$. Lastly, we visualize the gene-guided visual concepts as heatmaps to analyze feature interactions between WSIs and genomics, shown in Figures 2 and 3, and assess patterns that emerge from how morphological features attend to each gene. Our code is made available at: https://github.com/mahmoodlab/MCAT.

2. Related Work

2.1. Weak Supervision in Gigapixel Images

Recent work has demonstrated remarkable progress in using multiple Instance Learning (MIL) and other set-based deep learning approaches for learning tasks in gigapixel images [23, 5, 52, 51, 62, 69, 37, 36]. Edwards and Storkey [17] and Zaheer *et al*. [67] proposed one of the first neural network architectures for supervised learning on sets, followed by Ilse *et al*. [24] later extending set-based deep learning as a general framework for MIL, with applications to pathology. Xu *et al*. [60] proposed an MIL-based label enrichment approach for tissue semantic segmentation without pixel-annotations. Lu *et al*. [38], Zhu *et al*. [70], Yao *et al*. [61, 62], Zhao *et al*. [69] explored different strategies for global pooling over patch-based instances. Though demonstrating impressive results in cancer classification, MIL-based approaches in pathology have generally focused only on instance-level feature extraction, and have not yet explored modeling global, long-range interactions through permutation-equivariant feature aggregation techniques such as attention mechanisms.

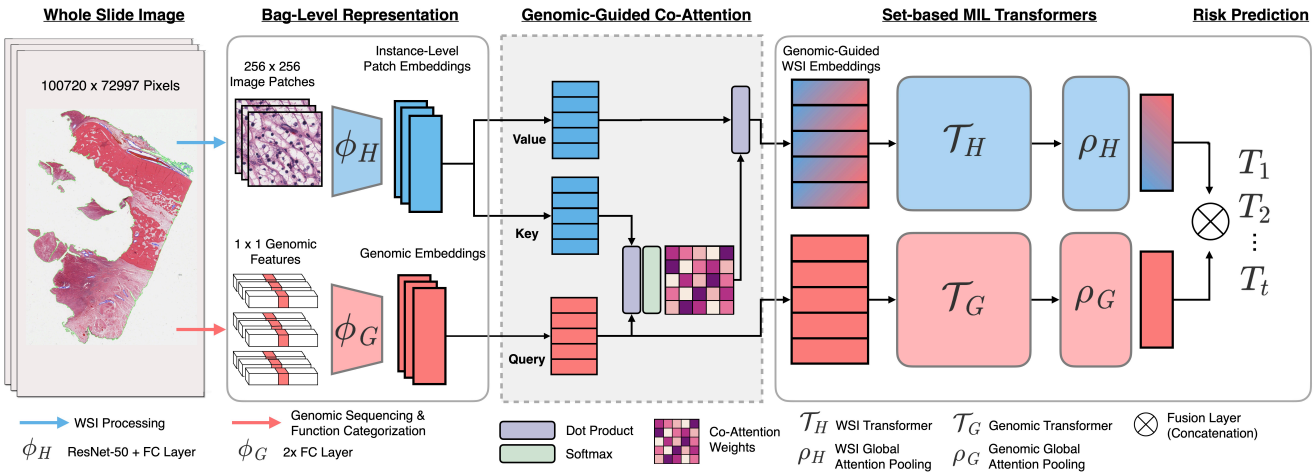

Figure 1: Overview of the Multimodal Co-Attention Transformer (MCAT) architecture. From gigapixel WSIs and genomic features, we formulate both modalities as bags representations, from which we use: 1) Genomic-Guided Co-Attention to capture multimodal interactions, and 2) set-based MIL Transformers as feature aggregation for survival outcome prediction.

2.2. Attention in Set-Based Deep Learning

Since the seminal work by Vaswani *et al*. [53], attention mechanisms have since seen widespread adoption across many different domains outside of neural machine translation such as language model pretraining [14, 55, 56], visual recognition [46, 16], visual question answering [34, 29, 64, 28, 44, 63], graph neural networks [54], and point clouds [31, 18]. Outside of language modeling, Lee *et al*. [31] developed the Set Transformer framework, which extends the original language Transformer to general set-structured data structures such as point clouds and counting problems similar to [67]. Dosovitskiy *et al*. [16] proposed using Transformer architectures for vision pretraining in natural images, in which 224×224 images were formulated as a sequence flattened 16×16 image patches. Recently, Kalra *et al*. [26] used Set Transformers for lung cancer subtyping using bags of 100 randomly sampled histology patches. For large-scale representation learning of entire WSIs, though WSIs can be naturally formulated as a sequence / bag of histology patches, in comparison to word embeddings that have a max sequence length of at most 512, the average bag size of a WSI contains approximately 15,000 256×256 image patches at $20\times$ magnification. with a max sequence length of 200,000 patches. Due to the large space complexity of WSI bags, using Transformers and other stacked self-attention network architectures in MIL-related tasks is computationally infeasible.

2.3. Multimodal Deep Learning

Learning joint representations via multimodal deep learning is a challenging task due to the heterogeneous statistical properties and noise levels across modalities [43, 2]. To learn shared representations, fusion operators such as vector concatenation, element-wise sum, element-wise multiplication (Hadamard Product), bilinear pooling (Kronecker Product), and co-attention mechanisms are often used in many multimodal learning tasks such as VQA [19, 29, 28, 20], sentiment analysis [65], survival analysis [42, 9, 10], and other tasks in medicine [8, 41]. In pathology, Mobadersany *et al*. [42] uses vector concatenation to integrate histology and genomic features for survival outcome prediction. Later, Chen *et al*. [9] uses Kronecker Product fusion to integrate image, graph, and genomic-based features. Though multimodal, many of these approaches are late fusion-based, in that features are only fused towards the penultimate network layers and provide limited interpretability of multimodal interactions. Moreover, in contrast with multimodal fusion approaches in VQA that can relate image features to word embeddings using co-attention learning, current multimodal work using WSIs do not have similar interpretability mechanisms that can relate histology features in WSIs to genomics.

3. Method

In this section, we present our overall framework, the Multimodal Co-Attention Transformer (MCAT), for weakly-supervised and multimodal learning using WSIs and genomics for survival outcome prediction, illustrated in Figure 1. In § 3.1, we present our formulation of WSI and genomic representations as bags using instance-level feature extraction. In § 3.2, we present our core method, the Genomic-Guided Co-Attention (GCA) layer, which learns a dense co-attention mapping between bag representations of WSIs and genomics that can visualize multimodal interactions (Figure 2). We also demonstrate how the GCA layer is able to reduce the space complexity of WSI bags, from which in § 3.3, we adapt set-based Transformers for MIL in survival outcome prediction. In § 3.4, we discuss

implementation details, with further information about our survival loss function described in the Supplementary Materials.

3.1. WSI and Genomic Bag Construction

Problem Formulation: Multiple Instance Learning (MIL) is a weakly-supervised learning task and framework that operates on set-based data structures. These set-based data structures are also known as "bags", in which each bag is an unordered (permutation-invariant) set of instances that can be of varying size with incomplete instance-level labels [67]. For single-label classification, given a bag $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_M\} \in \mathbb{R}^{M \times d_{\text{in}}}$ containing d_{in}-dimensional instances with label Y, the goal is to learn a permutation-invariant function $\mathcal{F}$ that predicts the bag label without detailed knowledge of the instances, and has the general form:

$$\mathcal{F}(X) = \zeta\big(\rho\left(\{\phi(\mathbf{x}_i) : \mathbf{x}_i \in X\}\right)\big) \tag{1}$$

where $\phi : \mathbb{R}^{d_{\text{in}}} \rightarrow \mathbb{R}^{d_{\text{out}}}$ is an instance-level function that processes features for each instance independently, $\rho : \mathbb{R}^{m \times d_{\text{out}}} \rightarrow \mathbb{R}^{d_{\text{out}}}$ is a symmetric, permutation-invariant aggregation function that pools the extracted features to a single bag-level feature embedding, and $\zeta : \mathbb{R}^{d_{\text{out}}} \rightarrow \mathbb{R}^{\#\text{ class}}$ is usually a bag-level classifier that further processes the bag-level features, which we use to estimate the hazard function in survival analysis.

In our task, let X represent patient data, $t_{\text{os}} \in \mathbb{R}^+$ be overall survival time (in months), $c \in \{0, 1\}$ be right uncensorship status (death observed) in a single triplet observation in a dataset $\{X_i, t_{i,\text{os}}, c_i\}_{i=1}$. In addition, let $\{W_{ij}\}_{j=1}^{K_i}$ be the set of K_i gigapixel WSIs collected for X_i, and $\mathbf{g}_{i,\text{attr}}$ be a vector of genomic attributes matched with X_i. For ease of notation, we drop i in referring to the ith observation. Our goal is to develop a set-based neural network architecture $\mathcal{F}$ that integrates $\{W_j\}_{j=1}^{K}$ and $\mathbf{g}_{\text{attr}}$ to estimate the hazard function $f_{\text{hazard}}(T = t \mid T \geq t, X) \in [0, 1]$, which measure the probability of the patient surviving after time point t (typically implemented as a Sigmoid activation after the last layer of $\mathcal{F}$) [12, 27, 66]. Instead of estimating t_{os} directly, survival models output a ordinal risk value obtained via the cumulative distribution function $f_{\text{surv}}(T \geq t, X) = \prod_{u=1}^{t}(1 - f_{\text{hazard}}(T = t \mid T \geq t, X))$. We include detailed preliminaries on survival analysis in the Supplementary Materials.

Instance Level Feature Extraction: To represent $\{W_j\}_{j=1}^{K_i}$ as a single bag data structure, we follow bag construction methods in conventional MIL approaches, in which instance-level feature representations are extracted from small image patches in the WSI. In contrast with previous approaches that sample image ROIs [59, 42, 11, 61, 62, 9], we construct our bag using all available tissue information across multiple WSIs for large-scale training. For all WSIs W_j, we patch the tissue-containing image regions into a set of non-overlapping 256×256 patches, which we use as input into an instance-level function ϕ_H, implemented as a ResNet-50 CNN + FC layer (pretrained on ImageNet) that extracts d_k-dim feature embeddings $\mathbf{h} \in \mathbb{R}^{d_k \times 1}$. For M total histology patches across all W_j, we pack the extracted patch embeddings into a bag $H_{\text{bag}} \in \mathbb{R}^{M \times d_k}$. In utilizing the entire tissue microenvironment across multiple WSIs, the average bag size during training and inference contains approximately $M = 15,231$ instances, with some bags having up to 17 gigapixel WSIs and $230,000$ instances. In conventional MIL approaches, from here, global aggregation techniques such as $\text{SUM}(\cdot)$ can be applied to form $\mathbf{h}_{\text{final}}$, followed by concatenation or bilinear pooling with genomic feature vector $\mathbf{g}_{\text{attr}}$ as late fusion.

Formulating Genes in an Embedding Space: Genomic features such as gene mutation status, copy number variation, and bulk RNA-Seq abundance are typically quantified as 1×1 measurements, or attributes, which alone do not incorporate any semantic information that would describe the functional impact of a gene in a biological system. To obtain more expressive, embedding-like feature representations similar to word embeddings in NLP, we categorize genes into N different sets with similar biological functional impact (*e.g.* - oncogenesis or cell differentiation). Let $\{B_n\}_{n=1}^{N}$ indicate unique functional categories obtained from [49, 32]. For each genomic attribute $\text{att}_i \in \mathbf{g}_{\text{attr}}$, we assign att_i to gene set $\mathbf{g}_n$ if $\text{att}_i \in B_n$, which we use as input to a genomics-based instance-level function ϕ_g parameterized using a FC layer. In applying ϕ_g instance-wise over all categorized gene sets, we obtain genomic embeddings $\{\mathbf{g}_n \in \mathbb{R}^{d_k \times 1}\}_{n=1}^{N}$, which we pack into a bag data structure $G_{\text{bag}} \in \mathbb{R}^{N \times d_k}$. In our implementation, we use $N = 6$ functional categories obtained from [32] to define the following genomic embeddings: 1) Tumor Supression, 2) Oncogenesis, 3) Protein Kinases, 4) Cellular Differentiation, 5) Transcription, and 6) Cytokines and Growth.

3.2. Genomic-Guided Co-Attention Layer

Due to the data heterogeneity gap between gigapixel WSI and genomic features, current multimodal approaches in pathology are limited to only incorporating late fusion, which does not capture interpretable genotype-phenotype interactions that exist in the tumor microenvironment. In reformulating both WSIs and genomic features as bag representations H_{bag} and G_{bag}, we can develop more complex feature aggregation strategies that directly model pairwise interactions between instance-level feature embeddings in H_{bag} and genomic embeddings in G_{bag}. In this section, we introduce Genomic-Guided Co-Attention (GCA), analogous to the standard Transformer attention that relate image-grid and word embeddings in VQA [53] (Fig. 2). GCA uses $G_{\text{bag}} \in \mathbb{R}^{N \times d_k}$ to guide the feature aggregation of $H_{\text{bag}} \in \mathbb{R}^{M \times d_k}$ into a clustered set of gene-guided visual

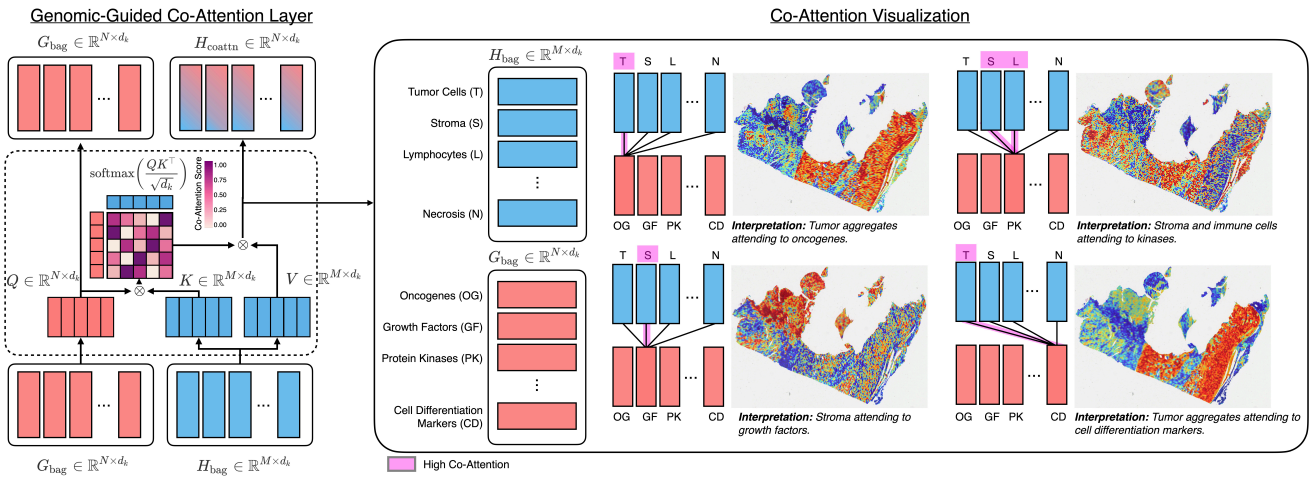

Figure 2: Overview of the Genomic-Guided Co-Attention (GCA) layer with co-attention visualization. The GCA layer uses G_{bag} (red) as queries to guide the aggregation of H_{bag} (blue) into $\widehat{H}_{\text{coattn}}$ (red/blue) using computed co-attention weights A_{coattn}. From A_{coattn}, we can visualize how each image patch in the gigapixel WSI attends to each genomic embedding.

concepts $\widehat{H}_{\text{bag}} \in \mathbb{R}^{N \times d_k}$, using the following mapping:

$$\begin{aligned}\text{CoAttn}_{G \to H}(G, H) &= \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right) \\ &= \text{softmax}\left(\frac{\mathbf{W}_q G H^\top \mathbf{W}_k^\top}{\sqrt{d_k}}\right)\mathbf{W}_v H \to A_{\text{coattn}}\mathbf{W}_v H \to \widehat{H}\end{aligned} \tag{2}$$

where $\mathbf{W}_q, \mathbf{W}_k, \mathbf{W}_v \in \mathbb{R}^{d_k \times d_k}$ are trainable weight matrices multiplied to the queries G_{bag} and key-value pair (H_{bag},H_{bag}), and $A_{\text{coattn}} \in \mathbb{R}^{N \times M}$ is the co-attention matrix for computing the weighted average of H_{bag}. Distinct from VQA is the complexity of gigapixel WSI and the disparate bag sizes, in which $M = 15,231$ and $N = 6$. For the task of multimodal survival outcome prediction, we find that: 1) the interpretability of GCA is able to scale up to hundreds of thousands of patches, providing and 2) we can use genomic embeddings in GCA to reduce the complexity of WSI bags.

Interpretation: Intuitively, for a single genomic embedding $\mathbf{g}_n \in G$, the GCA layer scores the pairwise similarity for how much $\mathbf{h}_m$ attends to $\mathbf{g}_n$ for all $\mathbf{h}_m \in H_{\text{bag}}$, written as a row vector $[a_{n1}, a_{n2}, \ldots, a_{nm}] \in A_{\text{coattn}}$. These attention weights are then applied element-wise to H_{bag}, which constructs a new WSI-level feature embedding $\widehat{\mathbf{h}}_n \in \mathbb{R}^{n \times 1}$ that reflects the biological function of $\mathbf{g}_n$. For example, if $\mathbf{g}_n$ is a genomic embedding that expresses the underlying biological pathways responsible for tumor formation, A_{coattn} computed by the GCA layer would saliently localize image patches containing tumor cells as high attention, which then aggregates $\widehat{\mathbf{h}}_n$ as a WSI-level representation primarily containing tumor cells. We describe the set of high-attention image patches that attend to a single genomic embedding $\mathbf{g}_n$ as a "gene-guided visual concept", in which patches that are similar in feature space to $\mathbf{g}_n$ would share similar phenotypic information. For N genomic embeddings in G_{bag}, the GCA layer captures up to N different gene-guided visual concepts, which we visualize as attention heatmaps in Figures 2 and 3.

Space Complexity: An important detail of GCA is that we set up Q, K, V such that the bag size of Q is much smaller than K, V. As a result, the query G_{bag} agggregates $H_{\text{bag}} \in \mathbb{R}^{M \times d_k}$ containing M instance-level patch embeddings as $\widehat{H}_{\text{coattn}} \in \mathbb{R}^{N \times d_v}$ containing N WSI-level embeddings, which makes the cost of applying subsequent self-attention layers have asymptoptic complexity $\mathcal{O}\left(N^2 d_v + N^2 d_v\right)$, which is quadratic with respect to N instead of M.

3.3. Set-based MIL Transformers with Survival Prediction

Following the observation in Zaheer *et al.* [67] that set-based network architectures remain permutation-invariant even if the encoder is a stack of permutation-equivariant layers, we can extend the original MIL framework using a set-based MIL Transformer, written as:

$$\begin{aligned}\mathcal{E}^{(l)}\left(H^{(l)}\right) &= \zeta^{(l)}\left(\psi^{(l)}\left(\left\{\phi^{(l)}\left(\mathbf{x}_i\right) : \mathbf{h}_i^{(l)} \in H^{(l)}\right\}\right)\right) \\ \mathcal{F}^{(L)}\left(H^{(L)}\right) &= \zeta^{(L)}\left(\rho^{(L)}\left(\left\{\phi^{(L)}\left(\mathbf{x}_i\right) : \mathbf{h}_i^{(L)} \in H^{(L)}\right\}\right)\right) \\ \mathcal{T}\left(X\right) &= \mathcal{F}^{(L)}\left(\mathcal{E}^{(L-1)}\left(\ldots \mathcal{E}^{(1)}\left(\left\{\left(\mathbf{x}_i\right) : \mathbf{x}_i \in X\right\}\right)\right)\right)\end{aligned} \tag{3}$$

in which $\mathbf{h}_i^{(l)}$ is an arbitrary embedding in the set input $H^{(l)}$ at hidden layer l, $\mathcal{E}^{(l)}$ is a stackable encoder block that replaces ρ in Equation 1 with a permutation-equivariant set function ψ, ϕ and ζ are permutation-invariant functions applied to feature embeddings (either instance-level or bag-level), $\mathcal{F}^{(L)}$ is the original MIL network but now applied

as a global pooling function as the last layer L, and $\mathcal{T}$ is the set-based MIL Transformer [53, 31] that uses stacked permutation-equivariant layers followed by a permutation-invariant pooling function. To show that $\mathcal{E}^{(l)}$ is the encoder block in Transformers, let $\zeta^{(l)}$ be a position-wise FC layer, $\psi^{(l)}$ be the self-attention layer in [53], and note that the position-wise residual mapping and $\mathrm{LayerNorm}$ operations retains permutation-invariance. We can also write $\psi^{(l)}$ more explicitly as the permutation-equivariant set function:

$$\psi^{(l)}\left(\left\{\mathbf{h}_i^{(l)}\right\}_{i=1}^{M}\right) = \left\{\sum_{i=1}^{M} \frac{\exp(\mathbf{h}_i^{(l)}\mathbf{h}_j^{(l)\top})}{d_k \sum_j \exp(\mathbf{h}_i^{(l)}\mathbf{h}_j^{(l)\top})} \cdot \mathbf{h}_i^{(l)} \rightarrow \mathbf{h}_i^{(l+1)}\right\} \tag{4}$$

in which permuting the set $\{\mathbf{h}_i^{(l)}\}$ permutes the update $\{\mathbf{h}_i^{(l+1)}\}$ from the output of $\psi^{(l)}$ in the same order. From this formulation, we can observe that Transformers are a generalization of the shallow set-based data structure commonly used in Equation 1, in which we can compose arbitrary hidden layers using permutation-equivariant functions before global pooling. Using $\widehat{H}_{\text{coattn}} \in \mathbb{R}^{N \times d_v}$ and $G_{bag} \in \mathbb{R}^{N \times d_k}$ as inputs, we construct two MIL Transformers $\mathcal{T}_H, \mathcal{T}_G$ to aggregate feature embeddings in $\widehat{H}_{\text{coattn}}, G_{bag}$. In the process for aggregating features in $\widehat{H}_{\text{coattn}}$, $\psi^{(l)}$ is used to model complex, long-range feature interactions between genomic-guided visual concepts that would otherwise be intractable using the original WSI bag with large M.

To implement ρ_H, ρ_G, following [24], we use the global attention pooling function $\mathcal{F}_{\text{attnpool}}(\cdot)$ to adaptively compute a weighted sum of all embeddings within each respective set to finally construct bag-level features $\mathbf{h}^{(L)}, \mathbf{g}^{(L)}$.

$$\begin{aligned}
&\phi^{(L)}(\mathbf{h}_i^{(l)}) = \mathbf{W}_\phi \mathbf{h}_i^{(l)} \\
&\rho^{(L)}\left(\left\{\mathbf{h}_i^{(L)}\right\}_{i=1}^{M}\right) = \sum_{i=1}^{M} a_i \phi^{(L)}(\mathbf{h}_i^{(L)}) \rightarrow \mathbf{h}^{(l)} \quad \text{where} \\
&a_i = \frac{\exp\left\{\mathbf{W}_\rho\left(\tanh\left(\mathbf{V}_\rho \mathbf{h}_i^{(L)\top}\right) \odot \mathrm{sigm}\left(\mathbf{U}_\rho \mathbf{h}_i^{(L)\top}\right)\right)\right\}}{\sum_{j=1}^{M} \exp\left\{\mathbf{W}_\rho\left(\tanh\left(\mathbf{V}_\rho \mathbf{h}_j^{(L)\top}\right) \odot \mathrm{sigm}\left(\mathbf{U}_\rho \mathbf{h}_j^{(L)\top}\right)\right)\right\}} \\
&\zeta^{(L)}(\mathbf{h}^{(L)}) = \mathbf{W}_\zeta \mathbf{h}^{(L)}
\end{aligned} \tag{5}$$

where $\mathbf{W}_\phi, \mathbf{W}_\rho, \mathbf{V}_\rho, \mathbf{U}_\rho, \mathbf{W}_\zeta \in \mathbb{R}^{d_v \times d_v}$ are trainable weight matrices, $\phi^{(L)}$ and $\rho^{(L)}$ are instance-level and bag-level FC layers respectively, $\rho^{(L)}$ is the global attention pooling operator, and a_i scores how much to weigh embedding $\mathbf{h}_i^{(L)}$ in in the bag-level features $\mathbf{h}^{(L)}$. As a final step, we integrate bag-level features from the output of $\mathcal{T}_G, \mathcal{T}_H$ using simple vector concatenation of $\left[\zeta_h^{(L)}(\mathbf{h}^{(L)}), \zeta_g^{(L)}(\mathbf{g}^{(L)})\right]$, which we process using several FC layers to obtain the final shared representation $\mathbf{h}_{\text{final}}$.

3.4. Implementation Details

MCAT is implemented in PyTorch and trained on a commercial workstation with 4 NVIDIA GTX 2080Ti GPUs. Functional categories used to define the genomic embeddings were obtained from [32], which categorizes genes into the aforementioned $N = 6$ categories based on similar biological functional impact. During training, we used Adam optimization with a learning rate of 2×10^{-4}, weight decay of 1×10^{-5}. Due to samples having varying bag sizes, we use a batch size of 1, with 32 gradient accumulation steps.

4. Experiments

4.1. Datasets & Evaluation Metrics

To validate our proposed method, we used the five largest cancer datasets from The Cancer Genome Atlas (TCGA), a public cancer data consortium that contains matched diagnostic WSIs and genomic data with labeled survival times and censorship statuses[1]. For this study, we used the following cancer types: Bladder Urothelial Carcinoma (BLCA) ($n = 437$), Breast Invasive Carcinoma (BRCA) ($n = 1022$), Glioblastoma & Lower Grade Glioma (GBMLGG) ($n = 1011$), Lung Adenocarcinoma (LUAD) ($n = 515$), and Uterine Corpus Endometrial Carcinoma (UCEC) ($n = 538$). For each patient sample, we collected all diagnostic WSIs used for primary diagnosis, which resulted in 4,370 WSIs collected with an average bag size of 15,231 256×256 patches per image (approx 5 TB of gigapixel images, 67 million patches). For each cancer dataset, we trained our proposed method in a 5-fold cross-validation, and used the cross-validated concordance index (c-Index) to measure the predictive performance of correctly ranking the predicted patient risk scores with respect to overall survival.

4.2. Comparisons with State-of-the-Art

Using the same 5-fold cross-validation splits for evaluating MCAT, we implemented and evaluated several state-of-the-art methods used for survival outcome prediction in computational pathology, training in total 275 models. For all methods, we use the same instance-level feature extraction pipeline for bag construction of WSIs, as well as identical training hyperparameters and loss function for supervision. Table 1 shows the results of all methods on all five cancer dataset benchmarks.

1. **SNN** [30]: As a unimodal baseline for only genomic features, we trained a feedforward network using the Self-Normalizing Network (SNN) architecture from Klambauer *et al*. [30], which has been used previously for survival outcome prediction in the TCGA [9].
2. **Deep Sets** [67]: One of the first neural network architectures for set-based deep learning, which proposes sum pooling over instance-level features.
3. **Attention MIL** [24]: A set-based neural network architecture that replaces sum pooling in Deep Sets with

[1] https://gdc.cancer.gov

Model	BLCA	BRCA	GBMLGG	LUAD	UCEC	Overall
SNN (Genomic Only) [30]	0.541 ± 0.016	0.466 ± 0.058	0.598 ± 0.054	0.539 ± 0.069	0.493 ± 0.096	0.527
Deep Sets (WSI Only) [67]	0.500 ± 0.000	0.500 ± 0.000	0.498 ± 0.014	0.496 ± 0.008	0.500 ± 0.000	0.499
Deep Sets (Concat)	0.604 ± 0.042	0.521 ± 0.079	0.803 ± 0.046	0.616 ± 0.027	0.598 ± 0.077	0.629
Deep Sets (Bilinear Pooling)	0.589 ± 0.050	0.522 ± 0.029	0.809 ± 0.027	0.558 ± 0.038	0.593 ± 0.055	0.614
Attention MIL (WSI Only) [24]	0.536 ± 0.038	0.564 ± 0.050	0.787 ± 0.028	0.559 ± 0.060	$\mathbf{0.625 \pm 0.057}$	0.614
Attention MIL (Concat)	0.605 ± 0.045	0.551 ± 0.077	0.816 ± 0.011	0.563 ± 0.050	0.614 ± 0.052	0.630
Attention MIL (Bilinear Pooling)	0.567 ± 0.034	0.536 ± 0.074	0.812 ± 0.005	0.578 ± 0.036	0.562 ± 0.058	0.611
DeepAttnMISL (WSI Only) [62]	0.504 ± 0.042	0.524 ± 0.043	0.734 ± 0.029	0.548 ± 0.050	0.597 ± 0.059	0.581
DeepAttnMISL (Concat)	0.611 ± 0.049	0.545 ± 0.071	0.805 ± 0.014	0.595 ± 0.061	0.615 ± 0.020	0.634
DeepAttnMISL (Bilinear Pooling)	0.575 ± 0.032	0.577 ± 0.063	0.813 ± 0.022	0.551 ± 0.038	0.586 ± 0.036	0.621
MCAT (Ours)	$\mathbf{0.624 \pm 0.034}$	$\mathbf{0.580 \pm 0.069}$	$\mathbf{0.817 \pm 0.021}$	$\mathbf{0.620 \pm 0.032}$	0.622 ± 0.019	**0.653**

Table 1: Ablation study results assessing c-Index performance of MCAT against several state-of-the-art deep learning-based methods across 5 different cancer datasets.

global attention pooling, in which instances are adaptively weighted using a Softmax function [60, 62].

4. **DeepAttnMISL** [62]: The current state-of-the-art for unimodal survival outcome prediction using WSIs. DeepAttnMISL first applies K-Means clustering to instance-level features, followed by processing each cluster using Siamese networks and then aggregating the cluster features using global attention pooling.
5. **Multimodal Comparisons**: As multimodal comparisons to MCAT, we augment the previous set-based network architectures with two common late fusion mechanisms to integrate bag-level WSI features and genomic features: 1) concatenation [42], and 2) bilinear pooling [19, 65, 57, 9].

Unimodal versus Multimodal: In comparing MCAT with the current state-of-the-art methods for WSI training in computational pathology, Attention MIL and DeepAttnMISL, MCAT achieves superior performance on all benchmarks, with an overall c-Index performance increase of 6.35% and 12.4% respectively. Against the genomic baseline, MCAT achieves a performance increase of 23.9%. In line with similar work in using multimodal fusion to augment supervised learning tasks [65], MCAT improves over its unimodal counterparts across all benchmarks.

MCAT versus Late Fusion: MCAT improves on all multimodal approaches (with varying unimodal WSI network backbones and fusion layers), with a $3.0\% - 6.87\%$ performance increase in overall c-Index. In one-versus-all comparisons in each cancer dataset, MCAT also achieves the highest c-Index performance in 4 out of 5 cancer types, which suggests that MCAT can be used in a general setting for any survival outcome prediction task in computational pathology. In the GBMLGG dataset which has distinct intertumoral heterogeneity due to both low-grade glioma (LGG) and high-grade glioblastoma (GBM) equally represented, MCAT achieves the highest c-Index performance of **0.817** across all models. In visualizing patient stratification results for GBMLGG (Supplementary Materials), we further observe that MCAT demonstrates strong separation of low and high risk cases.

4.3. Ablation Studies

To assess the impact of Transformers in solving MIL tasks, we perform an ablation study that evaluates different variations of MCAT that exclude $\mathcal{T}_H$ or $\mathcal{T}_G$ before global attention pooling, the most commonly-used layer used for feature aggregation in MIL. Table 2 in the Supplementary Materials shows results for MCAT models using: 1) only global attention pooling, 2) $\mathcal{T}_H$ present, 3) $\mathcal{T}_G$ present, and 4) both $\mathcal{T}_H, \mathcal{T}_G$ present (main method), in which we demonstrate Transformer layers improve over conventional global attention pooling in overall c-Index performance. The additive performance increases in adding Transformer layers in the MIL framework suggests that Transformers are able to correctly model pairwise feature interactions that are prognostic for cancer survival within both modalities.

4.4. Attention Visualization

To visualize the genomic-guided WSI embeddings used as input in our set-based MIL Transformer, we overlay the co-attention weights computed for each histology patch $\mathbf{h}_m$ attending to each genomic embedding $\mathbf{g}_n$ with visual assessment from two pathologists, shown in Figure 3 for both a low and high risk case in the BRCA dataset. In addition, we also used Integrated Gradients [50] to visualize the top-10 genes in each embedding with the highest absolute attribution value.

Overall, we observe that the genomic embeddings used in guided co-attention were able to reflect many known genotype-phenotype relationships in cancer pathology. In BRCA, genomic-guided WSI embeddings for tumor suppression, protein kinases and cellular differentiation generally reflected normal stroma, glands, and adipocytes. For

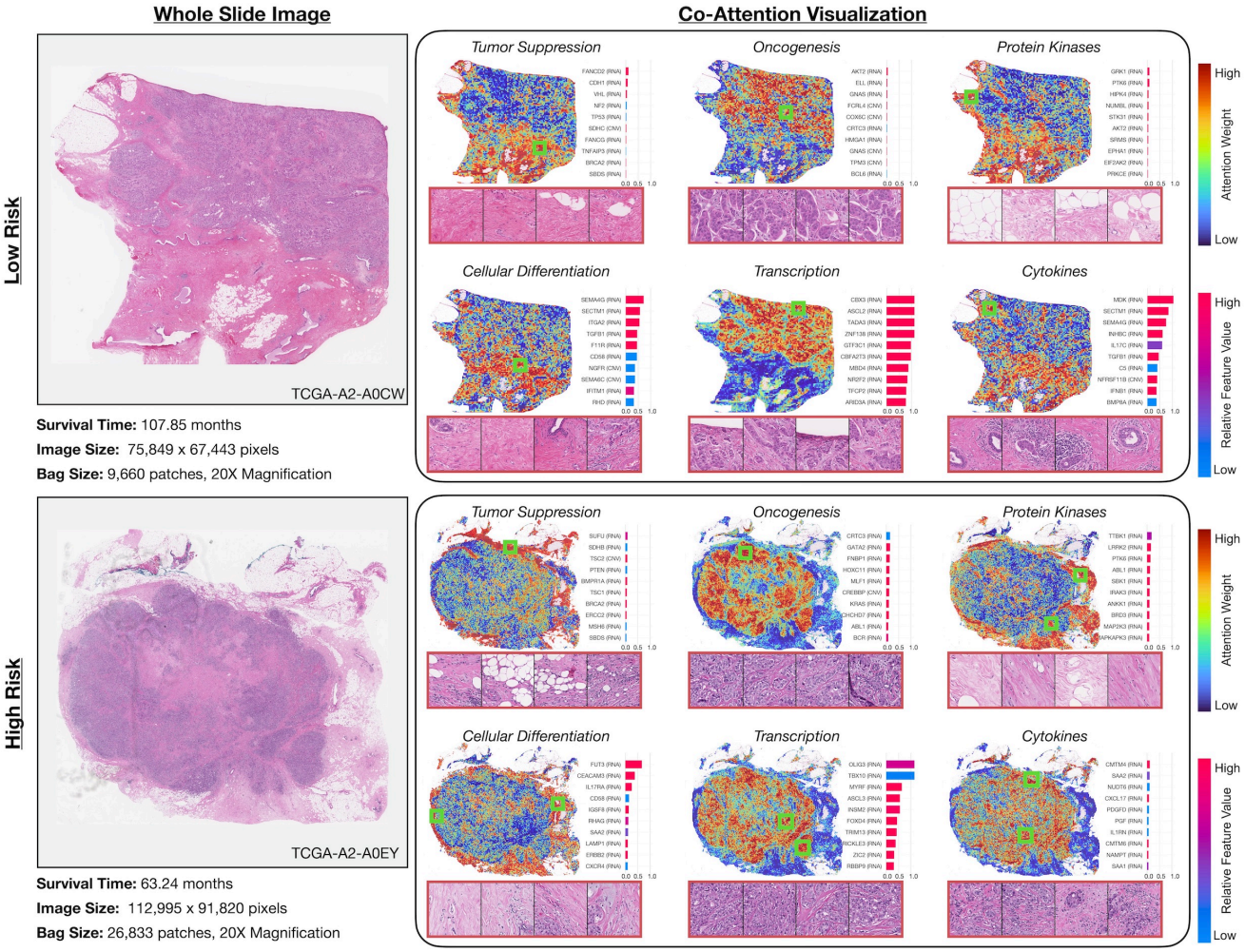

Figure 3: Co-attention visualization for low and high risk cases in BRCA, with corresponding high attention patches and high-attributed genes in each heatmap.

all cases, high attention in the cellular differentiation embedding focused on tumor-associated stroma, while both the tumor suppression and protein kinases embeddings had greater high attention given to stroma adjacent to adipocytes and glandular structures. These findings corroborate the attribution of genes such as FUT3 synthesized by epithelial cells, and TGFB1 used in regulating cell growth. In the oncogensis and transcription embedding, high attention weights localized image regions such as invasive, high-grade tumor morphology such as dense tumor cellularity and tumor-infiltrated stroma (opposite of tumor suppression). In the cytokines embedding, we observe that high attention regions were focused on immune cells infiltrating normal stroma and tumor cells. Additional visualizations can be found in the Supplementary Materials.

5. Conclusion

In this work, we present the Multimodal Co-Attention Transformer (MCAT) for survival outcome prediction in pathology. Our method formulates both gigapixel WSIs and genomic features as permutation-invariant sets, from which we develop more sophisticated feature aggregation strategies in MIL via transformer attention. A limitation in our current study is that we used a previously-curated gene set with potentially overlapping biological functional impact. Future work would focus on investigating early fusion of WSIs with more fine-grained, distinct biological gene sets, and further quantification of phenotype-genotype correspondences.

6. Acknowledgements

We thank Felix Yu and Quinn Zhen for their insightful feedback. This work was supported in part by internal funds from BWH Pathology, Nvidia GPU Grant Program, and NIGMS R35GM138216 (F.M.). R.J.C. and T.M. were supported by the NSF Graduate Fellowship.

References

[1] Khalid Abdul Jabbar, Shan E Ahmed Raza, Rachel Rosenthal, Mariam Jamal-Hanjani, Selvaraju Veeriah, Ayse Akarca, Tom Lund, David A Moore, Roberto Salgado, Maise Al Bakir, et al. Geospatial immune variability illuminates differential evolution of lung adenocarcinoma. *Nature Medicine*, pages 1–9, 2020.

[2] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):423–443, 2018.

[3] Peter Bandi, Oscar Geessink, Quirine Manson, Marcory Van Dijk, Maschenka Balkenhol, Meyke Hermsen, Babak Ehteshami Bejnordi, Byungjae Lee, Kyunghyun Paeng, Aoxiao Zhong, et al. From detection of individual metastases to classification of lymph node status at the patient level: the camelyon17 challenge. *IEEE Transactions on Medical Imaging*, 38(2):550–560, 2018.

[4] Babak Ehteshami Bejnordi, Mitko Veta, Paul Johannes Van Diest, Bram Van Ginneken, Nico Karssemeijer, Geert Litjens, Jeroen AWM Van Der Laak, Meyke Hermsen, Quirine F Manson, Maschenka Balkenhol, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA*, 318(22):2199–2210, 2017.

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

[6] Lyndon Chan, Mahdi S Hosseini, Corwyn Rowsell, Konstantinos N Plataniotis, and Savvas Damaskinos. Histosegnet: Semantic segmentation of histological tissue type in whole slide images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10662–10671, 2019.

[7] Hao Chen, Xiaojuan Qi, Lequan Yu, and Pheng-Ann Heng. Dcan: deep contour-aware networks for accurate gland segmentation. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 2487–2496, 2016.

[8] Richard Chen, Filip Jankovic, Nikki Marinsek, Luca Foschini, Lampros Kourtis, Alessio Signorini, Melissa Pugh, Jie Shen, Roy Yaari, Vera Maljkovic, et al. Developing measures of cognitive impairment in the real world from consumer-grade multimodal sensor streams. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2145–2155, 2019.

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

[10] Richard J Chen, Ming Y Lu, Drew FK Williamson, Tiffany Y Chen, Jana Lipkova, Muhammad Shaban, Maha Shady, Mane Williams, Bumjin Joo, Zahra Noor, et al. Pan-cancer integrative histology-genomic analysis via interpretable multimodal deep learning. *arXiv preprint arXiv:2108.02278*, 2021.

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

[12] David R Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202, 1972.

[13] David Roxbee Cox and David Oakes. *Analysis of survival data*, volume 21. CRC press, 1984.

[14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2018.

[15] James A. Diao, Jason K. Wang, Wan Fung Chui, Victoria Mountain, Sai Chowdary Gullapally, Ramprakash Srinivasan, Richard N. Mitchell, Benjamin Glass, Sara Hoffman, Sudha K. Rao, Chirag Maheshwari, Abhik Lahiri, Aaditya Prakash, Ryan McLoughlin, Jennifer K. Kerner, Murray B. Resnick, Michael C. Montalto, Aditya Khosla, Ilan N. Wapinski, Andrew H. Beck, Hunter L. Elliott, and Amaro Taylor-Weiner. Human-interpretable image features derived from densely mapped cancer pathology slides predict diverse molecular phenotypes. *Nature Communications*, 12(1), Mar. 2021.

[16] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. 2020.

[17] Harrison Edwards and Amos Storkey. Towards a neural statistician. *arXiv preprint arXiv:1606.02185*, 2016.

[18] Fabian B Fuchs, Daniel E Worrall, Volker Fischer, and Max Welling. Se (3)-transformers: 3d roto-translation equivariant attention networks. *Advances in Neural Information Processing Systems*, 2020.

[19] Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2016.

[20] Jevgenij Gamper and Nasir Rajpoot. Multiple instance captioning: Learning representations from histopathology textbooks and articles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

[21] Simon Graham, Quoc Dang Vu, Shan E Ahmed Raza, Ayesha Azam, Yee Wah Tsang, Jin Tae Kwak, and Nasir Rajpoot. Hover-net: Simultaneous segmentation and classi-

fication of nuclei in multi-tissue histology images. *Medical Image Analysis*, 58:101563, 2019.

[22] Mahdi S Hosseini, Lyndon Chan, Gabriel Tse, Michael Tang, Jun Deng, Sajad Norouzi, Corwyn Rowsell, Konstantinos N Plataniotis, and Savvas Damaskinos. Atlas of digital pathology: A generalized hierarchical histological tissue type-annotated database for deep learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11747–11756, 2019.

[23] Le Hou, Dimitris Samaras, Tahsin M Kurc, Yi Gao, James E Davis, and Joel H Saltz. Patch-based convolutional neural network for whole slide tissue image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2424–2433, 2016.

[24] Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *Proceedings of the 35th International Conference on Machine Learning*, pages 2132–2141, 2018.

[25] Hartland W Jackson, Jana R Fischer, Vito RT Zanotelli, H Raza Ali, Robert Mechera, Savas D Soysal, Holger Moch, Simone Muenst, Zsuzsanna Varga, Walter P Weber, et al. The single-cell pathology landscape of breast cancer. *Nature*, 578(7796):615–620, 2020.

[26] Shivam Kalra, Mohammed Adnan, Graham Taylor, and Hamid R Tizhoosh. Learning permutation invariant representations using memory networks. In *European Conference on Computer Vision*, pages 677–693. Springer, 2020.

[27] Jared L Katzman, Uri Shaham, Alexander Cloninger, Jonathan Bates, Tingting Jiang, and Yuval Kluger. Deepsurv: personalized treatment recommender system using a cox proportional hazards deep neural network. *BMC medical research methodology*, 18(1):1–12, 2018.

[28] Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. Bilinear attention networks. *Advances in Neural Information Processing Systems*, 2018.

[29] Jin-Hwa Kim, Kyoung Woon On, Woosang Lim, Jeonghee Kim, Jung-Woo Ha, and Byoung-Tak Zhang. Hadamard Product for Low-rank Bilinear Pooling. In *International Conference on Learning Representations*, 2017.

[30] Günter Klambauer, Thomas Unterthiner, Andreas Mayr, and Sepp Hochreiter. Self-normalizing neural networks. In *Advances in Neural Information Processing Systems*, pages 971–980, 2017.

[31] Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, pages 3744–3753. PMLR, 2019.

[32] Arthur Liberzon, Chet Birger, Helga Thorvaldsdóttir, Mahmoud Ghandi, Jill P Mesirov, and Pablo Tamayo. The molecular signatures database hallmark gene set collection. *Cell Systems*, 1(6):417–425, 2015.

[33] David N Louis, Arie Perry, Guido Reifenberger, Andreas Von Deimling, Dominique Figarella-Branger, Webster K Cavenee, Hiroko Ohgaki, Otmar D Wiestler, Paul Kleihues, and David W Ellison. The 2016 world health organization classification of tumors of the central nervous system: a summary. *Acta neuropathologica*, 131(6):803–820, 2016.

[34] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Hierarchical question-image co-attention for visual question answering. *Advances in Neural Information Processing Systems*, 2016.

[35] Ming Y Lu, Richard J Chen, Jingwen Wang, Debora Dillon, and Faisal Mahmood. Semi-supervised histology classification using deep multiple instance learning and contrastive predictive coding. *Advances in Neural Information Processing Systems (NeurIPS) Workshop in Machine Learning for Health*, 2019.

[36] Ming Y Lu, Tiffany Y Chen, Drew FK Williamson, Melissa Zhao, Maha Shady, Jana Lipkova, and Faisal Mahmood. Ai-based pathology predicts origins for cancers of unknown primary. *Nature*, 594(7861):106–110, 2021.

[37] Ming Y Lu, Dehan Kong, Jana Lipkova, Richard J Chen, Rajendra Singh, Drew FK Williamson, Tiffany Y Chen, and Faisal Mahmood. Federated learning for computational pathology on gigapixel whole slide images. *arXiv preprint arXiv:2009.10190*, 2020.

[38] Ming Y Lu, Drew FK Williamson, Tiffany Y Chen, Richard J Chen, Matteo Barbieri, and Faisal Mahmood. Data efficient and weakly supervised computational pathology on whole slide images. *Nature Biomedical Engineering*, 2020.

[39] Wenqi Lu, Simon Graham, Mohsin Bilal, Nasir Rajpoot, and Fayyaz Minhas. Capturing cellular topology in multi-gigapixel pathology images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 260–261, 2020.

[40] Faisal Mahmood, Daniel Borders, Richard J. Chen, Gregory N. Mckay, Kevan J. Salimian, Alexander Baras, and Nicholas J. Durr. Deep adversarial training for multi-organ nuclei segmentation in histopathology images. *IEEE Transactions on Medical Imaging*, 39(11), Nov. 2020.

[41] Faisal Mahmood, Ziyun Yang, Richard Chen, Daniel Borders, Wenhao Xu, and Nicholas J Durr. Polyp segmentation and classification using predicted depth from monocular endoscopy. In *Medical Imaging 2019: Computer-Aided Diagnosis*, volume 10950, page 1095011. International Society for Optics and Photonics, 2019.

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

[43] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. Multimodal deep learning. In *Proceedings of the 28th Tnternational Conference on Machine Learning*, pages 689–696, 2011.

[44] Duy-Kien Nguyen and Takayuki Okatani. Improved fusion of visual and language representations by dense symmetric co-attention for visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6087–6096, 2018.

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

[46] Prajit Ramachandran, Niki Parmar, Ashish Vaswani, Irwan Bello, Anselm Levskaya, and Jonathon Shlens. Stand-alone self-attention in vision models. *arXiv preprint arXiv:1906.05909*, 2019.

[47] Joel Saltz, Rajarsi Gupta, Le Hou, Tahsin Kurc, Pankaj Singh, Vu Nguyen, Dimitris Samaras, Kenneth R Shroyer, Tianhao Zhao, Rebecca Batiste, et al. Spatial organization and molecular correlation of tumor-infiltrating lymphocytes using deep learning on pathology images. *Cell reports*, 23(1):181–193, 2018.

[48] Muhammad Shaban, Syed Ali Khurram, Muhammad Moazam Fraz, Najah Alsubaie, Iqra Masood, Sajid Mushtaq, Mariam Hassan, Asif Loya, and Nasir M Rajpoot. A novel digital score for abundance of tumour infiltrating lymphocytes predicts disease free survival in oral squamous cell carcinoma. *Scientific Reports*, 9(1):1–13, 2019.

[49] Aravind Subramanian, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, Scott L Pomeroy, Todd R Golub, Eric S Lander, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550, 2005.

[50] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. *Proceedings of the 34th International Conference on Machine Learning*, 2017.

[51] David Tellez, Geert Litjens, Jeroen van der Laak, and Francesco Ciompi. Neural image compression for gigapixel histopathology image analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[52] Jeroen van der Laak, Francesco Ciompi, and Geert Litjens. No pixel-level annotations needed. *Nature biomedical engineering*, 3(11):855–856, 2019.

[53] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017.

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

[55] Apoorv Vyas, Angelos Katharopoulos, and François Fleuret. Fast transformers with clustered attention. *Advances in Neural Information Processing Systems*, 33, 2020.

[56] Sinong Wang, Belinda Li, Madian Khabsa, Han Fang, and Hao Ma. Linformer: Self-attention with linear complexity. *Advances in Neural Information Processing Systems*, 2020.

[57] Wei-Hung Weng, Yuannan Cai, Angela Lin, Fraser Tan, and Po-Hsuan Cameron Chen. Multimodal multitask representation learning for pathology biobank metadata prediction. *Advances in Neural Information Processing Systems ML4H Workshop*, 2019.

[58] Wing Hung Wong et al. Theory of partial likelihood. *The Annals of statistics*, 14(1):88–123, 1986.

[59] Ellery Wulczyn, David F Steiner, Zhaoyang Xu, Apaar Sadhwani, Hongwu Wang, Isabelle Flament-Auvigne, Craig H Mermel, Po-Hsuan Cameron Chen, Yun Liu, and Martin C Stumpe. Deep learning-based survival prediction for multiple cancer types using histopathology images. *PLoS One*, 15(6):e0233678, 2020.

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

[61] Jiawen Yao, Xinliang Zhu, and Junzhou Huang. Deep multi-instance learning for survival prediction from whole slide images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 496–504. Springer, 2019.

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

[63] Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. Deep modular co-attention networks for visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6281–6290, 2019.

[64] Zhou Yu, Jun Yu, Jianping Fan, and Dacheng Tao. Multi-modal factorized bilinear pooling with co-attention learning for visual question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1821–1830, 2017.

[65] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Tensor fusion network for multimodal sentiment analysis. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2017.

[66] Shekoufeh Gorgi Zadeh and Matthias Schmid. Bias in cross-entropy-based training of deep survival networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[67] Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Ruslan Salakhutdinov, and Alexander Smola. Deep sets. *Advances in Neural Information Processing Systems*, 2017.

[68] Xiaofan Zhang, Hai Su, Lin Yang, and Shaoting Zhang. Fine-grained histopathological image analysis via robust segmentation and large-scale retrieval. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5361–5368, 2015.

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

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