Hybrid Graph Convolutional Network With Online Masked Autoencoder for Robust Multimodal Cancer Survival Prediction

Wentai Hou, Chengxuan Lin, Lequan Yu, *Member, IEEE*, Jing Qin, *Senior Member, IEEE*, Rongshan Yu, *Senior Member, IEEE*, and Liansheng Wang, *Member, IEEE*

***Abstract*— Cancer survival prediction requires exploiting related multimodal information (*e.g.*, pathological, clinical and genomic features, *etc.*) and it is even more challenging in clinical practices due to the incompleteness of patient's multimodal data. Furthermore, existing methods lack sufficient intra- and inter-modal interactions, and suffer from significant performance degradation caused by missing modalities. This manuscript proposes a novel hybrid graph convolutional network, entitled HGCN, which is equipped with an online masked autoencoder paradigm for robust multimodal cancer survival prediction. Particularly, we pioneer modeling the patient's multimodal data into flexible and interpretable multimodal graphs with modality-specific preprocessing. HGCN integrates the advantages of graph convolutional networks (GCNs) and a hypergraph convolutional network (HCN) through node message passing and a hyperedge mixing mechanism to facilitate intra-modal and inter-modal interactions between multimodal graphs. With HGCN, the potential for multimodal data to create more reliable predictions of patient's survival risk is dramatically increased compared to prior methods. Most importantly, to compensate for missing patient modalities in clinical scenarios, we incorporated an online masked autoencoder paradigm into HGCN, which can effectively capture intrinsic dependence between modalities and seamlessly generate missing hyperedges for model inference. Extensive experiments and analysis on six cancer cohorts from TCGA show that our method significantly outperforms the state-of-the-arts in both complete and missing modal settings. Our codes are made available at https://github.com/lin-lcx/HGCN.**

***Index Terms*— Survival prediction, multi-modal learning, graph convolutional network, hypergraph convolutional network, masked autoencoder, decision fusion.**

Manuscript received 8 December 2022; revised 20 February 2023; accepted 3 March 2023. Date of publication 6 March 2023; date of current version 1 August 2023. This work was supported in part by the Ministry of Science and Technology, China, under Grant 2021ZD0201900 and Grant 2021ZD0201903; in part by the Project of Strategic Importance of The Hong Kong Polytechnic University under Project 1-ZE2Q; and in part by the Hong Kong Research Grants Council under Grant T45-401/22-N. *(Wentai Hou and Chengxuan Lin contributed equally to this work.) (Corresponding author: Liansheng Wang.)*

Wentai Hou is with the Department of Information and Communication Engineering, School of Informatics, Xiamen University, Xiamen 361005, China, and also with the National Institute for Data Science in Health and Medicine, Xiamen University, Xiamen 361102, China (e-mail: houwt@stu.xmu.edu.cn).

Chengxuan Lin, Rongshan Yu, and Liansheng Wang are with the Department of Computer Science, School of Informatics, Xiamen University, Xiamen 361005, China (e-mail: lincx@stu.xmu.edu.cn; rsyu@xmu.edu.cn; lswang@xmu.edu.cn).

Lequan Yu is with the Department of Statistics and Actuarial Science, The University of Hong Kong, Hong Kong, SAR, China (e-mail: lqyu@hku.hk).

Jing Qin is with the Center for Smart Health, School of Nursing, The Hong Kong Polytechnic University, Hong Kong, SAR, China (e-mail: harry.qin@polyu.edu.hk).

Digital Object Identifier 10.1109/TMI.2023.3253760

I. Introduction

THE ability to predict the survival risk of patients with cancer has the potential to assist clinical management decisions in treatment and monitoring [1]. The Cox proportional hazards model [2] (hereafter called "Cox model") is the most common model for cancer survival prediction. Previously, demographic and clinical covariates were combined using the Cox model to predict patient survival [3]. However, this method does not comprehensively characterize patients' disease pathology and the genetic factors at play, thus limiting the efficacy of survival prediction [3], [4], [5].

In recent years, the development of medical imaging technology [6], [7] and advanced genomic methods [8], [9] has brought survival prediction research into a new multimodal era. Survival prediction analysis is significantly different from cancer diagnosis and staging tasks, as it is a future state prediction task dependent on multiple factors. Furthermore, high-quality multimodal medical data that can be used for model development is relatively rare. Therefore, it is important to provide more detailed evidence-based support for survival prediction through exploration of intra- and inter-modal relationships. A number of feature-based fusion techniques, such as GSCNN [10], MultiSurv [11], Pathomic Fusion [12], metric learning [13] and MCAT [14], have been developed to fuse the feature representations of different modalities. However, most of the above methods only focus on enhancing the feature representation of each modality and adopting relatively simple fusion scheme (*i.e.*, concatenation [10], row-wise maximum [11], Outer product [12], similarity constraint [13], co-attentive fusion [14]), yielding insufficient excavating of the intra- and inter-modal interactions, thus limiting the performance of clinical output prediction.

Although multimodal data has the potential to more comprehensively reflect the health status of patients, as shown

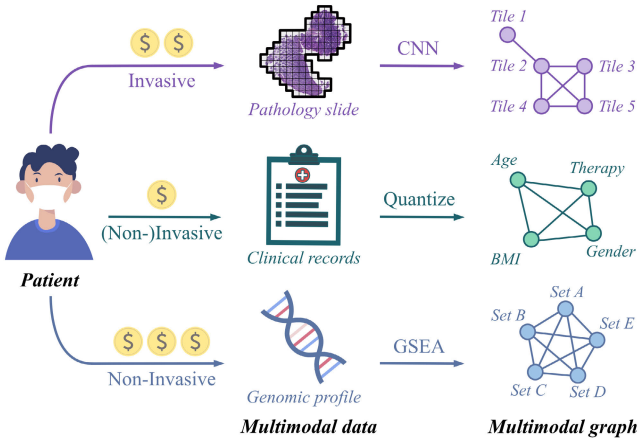

Fig. 1. Acquisition cost comparison and graph representation of a patient's multimodal data. Note that actual costs vary across providers, countries, and specific examination parameters. To improve visual clarity, all nodes are not depicted in the graphic above. CNN: Convolutional Neural Network. GSEA: Gene Set Enrichment Analysis.

in Fig. 1, costs and risks of acquiring these different modalities are significant. Therefore, considering the economic and physical condition of the patient, missing modalities have the potential to be common in the clinic, which leads to current methods becoming obsolete or invalid [15]. To this end, developing a multimodal cancer survival prediction model that is robust to account for unexpected or missing modalities during inference is of great significance and has the potential to reduce patient and doctor burden [16], [17], [18]. Some previously developed methods to tackle data incompleteness include: zero padding [19], [20], learning factorized representation [21] and autoencoder [22], [23]. However, these approaches are difficult to be adopted to multimodal cancer survival prediction due to several reasons. First, medical data dimensions are extremely varied. For example, clinical records are multidimensional text data, pathology slides are comprised of gigapixel images, while genomic profiles are comprised of ten thousand dimensional sequences. It is relatively non-trivial to effectively align multimodal features with unbalanced dimensions. Furthermore, compared with conventional visual-audio data, attributing differences between medical data sets is more difficult, which makes it difficult to use a traditional generative model to reconstruct missing feature representations.

In this manuscript, we proposed a novel **H**ybrid **G**raph **C**onvolutional **N**etwork (HGCN) equipped with an online masked autoencoder for multimodal cancer survival prediction, which can effectively exploit complementary information from different modalities and is robust in missing modality scenarios. The key idea of our framework was to model the multimodal medical data as full graph structures to facilitate inter- and intra-modal interactions, as well as construct a masked signal model to simulate missing modality scenarios to support modal interactions during inference. Specifically, as shown in Fig. 1, we modeled the patient's multimodal medical data as flexible graph structures through convolution neural network (CNN) [24], quantization, and gene set enrichment analysis (GSEA) [25]. We then adopted independent graph convolutional networks (GCNs) [26] to enable the messaging between adjacent nodes, realizing the intra-modal interaction for individual modalities. Furthermore, we designed a novel hypergraph convolutional network (HCN) [27] with a hyperedge mixing mechanism to mix up the high-order features (*i.e.*, hyperedges) of multimodal graphs, yielding inter-modal interactions for different modality data. More importantly, to facilitate the model inference under missing modalities, we incorporated an online masked autoencoder (MAE) paradigm into the HGCN, which has the capacity to learn intrinsic dependence between multiple modalities via transformer, and generates the missing hyperedges in a encoder-decoder manner. Final survival prediction results are then obtained by merging the output decisions of available modalities.

The main contributions of this work can be summarized as follows.

- We pioneer usage of graph structures for explicitly modeling patient's multimodal data for survival prediction problem in which factors and internal associations for each modality can be retained. This results in more flexible processing patterns and sufficient interpretability.
- We propose a novel hybrid graph convolution network, which consist of identical graph convolutional layers and an elaborately designed hypergraph convolutional layer with hyperedge mixing module, to better conduct inter- and intra-modal interactions of multimodal graphs.
- We design an online masked autoencoder paradigm to deal with the missing modality scenarios. This paradigm skillfully utilizes intrinsic dependence of multiple modalities learned via transformer, and thus generate the missing hyperedges during model inference.
- Extensive experiments with promising results on six public cancer cohort from TCGA validate the effectiveness and robustness of the proposed method for cancer survival prediction. Also, multiple visualizations and interpretable analysis are conducted to potentially aid clinical management and biomarkers discovery.

II. Related Works

A. Multimodal Cancer Survival Prediction

During the last several decades, various clinicians have made clinical cancer survival prediction predictions based on clinical covariates and experience [2]. With the advent of computer vision [28] and medical imaging (*e.g.*, pathology [29], CT [30], MRI [31], *etc.*), clinicians understand the importance of using AI technology as a decision support tool for cancer survival prediction. In recent years, to further improve the performance of AI-based survival prediction, some studies [10], [11], [12], [13], [14] attempted to combine clinical records and genomic profile data with medical imaging data, which indicated a shift in survival prediction research to a multimodal era. Overall, most of these studies focused on multimodal medical data modeling and feature fusion. For example, Mobadersany, et al., proposed a GSCNN architecture to combine histology images and genomic data into fully connected layers to predict patient's survival [10].

Vale, et al., proposed a MultiSurv that employs dedicated submodels to represent each modality as a one-dimensional feature, and conducts a row-wise maximum on each representation to fuse features together and predict survival outcome [11]. Chen, et al., proposed a Pathomic Fusion model that turns pathological images and gene data into graph data and one-dimensional vectors [12]. Feature representations are then obtained via graph neural network (GNN) and deep neural network (DNN) [12]. Then, Outer products are used to fuse these feature representations for cancer prognosis and prediction [12]. Cheerla, et al., proposed a similarity loss for the purpose of multimodal joint learning through constraining the feature of different modes in an embedded space [13]. Inspired by approaches in visual question answering, Chen, et al., presented a multimodal co-attention transformer (MCAT) framework that learns how histology patches attend to genes when predicting patient survival [14]. However novel, these methods have limited clinical application value because they ignore missing modality scenarios caused by patients who cannot physically undergo or pay for multi-omics examinations.

B. Multimodal Learning With Missing Modalities

Although the presence of multiple modalities provides valuable information, a key challenge for multimodal data learning is that models must be robust enough to respond to unexpected or missing modalities during the testing process. Padding feature representations of the missing modalities with the number zero is a widely used method to cope with incomplete modalities [19], [20]. Although this method can help multimodal models inference, it cannot alleviate information loss caused by missing modalities. To solve this problem, some researchers have adopted generative models. For example, Tsai, et al., proposed a multimodal factorized model (MFM) to learn factorized representations so that missing modalities can be reconstructed for final decision making [21]. Moreover, recent multimodal learning studies employ an autoencoder, which is a general architecture that has the capacity to learn feature representations in an unsupervised manner, and thus is capable of complementing the feature representation of missing modalities [22], [23]. However, due to significant differences in medical data attributes and differences in dimension, application of these methods to models predicting cancer survival is difficult to achieve.

C. Masked Autoencoder

"Masked signal modeling" is a task of self-supervised learning [32] that masks a portion of each input signal and tries to predict these masked signals through the use of visible evidence [33]. In the natural language processing field, masked signal modeling based models (*i.e.*, BERT [34] and its improvements [35], [36]) have spurred a rapid succession of contextual language representations. Recently, He, et al., [37] and Xie, et al., [33] successfully applied this philosophy to the field of computer vision field, and proposed a masked autoencoder (MAE) for image representation learning. However, MAE was regarded as a pretext task and trained independently. Additionally, after pre-training, the MAE decoder is discarded and does not participate in downstream task training. Therefore, we wanted to ask the question: *whether MAE has broader application prospects?* In this manuscript, we propose a new MAE training paradigm, and show how it can be applied to the field of multimodal learning with missing modalities.

III. Methodology

The proposed multimodal learning framework contains two characteristics: (1) learning complementary intra- modal and inter-modal interactions for predicting patients' survival risk and (2) dealing with unexpected or missing modalities during model inference. Fig. 1 and Fig. 2 illustrate input data processing and the forward flow of the proposed framework, respectively. First, the patients' multimodal medical data was modeled in a flexible graph structure for further analysis, as shown in Fig. 1. This graph is then adopted into independent graph convolutions to analyze intra-modal interaction. Subsequently, we extracted higher-order representations of modalities as hyperedges and mixed them to analyze inter-modal interaction. For missing modalities, we designed a masked autoencoder to generate missing hyperedges for intra- and inter-modal interactions. Knowledge generated by intra- and inter-model interactions were combined to assist decision-making of each modality and produce a set of final survival prediction results from these decisions. Future subsections will further detail the construction of multimodal graphs, hybrid graph convolutional network, online masked autoencoder paradigm, learning strategy, and the model inference.

A. Construction of Multimodal Graphs

For the purpose of this study, we consider that multimodal data of a batch of patients (batch size is B) has $M = \{p, c, g\}$ modalities, (*i.e.*, pathological slide, clinical records and genomic profile), which comprehensively depict the patient's physical conditions from body level to a molecule level. Here, we denote the survival of i-th patient as (t_i, δ_i), where t_i is the observed time and $\delta_i \in \{0, 1\}$ is the censorship status.

As shown in Fig. 1, pathological slides are cropped into non-overlapping 512×512 tiles utilizing a sliding window strategy. Subsequently, each tile is entered into KimiaNet [24], a pretrained CNN, which extracts a 1024-dimensional vectorized representation for each tile. According to the spatial coordinates of each tile, these vectorized representations were connected in a 8-adjacent manner so that the graph representation of pathological slide can be obtained and denoted as $\mathcal{G}_p = \{\mathbf{V}_p, \mathbf{A}_p\}$. For clinical records (*i.e.*, age, gender, BMI, *etc.*), we quantified each record and then obtained its vector representation through one-hot coding. Through fully connecting each one-hot vector, the graph representation of clinical records can be obtained and denoted as $\mathcal{G}_c = \{\mathbf{V}_c, \mathbf{A}_c\}$. For genomic profile, we employ GSEA to generate five genomic embeddings: 1) Tumor Supression, 2) Oncogenesis, 3) Protein Kinases, 4) Cellular Differentiation, and 5) Cytokines and Growth. These five genomic embedding are connected to each other, resulting in the graph representation of genomic profile, which were denoted as $\mathcal{G}_g = \{\mathbf{V}_g, \mathbf{A}_g\}$. Note that $\mathbf{V}_p$, $\mathbf{V}_c$,

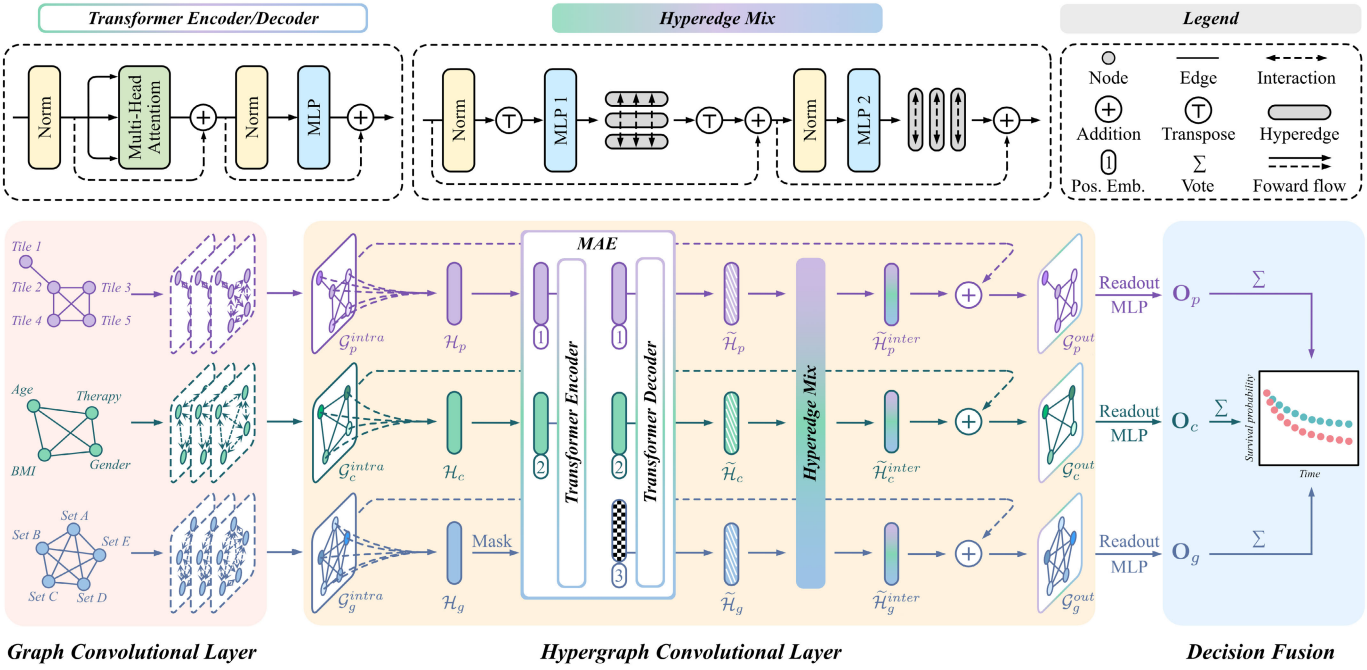

Fig. 2. Illustration of the proposed hybrid graph convolutional network (HGCN) with online masked autoencoder (MAE). Multimodal graphs are fist fed into the GCN Layers for intra-modal interaction and then fed into our designed hypergraph convolutional layer with hyperedge mixing module for inter-modal interaction. The online MAE are jointly trained with HGCN, which continuously simulates the different scenarios of missing modalities (such as genomic profile shown in the figure) through the random mask operation. After intra- and inter-modal interactions, the decisions of three modalities are fused to obtain the final survival prediction result. In the inference, the trained MAE can generate the hyperedge information of missing modalities to assist modal interaction and decision-making. Note that not all nodes are shown for visual clarity.

$\mathbf{V}_g$ respectively represent node feature sets, and $\mathbf{A}_p$, $\mathbf{A}_c$, $\mathbf{A}_g$ represent the adjacent matrix of each modality. Moreover, all node features were aligned to 1024 dimensions via zero padding to facilitate subsequent analysis.

Representation of patients' multimodal data in the form of a full graph has the following advantages. (I) *More flexibly represents the whole slide images:* Graph-based methods treat the WSIs as point-cloud structure data [38], [39]. This more accurately and compactly depicts pathological tissue topology, which not only reduces the amount of computation needed, but also increase the potential to investigate complex pathological structures (such as feature-based KNN, dynamic KNN). (II) *More effectively mines the tissue-level context:* Several studies shown that cancer survival prediction requires considering both local- and global-level interactions in the tumor and surrounding tissues [40], [41]. Compared with CNN based methods, GCN has fewer parameters and pays more attention to the inter-node interactions within a region, which is more in line with the intrinsic requirements of survival prediction. (III) *Handle missing features of tabular data:* In clinical cohorts, information missing in tabular data (*i.e.*, clinical records) is common due to various reasons such as sensor damage, data corruption, and human mistakes in recording. Benefiting from flexibility of graph representation and inductive graph convolution, patients with missing information can still be included in these studies without affecting network training and testing, which is difficult to achieve for fully connected layers. (IV) *Enhances the interpretability of tabular data:* Computer aided diagnosis and prognosis requires not only high performance, but also strong rationale to deliver judgments. In our framework, different types of information (*e.g.*, age, gender, BMI, *etc.*) in the tabular data are decoupled and represented as different nodes, which allows the information to be more easily analyzed via the attention score. Moreover, this method allows for the clinician to refine the graph data based on prior knowledge, potentially improving treatment decisions.

B. Hybrid Graph Neural Network

1) Intra-Modal Interaction With GCN Layer: The goal of intra-modal interaction is to mine modality-specific information for decision-making. As shown in Fig. 2, three identical GCN layers were adopted to implement message passing for graphic representation of each modality. Specifically, the intra-modal interaction for any modality $m \in \{p, c, g\}$ can be implemented as:

$$\mathcal{G}_m^{intra} = \sigma(\mathrm{GCN}(\mathcal{G}_m^{in})), \tag{1}$$

where $\mathcal{G}_m^{in}$ represents the input graph representation of m and $\mathcal{G}_m^{intra}$ represents the generated graph representation of each modality. It should be pointed out that all graph convolutions with generalizability to unknown graphs (*i.e.*, inductive graph convolutions) can be used to explore the inter-modal interaction of multimodal graphs. In this study, the standard inductive graph convolution was employed due to the low complexity of node features and intrinsic real-time requirements of clinical survival prediction (*i.e.*, GraphSAGE [42]), which allows for message passing as well as maps the channel of each node to a hidden dimension C. C is a hyperparameter, which is set to 512 by default. $\sigma(\cdot)$ represents the activation function, such

as ReLU. By using GCNs, the factors of each modality can communicate with each other and modality representation for intra-modal interactions are subsequently saved in $\mathcal{G}_m^{intra}$.

2) Hyperedge Generation: Due to the high variety and many dimensions of the medical data, establishing pairwise mapping for all factors is computationally intractable. In recent years, a generalized variant of ordinary graphs, *i.e.*, hypergraph, has been proposed to describe the relationship between an edge containing multiple nodes [27], [43]. Based on those previous studies, these studies employed a hypergraph method (after the intra-modal interaction) in which all nodes belonging to the same modality are linked by a hyperedge. Then, the high-order representation (*i.e.*, hyperedge) of each modality was extracted via attention pooling operation. The hyperedge of the modality m with N nodes is calculated as:

$$\mathcal{H}_m = \sum_{n=1}^{N} \text{Softmax}\left(\text{MLP}_1\left(\mathbf{V}_m^n\right)\right) \odot \mathbf{V}_m^n, \tag{2}$$

where $\mathbf{V}_m^n$ is the feature of node n in $\mathcal{G}_m^{intra}$. MLP is a Multilayer Perceptron to calculate the attention scores for $\mathbf{V}_m$. Therefore, $\mathcal{H}_m \in \mathbb{R}^{B\times 1\times C}$ and we consider $\mathcal{H}_m$ as the information carrier for the inter-modal interaction.

3) Inter-Modal Interaction With Hyperedge Mix: To promote inter-modal interaction, it is crucial to establish multi-view communication between hyperedges. However, a conventional hypergraph neural network cannot meet this requirement. Therefore, we designed a hyperedge mix module within the HCN. Based on a prior study [44], we adopted two MLP layers to mix the hyperedge from the token wise and the channel wise to promote an all-rounded inter-modal interaction between modalities. Specifically, we denote the set of original hyperedges as $\mathcal{H} \in \mathbb{R}^{B\times M\times C}$. This process can be written as follows:

$$\begin{aligned}\mathcal{H}^{inter} &:= \mathcal{H}^{\top} + \text{MLP}_2((\text{LayerNorm}(\mathcal{H}))^{\top}) \\ \mathcal{H}^{inter} &:= (\mathcal{H}^{inter})^{\top} + \text{MLP}_3(\text{LayerNorm}((\mathcal{H}^{inter})^{\top})),\end{aligned} \tag{3}$$

in which each MLP contains two fully-connected layers. By using hyperedge mix, the hyperedges in question receive sufficient communication between each other, thus ensuring that mixed hyperedges $\mathcal{H}^{inter}$ contains the information of multiple modalities after the inter-modal interaction.

As shown in Fig. 2, following the traditional hypergraph convolutional network [27], each hyperedge $\mathcal{H}_m^{inter} \in \mathbb{R}^{B\times 1\times C}$ is further scattered and added to the corresponding graph $\mathcal{G}_m^{intra} \in \mathbb{R}^{B\times N\times C}$, in order to obtain the output multimodal graphs $\mathcal{G}^{out}$. This process can be written as:

$$\mathcal{G}_m^{out} = \mathcal{G}_m^{intra} \oplus \mathcal{H}_m^{inter}, \tag{4}$$

where $\oplus$ denotes the broadcast addition, and $\mathcal{G}_m^{out}$ contains specific information of modality m and high-order interaction information with other modalities.

C. Online Masked Autoencoder

To deal with the incomplete hyperedges caused by the missing modalities during model inference, we design a masked autoencoder module that can reconstruct the original hyperedges given its partial observation. This study is inspired by MAE [37], an image representation method that first masks random patches of the input image and then encourages the model to reconstruct those missing pixels. The differences between the two works are: (I) The reconstruction target in this study are hyperedge features, rather than the pixel values of the image. (II) The reconstruction task in these studies are trained in an end-to-end manner simultaneously with cancer survival prediction and not served as a pretext task. These studies assume that one or more medical examinations are missing due to financial burden or physical condition of the patient. During the training process, we ***randomly*** sampled a subset (one or two) of modalities and masked the corresponding hyperedges, resulting in a partially masked set of hyperedges. As shown in Fig. 2, the partially masked set of hyperedges with the positional embedding was fed into a Transformer [45] based *asymmetric* autoencoder [37] to reconstruct the missing hyperedges according to the semantic of available hyperedges and contextual relations. Specifically, the Transformer encoder was only applied on *unmasked hyperedges*. The Transformer decoder is applied on the full set of tokens consisting of (I) encoded available hyperedges, and (II) a series of shared and learned vectors (*i.e.*, mask tokens) that indicate the presence of missing hyperedges to be reconstructed. Denote the reconstructed hyperedge set to $\widetilde{\mathcal{H}} \in \mathbb{R}^{B\times M\times C}$, the loss function of the masked autoencoder can be written as:

$$\mathcal{L}_{MAE} = \frac{1}{BMC} \sum_{i=1}^{B} \sum_{m\in[M]} \sum_{k=1}^{C} (\mathcal{H}_{i,m,k} - \widetilde{\mathcal{H}}_{i,m,k})^2. \tag{5}$$

In equation 5, $[M]$ is defined as an uncertain value, which denotes the set missing modalities. Similar with BERT [46] and original MAE [37], only the reconstruction loss of masked hyperedges was calculated. It should be noted that the generated hyperedges of available modalities were also adopted for subsequent analysis, so that all hyperedges maintain the same numerical distribution before inter-modal interaction.

D. Training and Inference Strategy

The survival output of modality m can be calculated as:

$$\mathbf{O}_m = \text{MLP}_4(\text{Readout}(\mathcal{G}_m^{output})). \tag{6}$$

Therefore, the Cox loss [47] of modality m is calculated by:

$$\mathcal{L}_{Cox}^m = \sum_{i=1}^{B} \delta_i \left(-\mathbf{O}_m(i) + \log \sum_{j:t_j>=t_i} \exp\left(\mathbf{O}_m(j)\right) \right), \tag{7}$$

where $\mathbf{O}_m(i)$ and $\mathbf{O}_m(j)$ denote the survival output of modality m of i-th and j-th patient, respectively. In summary, the total loss of the whole framework is defined as:

$$\mathcal{L}_{Total} = \epsilon_o \mathcal{L}_{MAE} + \sum_{m\in M} \epsilon_m \mathcal{L}_{Cox}^m, \tag{8}$$

where ϵ_o and ϵ_m are the trade-off parameters. Considering the feature complexity of each modality, the ϵ_o and ϵ_m is set to 5 and 1, respectively. The batch size B is set to 32 and the Adam optimizer with learning rate of $3e-5$ is used to optimize the whole framework. The architecture of the proposed framework is shown in TABLE I.

Algorithm 1 Model Inference

Input: Number of available modalities M', $M' \leq M$;
Available multimodal graphs $\{\mathcal{G}_1, \ldots, \mathcal{G}_m\}$, $m \in M'$.
Output: Final survival prediction $\mathbf{O}_{final}$.
Initialization: Trained HGCN with online MAE.

 # Compute intra-modal interactions of available graphs.
for m in M' **do**:
  $\mathcal{G}_m^{intra} \leftarrow \sigma(\text{GraphSAGE}(\mathcal{G}_m^{in}))$.
end for
 # Generate hyperedges of available graphs.
for m in M' **do**:
  $\mathcal{H}_m \leftarrow \sum_{n=1}^{N} \text{Softmax}\left(\text{MLP}_1\left(\mathbf{V}_m^n\right)\right) \odot \mathbf{V}_m^n$.
end for
 # Generate complete hyperedges by trained MAE.
$\widetilde{\mathcal{H}} \leftarrow \text{MAE}(\mathcal{H})$
 # Compute complete inter-modal interactions by hyperedge mix.
$\widetilde{\mathcal{H}}^{inter} \leftarrow \widetilde{\mathcal{H}}^{\top} + \text{MLP}_2((\text{LayerNorm}(\widetilde{\mathcal{H}}))^{\top})$
$\widetilde{\mathcal{H}}^{inter} \leftarrow (\widetilde{\mathcal{H}}^{inter})^{\top} + \text{MLP}_3(\text{LayerNorm}((\widetilde{\mathcal{H}}^{inter})^{\top}))$
 # Combine available intra-modal information with corresponding complete inter-modal information.
for m in M' **do**:
  $\mathcal{G}_m^{out} \leftarrow \mathcal{G}_m^{intra} \oplus \widetilde{\mathcal{H}}_m^{inter}$
end for
 # Compute the predicted survival of available modalities.
for m in M' **do**:
  $\mathbf{O}_m \leftarrow \text{MLP}_4(\text{Readout}(\mathcal{G}_m^{output}))$
end for
 # Decision fusion.
$\mathbf{O}_{final} \leftarrow \text{Mean}(\mathbf{O}_1, \ldots, \mathbf{O}_m)$

TABLE I
ARCHITECTURE OF PROPOSED FRAMEWORK. M' DENOTES THE NUMBER OF AVAILABLE MODALITIES. [N] IS AN UNCERTAIN VALUE FOR EACH SAMPLE, WHICH DENOTE THE NODE NUMBER OF AVAILABLE MODALITIES. FC: FULLY-CONNECTED LAYER

Module	Input size	Output size
GCN Layer	$M' \times [N] \times 1024$	$M' \times [N] \times 512$
Hyperedge gen.	$M' \times [N] \times 512$	$M' \times 1 \times 512$
MAE	$M' \times 1 \times 512$	$M \times 1 \times 512$
Hyperedge mix	$M \times 1 \times 512$	$M \times 1 \times 512$
Broadcast add	$M' \times [N] \times 512$, $M' \times 1 \times 512$	$M' \times [N] \times 512$
Readout	$M' \times [N] \times 512$	$M' \times 1 \times 512$
$\text{MLP}_4(\text{FC}_1)$	$M' \times 1 \times 512$	$M' \times 1 \times 128$
$\text{MLP}_4(\text{FC}_2)$	$M' \times 1 \times 128$	$M' \times 1 \times 1$
Decision fusion	$M' \times 1 \times 1$	$1(\text{O}_{final})$

During inference procedure, the final survival prediction is obtained by fusing all modality decisions. Moreover, benefiting from the online MAE paradigm, our model still retained the ability to perform intra- and inter-modal interaction even when some modalities were missing. **Algorithm 1** provide the pseudo-code of the inference process of our model.

IV. EXPERIMENTS

A. Experimental Setting

1) Datasets: To validate the proposed method, six cancer cohorts from The Cancer Genome Atlas (TCGA) were used. TCGA is a public cancer data consortium that contains matched clinical records, diagnostic WSIs and genomic data with labeled survival times and censorship statuses. For this study, we used the following cancer cohorts: **Kidney Clear Cell Carcinoma (KIRC)** (385 cases), **Liver Hepatocellular Carcinoma (LIHC)** (287 cases), **Esophageal Carcinoma (ESCA)** (153 cases), **Lung Squamous Cell Carcinoma (LUSC)** (438 cases), **Lung Adenocarcinoma (LUAD)** (452 cases), and **Uterine Corpus Endometrial Carcinoma (UCEC)** (387 cases).

TABLE II
CLINICAL RECORDS INCLUDED IN EACH CANCER COHORTS

Cohorts	Clinical records
KIRC	Race, Age at index, Gender, Prior malignancy, Pack years smoked, Years smoked, Radiation Therapy, Pharmaceutical Therapy
LIHC	Race, Age at index, Gender, BMI, Radiation Therapy, Pharmaceutical Therapy
ESCA	Race, Age at index, Gender, Alcohol history, Primary diagnosis, Site of resection or biopsy, Morphology, BMI, Cigarettes per day, Radiation Therapy, Pharmaceutical Therapy
LUSC	Race, Age at index, Gender, Prior malignancy, Site of resection, Pack years smoked, Years smoked, Radiation Therapy, Pharmaceutical Therapy
LUAD	Race, Age at index, Gender, Morphology, Prior malignancy, Site of resection or biopsy, Radiation Therapy, Pharmaceutical Therapy
UCEC	Race, Age at index, Primary diagnosis, Morphology, Radiation Therapy, Pharmaceutical Therapy

Note that because the nature and features of different cancer cohorts are different (*e.g.*, only women suffer from UCEC), the clinical records included in each cohort is inconsistent, which can be found in TABLE II. All pathological slides were normalized to the magnification of $\times 10$, with an average size of approximately 20840 $\times$ 15606. The dimension of original genomic profile data is 20472, which is screened and split by GSEA into 5 sets containing 1903 variables, including Tumor Supression, 2) Oncogenesis, 3) Protein Kinases, 4) Cellular Differentiation, and 5) Cytokines and Growth.

2) Evaluation Metric and Procedure: First, concordance index (CI) was used as the primary evaluation metric [48]. CI measures the fraction of all pairs of patients whose survival risks are correctly ordered. CI ranges from 0 to 1, where a larger CI indicates better performance. Second, Brier Score (BS) was used to evaluate the absolute prediction, which represents the average squared error between the observed survival status and the predicted survival probability [48]. BS is a value between 0 and 1, with 0 being the best possible value. Third, to evaluate the ability of patients stratification, Kaplan-Meier (KM) analysis was used. For each cohort, patients were stratified into high-risk and low-risk groups by the median score output using predictive models. A lower p-value using Logrank test represented better performance [48].

In this study, a 5-fold evaluation procedure was conducted to evaluate survival prediction performance for each method. Specifically, for each TCGA cohort, all samples were split into 5 folds, four of which were training sets with one acting as a test set. Once the training sets were further divided, 25% of their samples were used as validation sets for choosing checkpoints. After 5 iterations, each sample in the cohort had

TABLE III

THE RESULTS (CI ↑) OF INDIVIDUAL MODAL, COMPARISON WITH STATE-OF-THE-ARTS (SOTAS), AND ABLATION STUDY UNDER COMPLETE MODAL SETTING. RESULTS WITH NOT SIGNIFICANTLY WORSE THAN THE BEST ($P > 0.05$, TWO-SAMPLE T-TEST) ARE SHOWN IN BOLD

Type	Method	KIRC	LIHC	ESCA	LUSC	LUAD	UCEC
Individual modal	MIL(p)	0.661 ± 0.008	0.642 ± 0.008	0.618 ± 0.013	0.561 ± 0.004	0.567 ± 0.010	0.670 ± 0.010
	Cox(c)	0.698 ± 0.014	0.506 ± 0.027	0.591 ± 0.036	0.567 ± 0.010	0.567 ± 0.022	0.673 ± 0.014
	Cox(g)	0.671 ± 0.004	0.657 ± 0.014	0.518 ± 0.024	0.524 ± 0.020	0.621 ± 0.039	0.680 ± 0.025
SOTAs	GSCNN [10]	0.713 ± 0.006	0.655 ± 0.007	0.550 ± 0.013	0.546 ± 0.009	0.617 ± 0.014	0.726 ± 0.011
	MultiSurv [11]	0.676 ± 0.010	0.623 ± 0.013	0.586 ± 0.021	0.564 ± 0.011	0.626 ± 0.008	0.690 ± 0.030
	Outer product [12]	0.691 ± 0.011	0.662 ± 0.029	0.602 ± 0.025	0.560 ± 0.013	0.602 ± 0.014	0.676 ± 0.027
	Metric learning [13]	0.729 ± 0.019	0.668 ± 0.016	0.604 ± 0.031	0.582 ± 0.012	0.613 ± 0.017	0.690 ± 0.011
	MCAT [14]	0.672 ± 0.010	0.685 ± 0.010	0.576 ± 0.013	0.564 ± 0.012	0.608 ± 0.009	0.683 ± 0.027
Ablation study	w/o Intra-modal interaction	0.733 ± 0.009	0.663 ± 0.012	0.604 ± 0.028	0.581 ± 0.018	0.626 ± 0.011	0.730 ± 0.013
	w/o Inter-modal interaction	0.738 ± 0.008	0.684 ± 0.009	0.608 ± 0.031	0.582 ± 0.016	0.633 ± 0.011	0.737 ± 0.010
	Ours	$\mathbf{0.747 \pm 0.007}$	$\mathbf{0.693 \pm 0.010}$	$\mathbf{0.634 \pm 0.015}$	$\mathbf{0.598 \pm 0.012}$	$\mathbf{0.651 \pm 0.008}$	$\mathbf{0.747 \pm 0.017}$

TABLE IV

THE RESULTS (BS ↓) OF INDIVIDUAL MODAL, COMPARISON WITH STATE-OF-THE-ARTS (SOTAS), AND ABLATION STUDY UNDER COMPLETE MODAL SETTING. RESULTS WITH NOT SIGNIFICANTLY WORSE THAN THE BEST ($P > 0.05$, TWO-SAMPLE T-TEST) ARE SHOWN IN BOLD

Type	Method	KIRC	LIHC	ESCA	LUSC	LUAD	UCEC
Individual modal	MIL(p)	0.210 ± 0.006	0.229 ± 0.005	0.247 ± 0.005	0.254 ± 0.010	0.257 ± 0.011	0.218 ± 0.023
	Cox(c)	0.212 ± 0.008	0.258 ± 0.009	0.259 ± 0.009	0.252 ± 0.003	0.241 ± 0.005	0.192 ± 0.007
	Cox(g)	0.237 ± 0.002	0.258 ± 0.009	0.282 ± 0.008	0.304 ± 0.010	0.283 ± 0.004	0.241 ± 0.007
SOTAs	GSCNN [10]	0.233 ± 0.011	0.291 ± 0.002	0.265 ± 0.007	0.320 ± 0.009	0.256 ± 0.006	$\mathbf{0.138 \pm 0.003}$
	MultiSurv [11]	0.305 ± 0.034	0.287 ± 0.011	0.322 ± 0.020	0.385 ± 0.005	0.309 ± 0.008	0.281 ± 0.081
	Outer product [12]	0.231 ± 0.009	0.242 ± 0.014	0.248 ± 0.008	0.251 ± 0.007	0.266 ± 0.011	0.212 ± 0.019
	Metric learning [13]	0.188 ± 0.007	0.231 ± 0.008	$\mathbf{0.241 \pm 0.005}$	0.255 ± 0.007	0.235 ± 0.006	0.211 ± 0.022
	MCAT [14]	0.265 ± 0.004	0.276 ± 0.001	0.259 ± 0.002	0.281 ± 0.001	0.284 ± 0.003	0.274 ± 0.004
Ablation study	w/o Intra-modal interaction	0.194 ± 0.008	0.264 ± 0.009	0.245 ± 0.004	$\mathbf{0.250 \pm 0.006}$	0.245 ± 0.009	0.174 ± 0.011
	w/o Inter-modal interaction	0.198 ± 0.004	0.222 ± 0.002	$\mathbf{0.234 \pm 0.007}$	$\mathbf{0.248 \pm 0.004}$	$\mathbf{0.238 \pm 0.006}$	0.191 ± 0.013
	Ours	$\mathbf{0.175 \pm 0.006}$	$\mathbf{0.213 \pm 0.005}$	$\mathbf{0.240 \pm 0.003}$	$\mathbf{0.246 \pm 0.005}$	$\mathbf{0.239 \pm 0.006}$	0.157 ± 0.007

been used as a test sample one time, thus generating 5-fold performance. For each trial, a 5-fold evaluation procedure was ran 10 times. The result of $mean \pm std$ was reported and Two-sample T-tests were used to quantify the difference significance of each method.

B. Experiments With Complete Modalities

The results of single modality settings are shown first and include: Graph based multiple instance learning (MIL) on pathology slide, Cox modeling on clinical records, and Cox modeling on genomic profile, respectively. For comparison with complete modalities setting, we compared four state-of-the-art methods: GSCNN [10], MultiSurv [11], Metric learning [13], Outer product [12], and MCAT [14]. We also compare the proposed method with the following baselines for ablation analysis:

- *w/o intra-modal interaction:* Replace the GCN layers in our framework with linear layers.
- *w/o inter-modal interaction:* Directly input the multi-modal graphs into three GCN branches, then fuse the decisions to determine final survival.

As shown in TABLE III, we first compare the ranking ability (CI) of different methods. Generally, the individual modal performances are lower than multimodal methods, which indicates that diverse and complementary information among modalities is helpful for survival prediction tasks. As our method considers not only the intra-modal interaction but also the inter-modal interaction to facilitate survival analysis, our method achieves the mean CI of 0.747, 0.693, 0.634, 0.598, 0.651 and 0.747 on six cancer cohorts, which consistently outperform the SOTAs and baselines. These results prove the clinical value and effectiveness of this framework.

Second, as shown in TABLE IV, we analyze the BS of our method on each cancer cohort. The data shows that our method achieves a minimum BS of 0.175, 0.213, 0.240, 0.246 and 0.239 on KIRC, LIHC, ESCA, LUSC and LUAD, respectively. For the UCEC cohort, our method also achieves a satisfactory result of 0.157, only 0.019 more than the GSCNN. These results show that our method can relatively accurately predict the absolute survival risk of cancer patients.

Last but not least, each cancer cohort was split into high-risk and low-risk groups based on the median survival risk generated by our model. If the survival prediction was accurate, then KM curves of these two groups should be significantly different. Specifically, as shown in Fig. 3, Metric learning achieves SOTA performance on three cancer cohorts (*i.e.*, KIRC, ESCA and LUSC). MCAT, MultiSurv and GSCNN show SOTA performance on LIHC, LUAD and UCEC, respectively. P-Values of our method for all six cancer cohorts were found to be lower than the corresponding SOTAs above, and less than 0.01 overall, which indicate the effectiveness of our model for patient stratification.

C. Interpretation of Results

Cancer survival prediction requires not only high performance, but also strong rationale to deliver proper judgement.

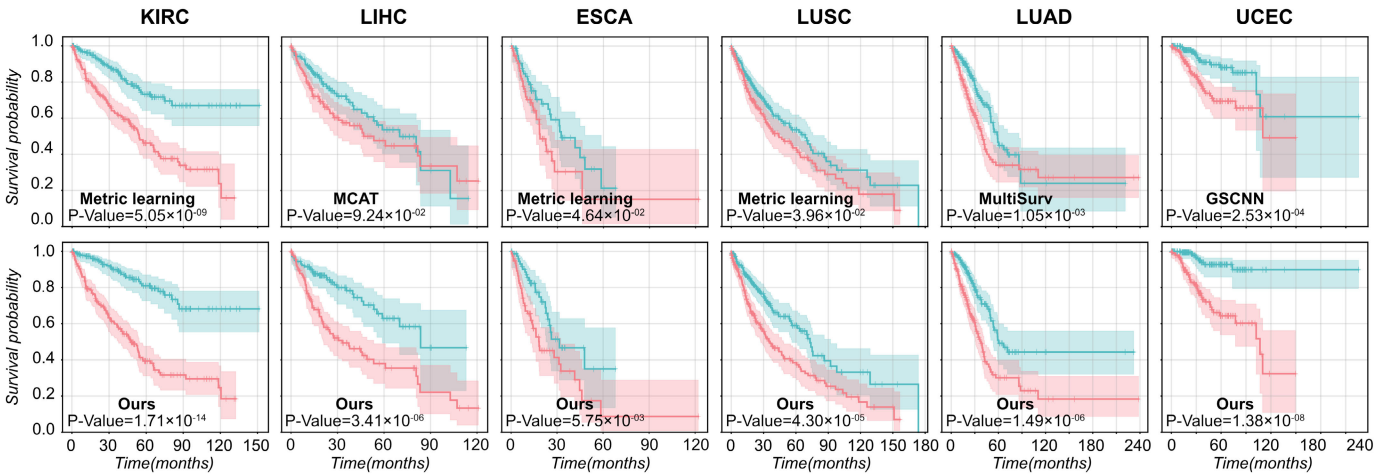

Fig. 3. KM analysis of best comparative method and our proposed framework for different dataset. All the patients across the five test folds are combined and analysis here. For each cohort, patients were stratified into high-risk (red curves) and low-risk (green curves) groups by the median score output by predictive models. The P-Value for each Log-rank test is shown in each figure.

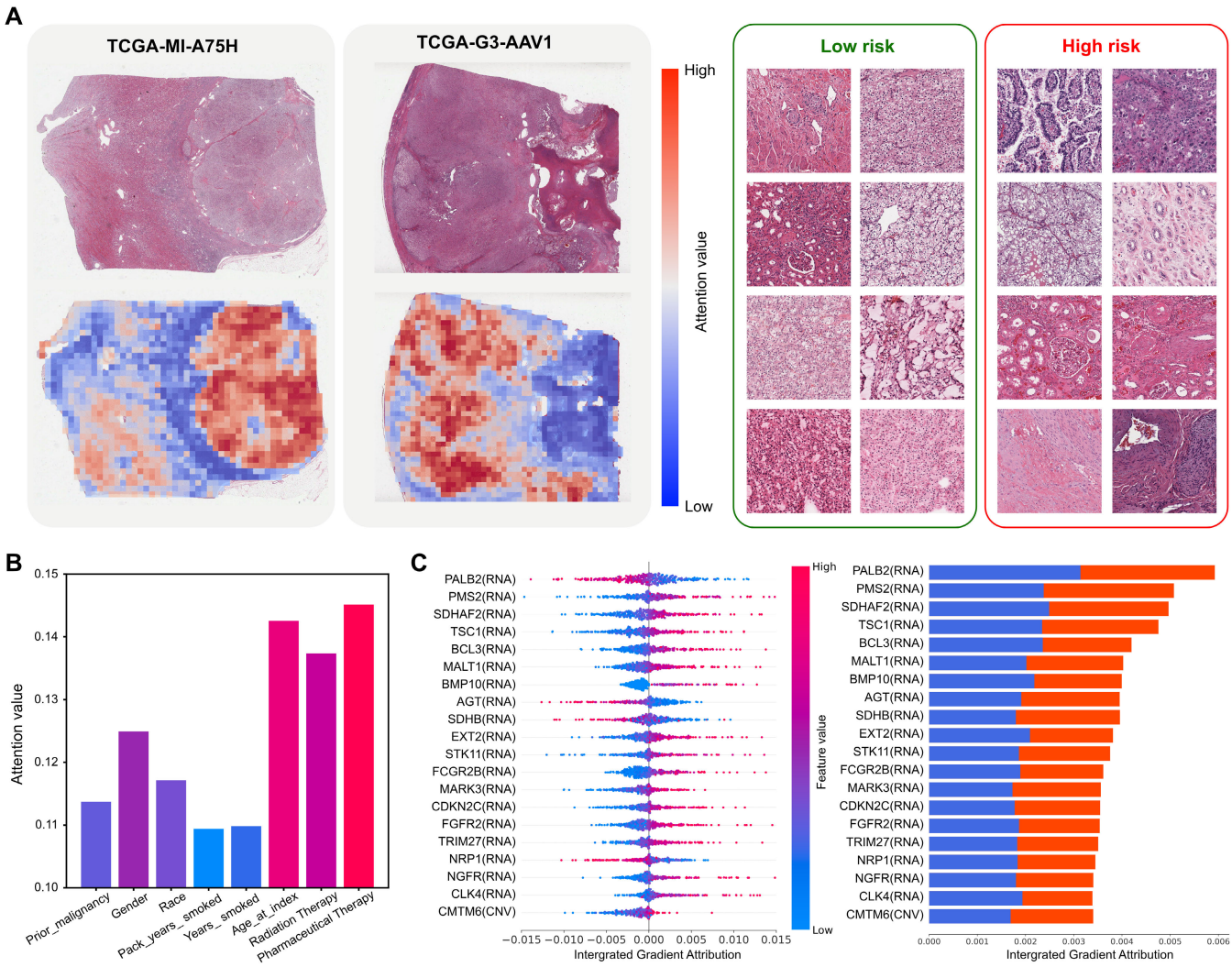

Fig. 4. Interpretation of the proposed framework on KIRC dataset. A: Heat map of pathological slides. B: Attention value of clinical records. C: Integrated gradient analysis of genomic profile.

Benefiting from the graph based multimodal data modeling, our framework could provide abundant interpretability with the readout operations [49], and integrated gradient analysis [50] on multimodal graphs. First, Fig. 4-A was used

to visualize pathological regions that are interpreted by the model, of which representative regions (red) of high/low risk patients may be used to discover new biomarkers. Second, Fig. 4-B shows the decision-making contributions of individual elements in the clinical record. For example, the data shows that both radiation therapy and pharmaceutical therapy have high-rank attention value for modal prediction. Moreover, patient age was also regarded as a significant factor, which is consistent with prior knowledge. The above observations may provide reference for clinicians to formulate treatment plans tailored to patient needs. Fig. 4-C expresses genomic profile data. The left portion of the figure shows the impact of changes in the top 20 genomic features on the prediction results. The right portion of the figure shows the value distribution of genomic features corresponding to patients within the high (red) and low (blue) risk groups, which may help to improve molecular knowledge of tumors.

D. Experiments With Missing Modalities

Due to the variant costs and trauma degree of multi-omics examination, the scenarios of missing modalities are very common. We further evaluate the robustness of the proposed model when validating with unexpected missing modalities, *e.g.*, clinical records only, clinical records with pathology slide, *etc.*Specifically, we follow the method of [11] to extract the feature representation of each modality and then implement a decision fusion model combined with the following techniques, which can also handle missing modalities to some extent:

- *Zero padding:* Pad the feature representations of the missing modalities with zeros [19], [20].
- *MFM:* Reconstruct the missed feature representations by multimodal factorized method [21].
- *Autoencoder:* Use the feature representations of available modalities as inputs of the autoencoder to reconstruct the feature representation of missing modalities [22], [23].
- *Ours (LB):* Remove the online MAE from our framework and training with the available modalities, which can be regarded as the lower bound (LB) of our model.

TABLE V and Fig. 5 detail the results of missing modal settings. First, reconstruction results of MFM were found to be sensitive to hyperparameters and easily over-fit the data, which suggests that MFM is not a useful tool for resolving missing modal scenarios and assisting in determining cancer prognosis. Second, zero padding is regarded as a simple data augmentation operation, which has the potential to support model performance and assist in inference. Third, off-line trained autoencoders can assist models by filling in knowledge gaps of missing modalities to a certain extent; however, they can only reconstruct fixed feature representations, which allows for amplification of reconstruction errors through network forward propagation. This has the potential to interfere with model decision combination. In contrast, the proposed online MAE method outlined in these studies synchronously learns higher-order features in an end-to-end manner, which has the potential to reduce decision-making interference when determining patient prognosis. As shown in TABLE V and Fig. 5, our method consistently outperformed baseline values for six missing modal settings, which ensures this method has clinical value.

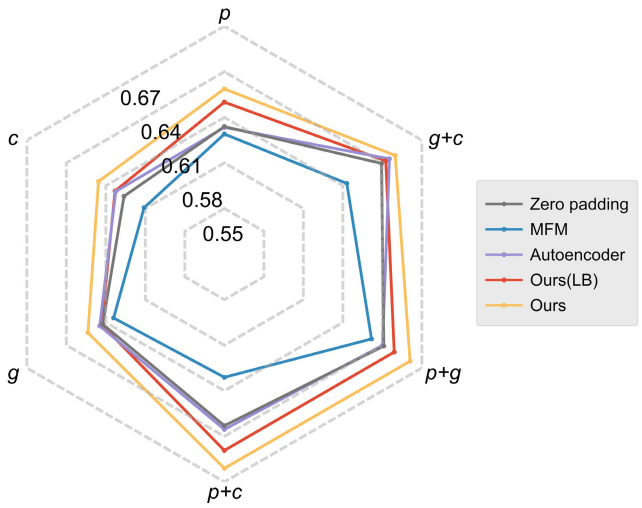

Fig. 5. Overall performance comparison (CI ↑) of the proposed method with four comparative methods using six different missing modal settings.

1) Quantitative Analysis of MAE: In our model, the online MAE module plays a key role in improving the prediction robustness. Therefore, in Fig. 6, the comparison of hyperedge values generated via complete modalities settings as well as missing modalities settings across validation folds was visualized to investigate the principle of online MAE. Moreover, we adopted the Pearson correlation coefficient (ρ) as a metric to measure the correlation between the two sets of generated results. As shown in Fig. 6, the dependence capturing ability of the transformer and generation ability of the autoencoder suggest that the results generated from these studies correlate strongly with missing modality and complete modality settings ($\rho > 0.8$). These results indicate that our proposed online MAE can effectively reconstruct the high-order information of missing modes, hence improving the robustness of the prediction results.

V. Discussion

With the rapid development of imaging and genomics in recent years, multimodal medical data are accumulating rapidly at an unprecedented scale. Abundant information contained in multimodal medical data brings opportunities to better predict patient survival and prognosis. However, the heterogeneity of this data brings challenges to the design of automated analysis methods. The primary challenge is how to flexibly and effectively model multi-modal data. To the best of our knowledge, this manuscript is the first to model multimodal medical data as full graph structures. Our method can intuitively preserve the internal correlation of elements in the mode, and is convenient for subsequent processing and interpretable analysis.

Based on our proposed multimodal graphs, we further proposed a novel HGCN to mine prognostic-related information. Furthermore, we adopted a node message passing mechanism for the graph network to analyze intra-modal interactions. Additionally, we designed an HGCN with a hyperedge mixing module to achieve advanced modal interaction. During the

TABLE V

COMPARISON RESULTS (CI ↑) WITH MISSING MODALITIES SETTINGS. ALL RESULTS ARE SHOWN WITH THREE DECIMAL PLACES. RESULTS WITH NOT SIGNIFICANTLY WORSE THAN THE BEST ($P > 0.05$, TWO-SAMPLE T-TEST) ARE SHOWN IN BOLD

Modality			Method	KIRC	LIHC	ESCA	LUSC	LUAD	UCEC
p	c	g							
✓	×	×	Zero padding	0.663 ± 0.007	0.608 ± 0.009	0.588 ± 0.022	0.555 ± 0.008	0.563 ± 0.007	0.649 ± 0.011
			MFM [21]	0.663 ± 0.007	**0.642 ± 0.016**	0.572 ± 0.028	0.547 ± 0.019	0.537 ± 0.011	0.633 ± 0.017
			Autoencoder	0.643 ± 0.008	**0.643 ± 0.006**	0.610 ± 0.014	0.536 ± 0.011	0.546 ± 0.010	0.643 ± 0.013
			Ours (LB)	0.661 ± 0.008	**0.642 ± 0.008**	**0.618 ± 0.013**	0.561 ± 0.004	0.567 ± 0.010	**0.670 ± 0.010**
			Ours	**0.677 ± 0.008**	**0.647 ± 0.005**	**0.623 ± 0.012**	**0.568 ± 0.008**	**0.577 ± 0.009**	**0.682 ± 0.021**
×	✓	×	Zero padding	0.701 ± 0.005	0.480 ± 0.008	0.563 ± 0.035	0.576 ± 0.009	0.571 ± 0.010	0.685 ± 0.013
			MFM [21]	0.687 ± 0.014	**0.498 ± 0.025**	0.553 ± 0.040	0.559 ± 0.022	0.565 ± 0.014	0.624 ± 0.067
			Autoencoder	0.702 ± 0.004	0.497 ± 0.015	0.605 ± 0.016	0.576 ± 0.009	0.568 ± 0.009	0.665 ± 0.015
			Ours (LB)	**0.709 ± 0.005**	**0.503 ± 0.019**	0.591 ± 0.016	**0.583 ± 0.008**	0.560 ± 0.014	0.671 ± 0.009
			Ours	**0.709 ± 0.005**	**0.508 ± 0.009**	**0.620 ± 0.007**	**0.582 ± 0.009**	**0.578 ± 0.006**	**0.695 ± 0.006**
×	×	✓	Zero padding	0.658 ± 0.007	**0.677 ± 0.010**	0.489 ± 0.044	**0.561 ± 0.011**	0.614 ± 0.012	0.677 ± 0.018
			MFM [21]	**0.663 ± 0.012**	0.640 ± 0.029	0.488 ± 0.036	0.530 ± 0.019	0.611 ± 0.010	**0.692 ± 0.027**
			Autoencoder	0.656 ± 0.008	**0.670 ± 0.017**	**0.523 ± 0.029**	**0.555 ± 0.011**	**0.616 ± 0.006**	0.668 ± 0.010
			Ours (LB)	**0.669 ± 0.009**	0.660 ± 0.016	**0.508 ± 0.035**	0.543 ± 0.021	0.605 ± 0.014	0.694 ± 0.011
			Ours	**0.666 ± 0.014**	**0.678 ± 0.016**	**0.525 ± 0.017**	0.543 ± 0.016	**0.622 ± 0.012**	**0.708 ± 0.021**
✓	✓	×	Zero padding	0.733 ± 0.007	0.591 ± 0.012	0.606 ± 0.023	0.579 ± 0.006	0.587 ± 0.007	0.701 ± 0.009
			MFM [21]	0.667 ± 0.008	0.633 ± 0.019	0.572 ± 0.028	0.549 ± 0.015	0.544 ± 0.016	0.643 ± 0.022
			Autoencoder	0.725 ± 0.005	0.639 ± 0.006	0.611 ± 0.022	0.569 ± 0.009	0.574 ± 0.011	0.694 ± 0.012
			Ours (LB)	0.725 ± 0.010	**0.649 ± 0.006**	**0.655 ± 0.014**	0.582 ± 0.010	0.587 ± 0.015	0.699 ± 0.017
			Ours	**0.750 ± 0.007**	**0.644 ± 0.005**	**0.664 ± 0.014**	**0.590 ± 0.010**	**0.601 ± 0.006**	**0.717 ± 0.012**
✓	×	✓	Zero padding	0.704 ± 0.005	0.677 ± 0.010	0.560 ± 0.025	**0.584 ± 0.008**	0.619 ± 0.010	0.703 ± 0.011
			MFM [21]	**0.709 ± 0.011**	0.665 ± 0.015	0.571 ± 0.027	0.569 ± 0.012	0.601 ± 0.015	0.675 ± 0.012
			Autoencoder	0.692 ± 0.009	**0.691 ± 0.009**	0.574 ± 0.021	0.563 ± 0.011	0.617 ± 0.005	0.705 ± 0.016
			Ours (LB)	0.704 ± 0.011	0.683 ± 0.008	**0.602 ± 0.012**	0.572 ± 0.010	0.619 ± 0.013	0.714 ± 0.009
			Ours	**0.712 ± 0.010**	**0.695 ± 0.010**	**0.606 ± 0.016**	**0.582 ± 0.011**	**0.639 ± 0.011**	**0.733 ± 0.021**
×	✓	✓	Zero padding	**0.731 ± 0.008**	**0.671 ± 0.007**	0.513 ± 0.047	**0.578 ± 0.010**	**0.631 ± 0.012**	0.713 ± 0.008
			MFM [21]	0.657 ± 0.016	0.641 ± 0.020	0.516 ± 0.033	0.558 ± 0.011	0.610 ± 0.014	0.696 ± 0.020
			Autoencoder	**0.727 ± 0.009**	0.658 ± 0.017	**0.571 ± 0.019**	0.567 ± 0.013	**0.631 ± 0.009**	0.719 ± 0.011
			Ours(LB)	**0.732 ± 0.005**	0.656 ± 0.016	**0.560 ± 0.019**	0.562 ± 0.013	0.623 ± 0.013	**0.723 ± 0.010**
			Ours	**0.728 ± 0.011**	**0.663 ± 0.014**	**0.565 ± 0.017**	**0.571 ± 0.012**	**0.636 ± 0.012**	**0.735 ± 0.015**

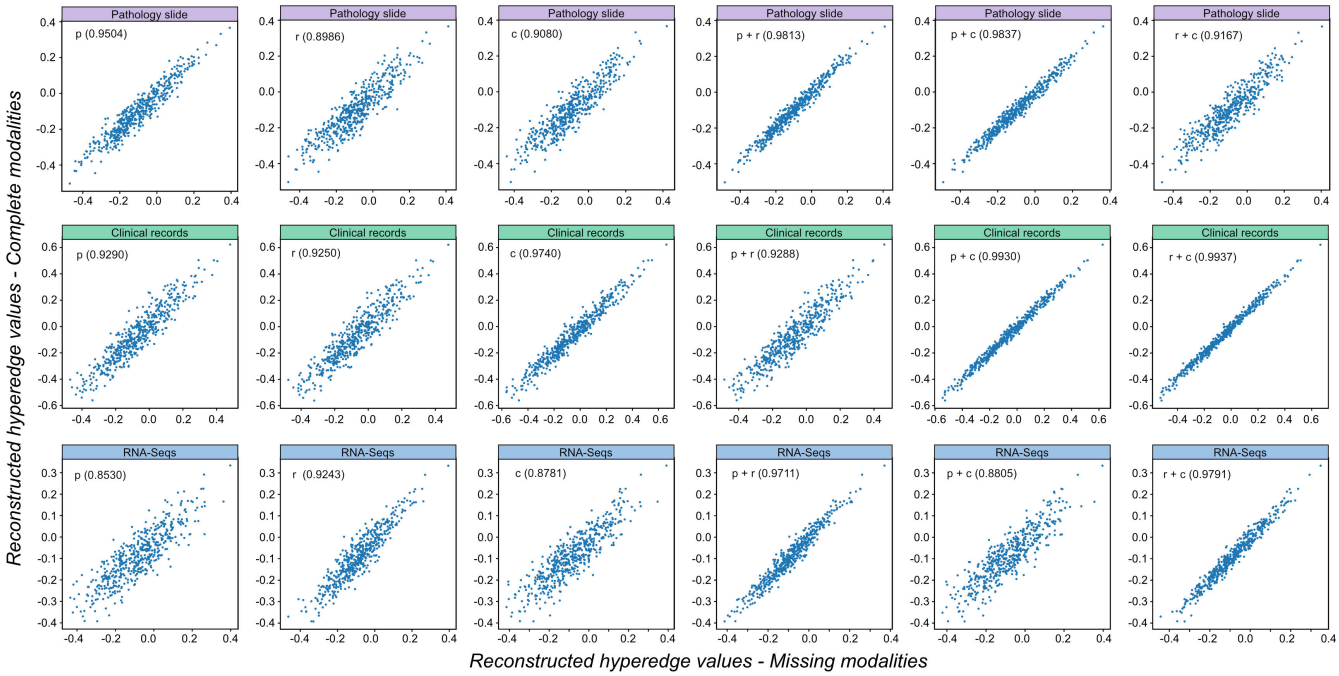

Fig. 6. Comparison of average hyperedge reconstruction results across the test folds of KIRC cohort. Each point represents one dimension of the hyperedge, with values generated by different missing modal settings (x axis) and complete modal setting (y axis). Pearson correlation coefficients (ρ) are shown in the legend of each image.

model inference stage, final survival predictions were obtained by combining the output of all available modalities. Our method involves a multi-view and multi-stage fusion scheme, and thus can perform more reliable survival predictions.

This manuscript systematically investigated the robustness of a multimodal survival prediction model, which is a topic of great clinical value that had remained ignored in existing studies. To handle missing modality scenarios due to financial stress or physical limitations of the patient, we proposed an online MAE paradigm and embedded it in our HGCN. Experimental results and analysis verified our online MAE can effectively capture intrinsic dependencies between modalities and generate hyperedges for missing modalities. Our research provides new insights, specifically with regards to how masked signal modeling not only has the potential to serve as a pretext task for self-supervised learning, but also can be trained end-to-end and applied to other scientific fields.

Our method does have some limitations. First, in this study, all split data, which include training sets, validation sets and test sets, adopted a consistent pre-processing method. Therefore, pathological image file loss and errors in data recording caused by human error may have affected the prediction performance of our model. Generally, this issue can be addressed by stronger data augmentation. Moreover, in clinical practice, this problem can be addressed through gradual covering of these samples for model fine-tuning. Second, to cope with missing modalities, we adopted a simple but effective voting mechanism for data combination to elicit a final decision. When individual performances of the fused modalities contain large gaps (about 15%), decision fusion should be considered a failure. Therefore, when our method is expanded to incorporate more modalities (*e.g.*, X-ray, CT, *etc.*), it needs to be evaluated in advance through fine tuning. In future studies, we will focus on developing more flexible and adaptive decision fusion mechanisms to address this challenge. Third, in our study, the performance of each method was evaluated according to the 5-fold evaluation procedure. The generalizability and application of our model can be reliably evaluated in this manner to a large extent; however, clinical performance of the algorithm still needs to be evaluated on external test sets and multi-center test sets, which will be the goal of our future works.

VI. Conclusion

In this manuscript, we studied the multimodal cancer survival prediction, a challenging yet clinically important task and proposed effective techniques to tackle missing modalities during network inference. Specifically, we pioneered the usage of full graph structures to model patients' multimodal data, resulting in more flexible processing patterns and more sufficient interpretability. Then we propose a novel HGCN to realize intra- and inter-modal interaction between multimodal graphs. By combining the knowledge generated by the intra- and inter-model interactions, the potential of multimodal data can be better unleashed, leading to more reliable predictions. More importantly, to handle the missing modalities, we elaborately designed a masked autoencoder to generate the missing hyperedges for model inference. The effectiveness and robustness of the proposed method were evaluated on six public cancer cohorts. Furthermore, we conducted multiple explanations and analyses for generated prediction results, which indicated overall ease of interpretability for our model. In the future, we will focus on exploring more effective and interpretable graph-based methods to model the patients' multimodal data and further quantify the intra- and inter-modal interactions. Moreover, we are dedicated to the collection of more large multimodal cancer cohorts to evaluate our model on more complex clinical tasks (*e.g.*, efficacy prediction) and scenarios (*e.g.*, independent choice of treatment type).

References

[1] H. Sung et al., "Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA, A Cancer J. Clinicians*, vol. 71, no. 3, pp. 209–249, May 2021.

[2] D. R. Cox, "Regression models and life-tables," *J. Roy. Stat. Soc. B, Methodol.*, vol. 34, no. 2, pp. 187–202, 1972.

[3] X. Huang, S.-J. Soong, W. H. McCarthy, M. M. Urist, and C. M. Balch, "Classification of localized melanoma by the exponential survival trees method," *Cancer*, vol. 79, no. 6, pp. 1122–1128, Mar. 1997.

[4] M. Radespiel-Tröger, W. Hohenberger, and B. Reingruber, "Improved prediction of recurrence after curative resection of colon carcinoma using tree-based risk stratification," *Cancer*, vol. 100, no. 5, pp. 958–967, Mar. 2004.

[5] M. Banerjee, D. Biswas, W. Sakr, and D. P. Wood, "Recursive partitioning for prognostic grouping of patients with clinically localized prostate carcinoma," *Cancer*, vol. 89, no. 2, pp. 404–411, Jul. 2000.

[6] M. Watanabe et al., "Recent progress in multidisciplinary treatment for patients with esophageal cancer," *Surgery Today*, vol. 50, no. 1, pp. 12–20, Jan. 2020.

[7] V. Jindal, K. K. Sahu, S. Gaikazian, A. D. Siddiqui, and I. Jaiyesimi, "Cancer treatment during COVID-19 pandemic," *Med. Oncol.*, vol. 37, pp. 1–3, Jul. 2020.

[8] Y. L. Orlov and A. V. Baranova, "Editorial: Bioinformatics of genome regulation and systems biology," *Frontiers Genet.*, vol. 11, p. 625, Jul. 2020.

[9] W. G. Feero, "Bioinformatics, sequencing accuracy, and the credibility of clinical genomics," *JAMA*, vol. 324, no. 19, pp. 1945–1947, 2020.

[10] P. Mobadersany et al., "Predicting cancer outcomes from histology and genomics using convolutional networks," *Proc. Nat. Acad. Sci. USA*, vol. 115, no. 13, pp. E2970–E2979, Mar. 2018.

[11] L. A. Vale-Silva and K. Rohr, "Long-term cancer survival prediction using multimodal deep learning," *Sci. Rep.*, vol. 11, no. 1, pp. 1–12, Jun. 2021.

[12] R. J. Chen et al., "Pathomic fusion: An integrated framework for fusing histopathology and genomic features for cancer diagnosis and prognosis," *IEEE Trans. Med. Imag.*, vol. 41, no. 4, pp. 757–770, Apr. 2022.

[13] A. Cheerla and O. Gevaert, "Deep learning with multimodal representation for pancancer prognosis prediction," *Bioinformatics*, vol. 35, no. 14, pp. i446–i454, Jul. 2019.

[14] R. J. Chen et al., "Multimodal co-attention transformer for survival prediction in gigapixel whole slide images," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 4015–4025.

[15] S. Haneuse, D. Arterburn, and M. J. Daniels, "Assessing missing data assumptions in EHR-based studies: A complex and underappreciated task," *JAMA Netw. Open*, vol. 4, no. 2, Feb. 2021, Art. no. e210184.

[16] M. L. Bell, M. Fiero, N. J. Horton, and C.-H. Hsu, "Handling missing data in RCTs; a review of the top medical journals," *BMC Med. Res. Methodol.*, vol. 14, no. 1, pp. 1–8, Dec. 2014.

[17] F. Cismondi, A. S. Fialho, S. M. Vieira, S. R. Reti, J. M. C. Sousa, and S. N. Finkelstein, "Missing data in medical databases: Impute, delete or classify?" *Artif. Intell. Med.*, vol. 58, no. 1, pp. 63–72, 2013.

[18] J. R. Carpenter and M. Smuk, "Missing data: A statistical framework for practice," *Biometrical J.*, vol. 63, no. 5, pp. 915–947, Jun. 2021.

[19] Z. Chen, S. Wang, and Y. Qian, "Multi-modality matters: A performance leap on VoxCeleb," in *Proc. INTERSPEECH*, Oct. 2020, pp. 2252–2256.

[20] G. Shen, X. Wang, X. Duan, H. Li, and W. Zhu, "MEmoR: A dataset for multimodal emotion reasoning in videos," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 493–502.

[21] Y.-H. H. Tsai, P. P. Liang, A. Zadeh, L.-P. Morency, and R. Salakhutdinov, "Learning factorized multimodal representations," in *Proc. Int. Conf. Learn. Represent.*, 2019, pp. 1–20.

[22] L. Tran, X. Liu, J. Zhou, and R. Jin, "Missing modalities imputation via cascaded residual autoencoder," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1405–1414.

[23] Y. Liu et al., "Incomplete multi-modal representation learning for Alzheimer's disease diagnosis," *Med. Image Anal.*, vol. 69, Apr. 2021, Art. no. 101953.
[24] A. Riasatian et al., "Fine-tuning and training of densenet for histopathology image representation using TCGA diagnostic slides," *Med. Image Anal.*, vol. 70, May 2021, Art. no. 102032.
[25] A. Subramanian et al., "Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles," *Proc. Nat. Acad. Sci. USA*, vol. 102, no. 43, pp. 15545–15550, Aug. 2005.
[26] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *Proc. Int. Conf. Learn. Represent.*, 2017, pp. 1–14. [Online]. Available: https://openreview.net/forum?id=SJU4ayYgl
[27] Y. Feng, H. You, Z. Zhang, R. Ji, and Y. Gao, "Hypergraph neural networks," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, 2019, pp. 3558–3565.
[28] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[29] W. Shao, T. Wang, Z. Huang, Z. Han, J. Zhang, and K. Huang, "Weakly supervised deep ordinal cox model for survival prediction from whole-slide pathological images," *IEEE Trans. Med. Imag.*, vol. 40, no. 12, pp. 3739–3747, Dec. 2021.
[30] Z. Lin et al., "CT-guided survival prediction of esophageal cancer," *IEEE J. Biomed. Health Informat.*, vol. 26, no. 6, pp. 2660–2669, Jun. 2022.
[31] S. Bae et al., "Radiomic MRI phenotyping of glioblastoma: Improving survival prediction," *Radiology*, vol. 289, no. 3, pp. 797–806, Dec. 2018.
[32] X. Liu et al., "Self-supervised learning: Generative or contrastive," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 1, pp. 857–876, Jun. 2021.
[33] Z. Xie et al., "SimMIM: A simple framework for masked image modeling," 2021, *arXiv:2111.09886*.
[34] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.
[35] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le, "XLNet: Generalized autoregressive pretraining for language understanding," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 1–11.
[36] J. Lee et al., "BioBERT: A pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, Feb. 2020.
[37] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," 2021, *arXiv:2111.06377*.
[38] R. J. Chen et al., "Whole slide images are 2D point clouds: Context-aware survival prediction using patch-based graph convolutional networks," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2021, pp. 339–349.
[39] W. Hou et al., "H^2-MIL: Exploring hierarchical representation with heterogeneous multiple instance learning for whole slide image analysis," in *Proc. 36th AAAI Conf. Artif. Intell., (AAAI), 34th Conf. Innov. Appl. Artif. Intell., (IAAI), 12th Symp. Educ. Adv. Artif. Intell., (EAAI)*. Palo Alto, CA, USA: AAAI Press, Mar. 2022, pp. 933–941. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/19976
[40] F. R. Balkwill, M. Capasso, and T. Hagemann, "The tumor microenvironment at a glance," *J. Cell Sci.*, vol. 125, no. 23, pp. 5591–5596, Dec. 2012.
[41] J. Saltz et al., "Spatial organization and molecular correlation of tumor-infiltrating lymphocytes using deep learning on pathology images," *Cell Rep.*, vol. 23, no. 1, pp. 181–193, 2018.
[42] W. L. Hamilton, R. Ying, and J. Leskovec, "Inductive representation learning on large graphs," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 1025–1035.
[43] S. Bai, F. Zhang, and P. H. S. Torr, "Hypergraph convolution and hypergraph attention," *Pattern Recognit.*, vol. 110, Feb. 2021, Art. no. 107637.
[44] I. Tolstikhin et al., "MLP-mixer: An all-MLP architecture for vision," 2021, *arXiv:2105.01601*.
[45] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.
[46] J. D. M.-W. C. Kenton and L. K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Naacl-HLT*, 2019, pp. 4171–4186.
[47] J. Yao, X. Zhu, J. Jonnagaddala, N. Hawkins, and J. Huang, "Whole slide images based cancer survival prediction using attention guided deep multiple instance learning networks," *Med. Image Anal.*, vol. 65, Oct. 2020, Art. no. 101789.
[48] P. Wang, Y. Li, and C. K. Reddy, "Machine learning for survival analysis: A survey," *ACM Comput. Surv.*, vol. 51, no. 6, pp. 1–36, 2019.
[49] Y. Li, D. Tarlow, M. Brockschmidt, and R. Zemel, "Gated graph sequence neural networks," 2015, *arXiv:1511.05493*.
[50] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *Proc. 34th Int. Conf. Mach. Learn.*, in Proceedings of Machine Learning Research, vol. 70, D. Precup and Y. W. Teh, Eds., Aug. 2017, pp. 3319–3328.