2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) | 979-8-3503-8622-6/24/$31.00 ©2024 IEEE | DOI: 10.1109/BIBM62325.2024.10821720

ContraMAE: **Contr**astive **a**lignment **m**asked **a**uto**e**ncoder framework for cancer survival prediction

1st Suixue Wang
School of Information and Communication Engineering
Hainan University
Haikou, China
wangsuixue@hainanu.edu.cn

2nd Huiyuan Lai
University of Groningen
Groningen, Netherlands
h.lai@rug.nl

3rd Shuling Wang
Department of Neurology
Affiliated Haikou Hospital of Xiangya School of Medicine,
Central South University
Haikou, China
18184657175@163.com

4th Qingchen Zhang
School of Computer Science and Technology
Hainan University
Haikou, China
zhangqingchen@hainanu.edu.cn
The corresponding author

***Abstract*—With the rapid advancement in multimodal fusion technology, the integration of pathological images with genomics data has achieved promising results in cancer survival prediction. However, most existing multimodal models are not pre-trained by combining pathology and genomics modalities, ignoring the inherent task-agnostic associations between different modalities. While some self-supervised methods align multimodal information through pre-training objectives such as correlation and mean square error, they lack in-depth multimodal interaction. To address these issues, we propose ContraMAE, a contrastive alignment masked autoencoder framework, to fuse pathological images and genomics data for cancer survival prediction. Concretely, we introduce a contrastive objective to align multimodality and construct their intrinsic consistency. Besides, we design two reconstruction objectives to capture the complex relationships between multi-modalities by mutually compensating for the information that each side lacks. In survival prediction, the pathology and genomics encodings from the ContraMAE encoder are concatenated as the final representation to generate a survival risk score. Experimental results demonstrate that ContraMAE outperforms existing state-of-the-art methods on five cancer datasets sourced from The Cancer Genome Atlas (TCGA). The code is available at https://github.com/SuixueWang/ContraMAE.**

***Index Terms*—Masked autoencoder, Contrastive learning, Multimodal pre-training, Pathology-genomics Alignment, Survival prediction**

I. Introduction

Cancer is a significant global health issue. It accounts for nearly 10 million deaths worldwide each year, making it a leading cause of death [8]. Among the diverse malignancies, breast, lung, colorectal, liver, and stomach cancers emerge as the most prevalent. Projections indicate that by 2040, the global incidence of cancer will rise by 47%, from 19.3 million in 2020 to 28.4 million. Typically, effective cancer treatments rely heavily on subtype diagnosis and survival prediction [8], [13]. Therefore, accurate survival prediction is important for clinicians to develop reasonable treatment regimens and thus improve patient survival rates.

In recent years, multimodal fusion methods have advanced cancer survival prediction by integrating pathological images and genomics data [3], [12], [14], [15]. For example, Wang et al. [14] propose a GPDBN model with a bilinear feature encoding module to enhance performance. Wu et al. [15] introduce the CAMR model, which uses cross-aligned learning to project diverse data types into a shared space and acquire the respective representations for each data type. Chen et al. [2] develop the MCAT framework, employing genomic-guided co-attention and attention pooling for multimodal integration. Zhou et al. [19] propose the CMTA framework to explore the intrinsic cross-modal relationships. However, these methods focus on multimodal fusion for survival prediction without considering task-agnostic associations between modalities via pre-training. In contrast, based on the pre-training paradigm, Yao et al. [17] present DeepCorrSurv, which maximizes correlation to align modalities. Ding et al. [3] use PathOmics to align modalities by minimizing mean square error. Thus, the existing multimodal fusion methods have two main limitations: 1) Most methods integrate multimodal data without pre-training, overlooking task-agnostic relationships across modalities; 2) Methods using self-supervised learning primarily align multimodal data through objectives like correlation or mean square error [3], [17], which may fail to capture complex relationships between pathological images and genomics data.

In this paper, we propose ContraMAE, a contrastive alignment masked autoencoder framework to fuse pathological images and genomics data for cancer survival prediction. To

learn the intrinsic task-agnostic interactions between multi-modalities, we present an improved masked autoencoder with three pre-training objectives. Specifically, we first introduce a pathology-genomics contrastive objective to align multi-modality and construct the inherent consistency between different modalities. Then, we design two reconstruction objectives, genomics reconstruction (GR), and pathology reconstruction (PG), aiming to capture the complex relationships between multi-modalities by mutually compensating for the information that each side lacks. In survival prediction, we concatenate the pathology and genomics encodings from the ContraMAE encoder as the final representation to generate a risk score for survival prediction. We conduct experiments to compare ContraMAE with existing state-of-the-art methods on five cancer datasets from The Cancer Genome Atlas (TCGA). The experimental results demonstrate that ContraMAE achieves the highest C-index values on all five cancer datasets, demonstrating the effectiveness of our framework.

In summary, our contributions are summarized as follows:

1) We propose a contrastive alignment masked autoencoder framework, ContraMAE, that employs three pre-training objectives to learn intrinsic task-agnostic relationships between modalities for improving performance on the downstream survival prediction task.
2) We introduce a contrastive objective that aligns information and establishes intrinsic consistency across various modalities, which is beneficial for subsequent multimodal data reconstructions.
3) We design two reconstruction objectives to capture the complex interactions across multiple modalities, where each modality compensates for the missing information in the other.

II. Method

In this section, we present the overview of the contrastive alignment masked autoencoder (ContraMAE) framework for survival prediction, as illustrated in Figure 1. We first delineate ContraMAE architecture, comprising encoder and decoder modules. Next, we introduce three pre-training objectives utilized in ContraMAE, followed by the downstream survival prediction task. Lastly, we detail the implementation settings.

A. ContraMAE Architecture

After preprocessing (described in Subsection III-A), let $\boldsymbol{P} \in \mathbb{R}^{r \times r \times 3}$ denote the representative region of a whole slide image (WSI) in pathology, let $\boldsymbol{G} \in \mathbb{R}^{3 \times d}$ represent genomics data containing RNA-Seq, miRNA, and DNA methylation. The survival time and survival status are denoted as t and e.

ContraMAE Encoder. We crop the representative region of WSI $\boldsymbol{P}$ into image patches $\boldsymbol{P}_{pat} \in \mathbb{R}^{L \times (r' \times r' \times 3)}$, where L is the number of patches. Then, the image patches are embedded using a lightweight CNN, with the kernel size matching the patch size and the number of output channels set to d. This process yields L patch embeddings, each with d dimensions. Concurrently, we embed the three types of genomic data using a shared linear fully connected network.

$$\boldsymbol{P}_{init} = \text{PatchEmbed}\left(\boldsymbol{P}_{pat}\right), \boldsymbol{P}_{init} \in \mathbb{R}^{L \times d} \tag{1}$$

$$\boldsymbol{G}_{init} = \text{Linear}\left(\boldsymbol{G}\right), \boldsymbol{G}_{init} \in \mathbb{R}^{3 \times d} \tag{2}$$

Subsequently, we concatenate the CLS token embedding, patch embeddings, and genomics embeddings, incorporating their respective positional encodings. These combined embeddings serve as the input fed into the cross-modal encoder to interact with two modalities:

$$\boldsymbol{H}_{in} = [\boldsymbol{c}_{init}, \boldsymbol{G}_{init}, \boldsymbol{P}_{init}] + [\boldsymbol{E}_{cls}, \boldsymbol{E}_{G}, \boldsymbol{E}_{P}] \tag{3}$$

where $\boldsymbol{E}_{cls}$ is the position encoding of cls token. $\boldsymbol{E}_{G}$ represents position encodings of genomics data corresponding to three tokens: RNA-Seq, miRNA, and DNA methylation. $\boldsymbol{E}_{cls}$ and $\boldsymbol{E}_{G}$ are initilized to random vectors. $\boldsymbol{E}_{P}$ represents position encodings of image patches. Following ViT [4], $\boldsymbol{E}_{P}$ is a 2D-aware position encoding of the image patches, which connects the X-axis encoding with the Y-axis encoding to model the spatial relationships between pixels in an image by introducing relative position and orientation information. As a result, the input dimension is $\boldsymbol{H}_{in} \in \mathbb{R}^{(1+3+L) \times d}$. Following that, $\boldsymbol{H}_{in}$ are fed into a cross-modal encoder implemented with a standard ViT. It consists of several Transformer blocks, each block contains a multi-head self-attention module and a position-wise fully connected feed-forward network. Moreover, the scaled dot-product is introduced into the multi-head self-attention module to compute adaptive weight from the input embedding $\boldsymbol{H}_{in}$. The processes in one Transformer block can be formulated as:

$$\begin{aligned} \text{head}_i &= \text{Attn}\left(W_Q^{(i)}\boldsymbol{H}_{in}, W_K^{(i)}\boldsymbol{H}_{in}, W_V^{(i)}\boldsymbol{H}_{in}\right) \\ &= \text{Softmax}\left(\frac{W_Q^{(i)}\boldsymbol{H}_{in}\boldsymbol{H}_{in}^{\top}W_K^{(i)\top}}{\sqrt{d}}\right)W_V^{(i)}\boldsymbol{H}_{in} \end{aligned} \tag{4}$$

$$\text{MSA}\left(\boldsymbol{H}_{in}\right) = \text{Concat}\left(\text{head}_1, ..., \text{head}_m\right)W_O \tag{5}$$

$$\boldsymbol{H}_{msa} = \text{LN}\left(\text{MSA}\left(\boldsymbol{H}_{in}\right) + \boldsymbol{H}_{in}\right) \tag{6}$$

$$\boldsymbol{H}_{block} = \text{LN}\left(\text{FFN}\left(\boldsymbol{H}_{msa}\right) + \boldsymbol{H}_{msa}\right) \tag{7}$$

where W_Q, W_K, and W_V are three learnable weight matrices multiplied by the queries $\boldsymbol{H}_{in}$, keys $\boldsymbol{H}_{in}$, and values $\boldsymbol{H}_{in}$. MSA, Concat, FFN, and LN denote the operations of multi-head self-attention, concatenation, feed-forward network, and layer normalization, respectively. m is the number of heads. As a result, the encodings $\boldsymbol{H}'_{enc} = [\boldsymbol{h}^{cls}, \boldsymbol{H}_{enc}]$ can be obtained by iterating through k rounds if there are k Transformer blocks in the encoder, where $\boldsymbol{h}^{cls}$ represents the encoding of cls token and $\boldsymbol{H}_{enc}$ denotes encodings corresponding to genomics and pathology.

ContraMAE Decoder. The cross-modal decoder aims to (i) decode the encodings $\boldsymbol{H}_{enc}$ and recover the original input data; and (ii) further integrate the multimodal information. Following MAE [5], we design a lightweight decoder with the same structure as the ContraMAE encoder but with fewer Transformer blocks and embedding dimensions. In the decoder, we also take Equation (4)-(7) to calculate each block's

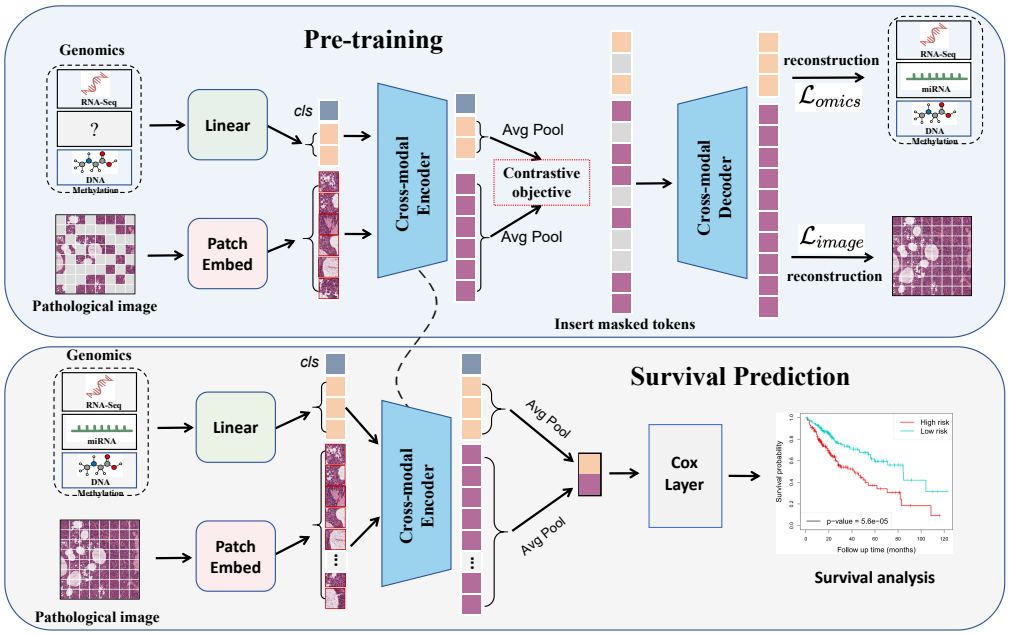

Fig. 1. Illustration of the contrastive alignment masked autoencoder (ContraMAE) framework.

representation, let $\boldsymbol{H}_{dec}$ denote the final representations generated from the decoder.

B. Pre-training Objectives

We introduce three objectives, pathology-genomics contrastive learning (PGC), genomics reconstruction (GR), and pathology reconstruction (PR), to pre-train our framework.

Pathology-Genomics Contrastive Learning. We introduce a PGC objective behind the ContraMAE encoder, which aims to align the modalities before the decoder and reconstruction. Concretely, we split the encodings $\boldsymbol{H}_{enc}$ from the encoder into two parts: $\boldsymbol{H}_{enc}^{(g)}$ and $\boldsymbol{H}_{enc}^{(p)}$, where $\boldsymbol{H}_{enc}^{(g)}$ corresponds to the encoding of genomics and $\boldsymbol{H}_{enc}^{(p)}$ corresponds to the encoding of pathology. Then, we aggregate the respective encoding via average pooling:

$$\boldsymbol{h}^{(g)} = \text{AvgPool}\left(\boldsymbol{H}_{enc}^{(g)}\right), \quad \boldsymbol{h}^{(g)} \in \mathbb{R}^{1\times d} \tag{8}$$

$$\boldsymbol{h}^{(p)} = \text{AvgPool}\left(\boldsymbol{H}_{enc}^{(p)}\right), \quad \boldsymbol{h}^{(p)} \in \mathbb{R}^{1\times d} \tag{9}$$

Assuming that a training batch has N pathology-genomics pairs, they can assemble N positive and $N^2 - N$ negative pathology-genomics pairs, where the positive pair has a value of 1 and the negative pair has a value of 0, which serve as the ground truth. Next, we calculate softmax-normalized pathology-to-genomics and genomics-to-pathology similarities within a training batch as follows:

$$\boldsymbol{s}_i^{p2g} = \frac{\exp\left(\boldsymbol{h}_i^{(p)\top}\boldsymbol{h}_i^{(g)}/\sigma\right)}{\sum_{j=1}^{N}\exp\left(\boldsymbol{h}_i^{(p)\top}\boldsymbol{h}_j^{(g)}/\sigma\right)} \tag{10}$$

$$\boldsymbol{s}_i^{g2p} = \frac{\exp\left(\boldsymbol{h}_i^{(g)\top}\boldsymbol{h}_i^{(p)}/\sigma\right)}{\sum_{j=1}^{N}\exp\left(\boldsymbol{h}_i^{(g)\top}\boldsymbol{h}_j^{(p)}/\sigma\right)} \tag{11}$$

where σ is a learnable temperature parameter. Finally, the cross-entropy loss function is used to compute contrastive losses between predicted similarities and the ground truth.

Genomincs Reconstruction and Pathology Reconstruction. We design two reconstruction objectives: GR and PR. GR randomly masks one of the three genomics types, and PR randomly masks a portion of the pathology image patches. The reconstruction tasks aim to predict the original genomics data and image patches. This makes multimodal data deeply interact, for example, pathology reconstruction partly relies on unmasked genomics data, which provides biomolecular information that pathology lacks. In addition, genomics reconstruction also partially depends on unmasked image patches, which include information regarding the microstructure of cells, tissues, and organs that genomics cannot provide. These two reconstruction objectives facilitate the mutual compensation of multimodal information, fully capturing intricate relationships between modalities.

Notably, the masked tokens in genomics data and image patches are initialized with zero vectors and then inserted into the encodings from the ContraMAE encoder. The order of all the encodings (which include visible tokens and masked tokens) is restored according to the original order.

We divide the final representations $\boldsymbol{H}_{dec}$ into genomics representation $\boldsymbol{h}_{dec}^{(g)}$ and pathology representation $\boldsymbol{h}_{dec}^{(p)}$. Subsequently, we employ two linear networks to map $\boldsymbol{h}_{dec}^{(g)}$ and $\boldsymbol{h}_{dec}^{(p)}$ to the same dimensions as $\boldsymbol{G}$ and $\boldsymbol{P}_{pat}$, as the reconstructed genomics data and pathological image patches. Finally, cross-entropy losses between reconstructed results and original input data are computed to train the model.

C. Survival prediction

In survival prediction, the ContraMAE decoder is discarded, the complete set of genomics data and pathological image patches are fed into the ContraMAE encoder without masking any tokens, and the representations of genomics data and pathology can be defined using Equation (8)-(9), respectively. We concatenate the representations of genomics data and pathology to yield the final representation. Subsequently, a linear network in the Cox layer is utilized to map the final representation to a single-node layer as the risk score for survival prediction, which can be written as:

$$z_i = \text{Linear}([\boldsymbol{h}^{(g)}, \boldsymbol{h}^{(p)}]) \tag{12}$$

Additionally, the average negative log partial likelihood [15] is used as the objective function in the Cox layer, which is formulated as follows:

$$\mathcal{L}_{surv} = -\frac{1}{n_E} \sum_{i:E_i=1} \left(z_i - \log \sum_{j:T_j>T_i} e^{z_j} \right) \tag{13}$$

where n_E is the total number of uncensored samples.

D. Implementation Details

We execute experiments using the PyTorch library on a Linux platform and run on a workstation with three NVIDIA A100 80 GB GPUs. The number of Transformer blocks, embedding dimension, and attention heads in the ContraMAE encoder and decoder are set to (12, 768, 12) and (8, 512, 16), respectively. We use the AdamW optimizer to train ContraMAE, with hyperparameters for learning rate, batch size, and training rounds differing in pre-training (5e-3, 100, 2000) and survival prediction (8e-4, 50, 80). The size of the representative region of the WSI is set to 1024×1024. Following MAE [5], we randomly mask 75% of pathology image patches during pre-training.

III. Experiments

TABLE I
GENOMIC FEATURES AT VARIOUS PREPROCESSING STAGES.

Genomic data	Initial genes	Preprocess		Final genes
		Methods	Genes	
RNA-Seq	59427	Rem, Var, Des	9249	300
miRNA	1881	Rem, Var, Des	1527	300
DNA methylation	485577	Kim, Var, Des	24889	300

Note: 'Kim' represents the missing value interpolation method (KNNImputer), 'Rem' denotes removing genes with missing values, 'Eli' signifies eliminating genes with zero variance, and 'Des' means using the DESeq2 tool for gene differential expression analysis.

A. Data Preprocessing and Evaluation Metrics

We experiment with five cancer datasets from TCGA [10], encompassing breast invasive carcinoma (BRCA), lung adenocarcinoma (LUAD), hepatocellular carcinoma (LIHC), stomach adenocarcinoma (STAD), and lower grade glioma (LGG). Each patient sample includes complete data types: RNA-Seq, miRNA, DNA methylation, and pathological images.

Genomics. The numbers of genomics features on various preprocessing stages are shown in Table I. For each cancer dataset, we directly remove the genes that appear missing values in RNA-Seq and miRNA, but employ KNNImputer [11] to interpolate the missing values in DNA methylation. Next, we eliminate genes with zero variance because they do not provide any information. After that, we perform differential gene expression analysis by Pydeseq2 [7] tool and genes importance analysis by random survival forest model (RSF) [6] to select the top 300 genes for each genomics type.

Pathology. The representative region of WSI is identified based on previous work [18]. Pathological images captured at $5\times$ magnification are cropped into overlapping tiles of 1024×1024 pixels using a sliding window strategy, with a stride of 100 pixels. The image density of each tile is computed by summing RGB values. Lastly, the tile with the highest density is considered the subregion with the highest diagnostic value and chosen as the representative WSI region.

Evaluation Metrics. We employ the concordance index (C-index) [2] as the evaluation metric with values ranging from 0 to 1. A higher C-index value indicates better predictive performance. A C-index value of 0.5 means that the model's predictions are equivalent to random chance. Moreover, we assess all investigated methods with 5-fold cross-validation splits on each cancer dataset.

B. Comparison with State-of-the-art Methods

We conduct experiments to compare ContraMAE with existing methods, which include the traditional methods, such as LASSO-Cox and EN-Cox, and deep learning-based methods, such as DeepCorrSurv, PathOmics, GPDBN, CAMR, MCAT, and CMTA. Original HIPT only supports pathology modality, we add a genomics-guided cross-attention, a Transformer, and attention pooling operations to enable HIPT to fuse pathological images with genomics data.

Table II illustrates all comparison models' experimental results. Compared to traditional methods, all deep learning-based methods obtain better performance. Notably, ContraMAE achieves superior performance on all five cancer datasets, with an overall C-index performance increase of 12.1% on LASSO-Cox, 9.9% on EN-Cox, 6.7% on GPDBN, 5.5% on CAMR, 2.1% on MCAT, 2.6% on HIPT, 2.5% on CMTA, 5.2% on DeepCorrSurv, and 2.9% on PathOmics, respectively. Moreover, we analyze the Kaplan-Meier curves of the LIHC dataset for further performance evaluation. In detail, we use the median of the risk scores as a risk indicator to divide LIHC patients into low-risk and high-risk groups. The Kaplan-Meier curves and corresponding log-rank test p-values of deep learning-based methods are presented in Figure 2. We can observe that ContraMAE better differentiates the survival curves between the low-risk and high-risk groups, and ContraMAE obtains competitive performance, with a p-value of 9.79e-09, which is comparable to the best performance of CAMR (p-value of 1.01e-09). The experimental results demonstrate that our ContraMAE achieves excellent performance and more precise discriminatory capability than existing methods.

TABLE II
PERFORMANCE COMPARISON OF CONTRAMAE AND OTHER METHODS USING THE C-INDEX VALUE ON FIVE CANCER DATASETS

Method	BRCA	LUAD	LIHC	STAD	LGG	Overall
LASSO-Cox [9]	0.582±0.023	0.590±0.055	0.618±0.033	0.542±0.045	0.725±0.054	0.599
EN-Cox [16]	0.622±0.034	0.615±0.046	0.626±0.042	0.550±0.052	0.753±0.063	0.621
GPDBN [14]	0.636±0.047	0.615±0.053	0.643±0.019	0.587±0.025	0.844±0.025	0.653
CAMR [15]	0.656±0.072	0.647±0.059	0.691±0.052	0.587±0.029	0.803±0.044	0.665
MCAT [2]	0.663±0.041	0.664±0.026	0.711±0.029	0.622±0.034	0.844±0.032	0.699
HIPT [1]	0.651±0.075	0.653±0.042	0.694±0.042	0.631±0.048	0.842±0.028	0.694
CMTA [19]	0.680±0.063	0.670±0.046	0.708±0.034	0.631±0.039	0.840±0.038	0.695
DeepCorrSurv [17]	0.659±0.018	0.662±0.032	0.700±0.048	0.609±0.049	0.828±0.034	0.668
PathOmics [3]	0.694±0.074	0.662±0.027	0.690±0.013	0.622±0.034	0.848±0.032	0.691
ContraMAE (Ours)	**0.702±0.023**	**0.679±0.028**	**0.725±0.019**	**0.641±0.020**	**0.853±0.041**	**0.720**

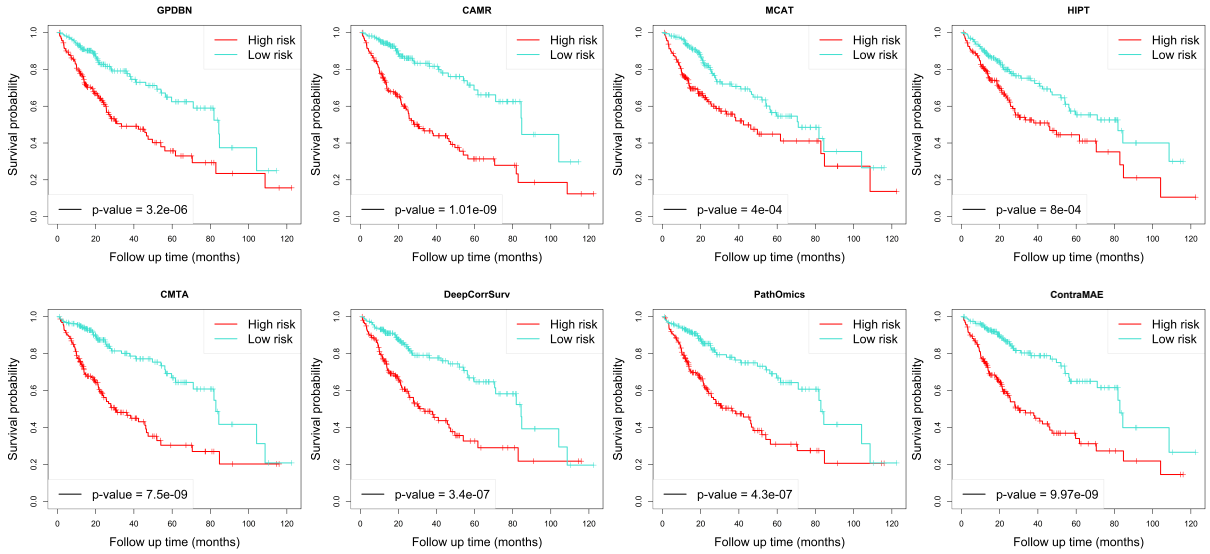

Fig. 2. Performance comparison of ContraMAE and other methods using Kaplan-Meier curves.

C. Ablation Study

Ablation study of modalities. To investigate the influences of various modalities on survival prediction performance, we perform an ablation study on BRCA, LUAD, and LIHC datasets by removing genomics or pathology modalities, respectively. As shown in Table III, the multimodal approach outperforms unimodal across three datasets, with a C-index improvement of 2.9% for genomics and 6.3% for pathology.

TABLE III
ABLATION STUDY OF MODALITIES.

Modality	BRCA	LUAD	LIHC	Overall
Genomics	0.652±0.042	0.660±0.032	0.706±0.028	0.673
Pathology	0.605±0.048	0.645±0.051	0.668±0.054	0.639
Multimodal	**0.702±0.023**	**0.679±0.028**	**0.725±0.019**	**0.702**

Ablation study of pre-training objectives. To assess the impact of different pre-training objectives on model performance, we conduct an ablation study comparing contrastive learning, reconstruction, and training from scratch. As shown in Table IV, the model without pre-training performs the worst. The combination of two reconstruction objectives outperforms contrastive learning, indicating that reconstruction contributes more to multimodal fusion. Moreover, using all three objectives yields the best performance, demonstrating the effectiveness of our pre-training strategy.

TABLE IV
ABLATION STUDY OF PRE-TRAINING OBJECTIVES.

Objectives	BRCA	LUAD	LIHC	Overall
From scratch	0.665±0.038	0.635±0.051	0.667±0.048	0.656
Contrastive learning	0.670±0.028	0.661±0.034	0.685±0.019	0.672
Two reconstructions	0.668±0.036	0.666±0.038	0.698±0.032	0.677
All objectives	**0.702±0.023**	**0.679±0.028**	**0.725±0.019**	**0.702**

D. Study of Risk Score Computing Methods

To explore the most effective risk computing method, we compare the following methods:

- M1: $\boldsymbol{H}_{enc}$ are average-pooled and linearly mapped to a single-node layer to produce the survival risk score.
- M2: $\boldsymbol{h}^{(g)}$ and $\boldsymbol{h}^{(p)}$ are linearly projected onto separate single-node layers, and their outputs are summed to generate the survival risk score.
- M3: The cls token encoding from ContraMAE encoder, $\boldsymbol{h}^{(cls)}$, is linearly mapped to a single-node layer to produce the survival risk score.

- M4 (Ours): Our survival risk score computing method.

As shown in Table V, splitting the encoder outputs into two branches for survival risk calculation (e.g., M2 and M4) improves performance. Notably, M4 achieves the highest C-index of 0.702, indicating it as the most effective fusion strategy for survival prediction.

TABLE V
THE STUDY OF RISK SCORE COMPUTATION METHODS IN THE COX LAYER.

Computing methods	BRCA	LUAD	LIHC	Overall
M1	0.668±0.056	0.636±0.047	0.702±0.056	0.669
M2	0.672±0.037	0.660±0.045	0.713±0.063	0.682
M3	0.668±0.029	0.642±0.036	0.692±0.052	0.667
M4	**0.702±0.023**	**0.679±0.028**	**0.725±0.019**	**0.702**

E. Study of Image Patch Size

The representative region size of the WSI is set as 1024×1024. However, various patch sizes, such as 64×64, 128×128, 256×256, and 512×512, should be explored for cropping. To identify the optimal patch size, we evaluate performance across different patch sizes. As shown in Figure 3, ContraMAE consistently achieves the best performance with 128×128 patches. Notably, both increasing the size to 256×256 or 512×512 and reducing it to 64×64 result in a significant drop in survival prediction performance.

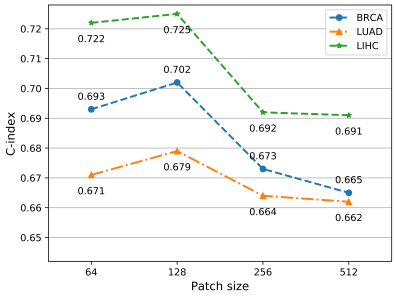

Fig. 3. The influence of patch sizes on the survival prediction performance.

IV. CONCLUSION

We propose ContraMAE, a contrastive alignment masked autoencoder framework, to integrate pathological images and genomics data for cancer survival prediction. We pre-train ContraMAE with three objectives to learn inherent task-agnostic relationships between modalities. Specifically, we first introduce a contrastive objective to align modalities and establish their intrinsic consistency. Then, we design two reconstruction objectives to capture the intricate interactions between modalities by mutually compensating for the information that each side does not possess. In survival prediction, we concatenate the pathology and genomics encodings from the ContraMAE encoder to generate the risk scores. Experimental results demonstrate that ContraMAE outperforms the existing methods on all five datasets sourced from TCGA.

ACKNOWLEDGMENT

This work is supported by two grants, No. 62162023 and No. KYQD(ZR)-21079.

REFERENCES

[1] Richard J Chen, Chengkuan Chen, Yicong Li, et al. Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16144–16155, 2022.

[2] Richard J Chen, Ming Y Lu, Wei-Hung Weng, et al. Multimodal co-attention transformer for survival prediction in gigapixel whole slide images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4025, 2021.

[3] Kexin Ding, Mu Zhou, Dimitris N Metaxas, et al. Pathology-and-genomics multimodal transformer for survival outcome prediction. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 622–631, 2023.

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020.

[5] Kaiming He, Xinlei Chen, Saining Xie, et al. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022.

[6] Hemant Ishwaran, Udaya B Kogalur, Eugene H Blackstone, et al. Random survival forests. *The Annals of Applied Statistics*, 2(3):841–860, 2008.

[7] Boris Muzellec, Maria Teleńczuk, Vincent Cabeli, et al. Pydeseq2: a python package for bulk rna-seq differential expression analysis. *bioRxiv*, 2022.

[8] Hyuna Sung, Jacques Ferlay, Rebecca L Siegel, et al. Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, 71(3):209–249, 2021.

[9] Robert Tibshirani. The lasso method for variable selection in the cox model. *Statistics in Medicine*, 16(4):385–395, 1997.

[10] Katarzyna Tomczak, Patrycja Czerwińska, and Maciej Wiznerowicz. Review the cancer genome atlas (tcga): an immeasurable source of knowledge. *Contemporary Oncology/Współczesna Onkologia*, 2015(1):68–77, 2015.

[11] Olga Troyanskaya, Michael Cantor, Gavin Sherlock, et al. Missing value estimation methods for dna microarrays. *Bioinformatics*, 17(6):520–525, 2001.

[12] Suixue Wang, Xiangjun Hu, and Qingchen Zhang. Hc-mae: Hierarchical cross-attention masked autoencoder integrating histopathological images and multi-omics for cancer survival prediction. In *2023 IEEE International Conference on Bioinformatics and Biomedicine*, pages 642–647. IEEE, 2023.

[13] Suixue Wang, Shuling Wang, and Zhengxia Wang. A survey on multi-omics-based cancer diagnosis using machine learning with the potential application in gastrointestinal cancer. *Frontiers in Medicine*, 9:1109365, 2023.

[14] Zhiqin Wang, Ruiqing Li, Minghui Wang, et al. Gpdbn: deep bilinear network integrating both genomic data and pathological images for breast cancer prognosis prediction. *Bioinformatics*, 37(18):2963–2970, 2021.

[15] Xingqi Wu, Yi Shi, Minghui Wang, et al. Camr: cross-aligned multi-modal representation learning for cancer survival prediction. *Bioinformatics*, 39(1):btad025, 2023.

[16] Yi Yang and Hui Zou. A cocktail algorithm for solving the elastic net penalized cox's regression in high dimensions. *Statistics and its Interface*, 6(2):167–173, 2013.

[17] Jiawen Yao, Xinliang Zhu, Feiyun Zhu, et al. Deep correlational learning for survival prediction from multi-modality data. In *International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pages 406–414, 2017.

[18] Kun-Hsing Yu, Ce Zhang, Gerald J Berry, et al. Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features. *Nature Communications*, 7(1):12474, 2016.

[19] Fengtao Zhou and Hao Chen. Cross-modal translation and alignment for survival analysis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21485–21494, 2023.