2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) | 979-8-3503-3748-8/23/$31.00 ©2023 IEEE | DOI: 10.1109/BIBM58861.2023.10385635

HC-MAE: Hierarchical Cross-attention Masked Autoencoder Integrating Histopathological Images and Multi-omics for Cancer Survival Prediction

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

2nd Xiangjun Hu
School of Information and Communication Engineering
Hainan University
Haikou, China
huxiangjun@hainanu.edu.cn
The co-first author

3rd Qingchen Zhang
School of Computer Science and Technology
Hainan University
Haikou, China
Software College
Shenyang Normal University
Shenyang, China
zhangqingchen@hainanu.edu.cn
The corresponding author

***Abstract*—Accurate cancer survival prediction enables clinicians to tailor treatment regimens based on individual patient prognoses, effectively mitigating over-treatment and inefficient medical resource allocation. Recently, the integration of histopathological images and multi-omics data, together with deep learning, has become increasingly applied to predict cancer survival. However, current deep learning-based integration methods ignore the spatial relationships across various fields of view within gigapixel histopathological images, since they mainly focus on a specific field of view. Inspired by the hierarchical image pyramid transformer (HIPT), we propose a hierarchical cross-attention masked autoencoder (HC-MAE) to integrate histopathological images and multi-omics data for cancer survival prediction. Specifically, HC-MAE aggregates the representations learned from different fields of view, effectively capturing the fine-grained details and the spatial relationships within histopathological images. We conduct experiments to compare the HC-MAE method with current state-of-the-art methods on six cancer datasets sourced from The Cancer Genome Atlas (TCGA). The experimental results demonstrate that HC-MAE achieves superior performance on five out of six cancer datasets, significantly outperforming the compared methods. The code is available at https://github.com/SuixueWang/HC-MAE.**

***Index Terms*—Hierarchical Cross-attention, Masked Autoencoder, Histopathological Image, Multi-omics, Survival Prediction**

I. Introduction

Cancer is widely acknowledged as a significant global public health threat, with increasing incidence and mortality rates. In 2020, it was reported that there were 19.3 million newly diagnosed cancer cases and approximately 10.0 million cancer-related deaths [1]. Among them, lung cancer was the most common cause of cancer mortality, accounting for approximately 1.8 million deaths, followed by cancers of the colorectum, liver, and stomach, as well as the breast in women [1]. Currently, surgery, radiotherapy, chemotherapy, and immunotherapy are the primary treatment regimens for cancer [2]. Generally, accurate survival prediction plays a crucial role in assisting clinicians to select the most effective treatment regimen, which prevents over-treatment and optimizes the allocation of medical resources. However, cancer is a complex disease that poses a substantial challenge toward achieving precise diagnoses and accurate prognoses [3].

In recent years, deep learning-based methods have made great progress in cancer survival prediction, some of which utilize the abundant and intricate details extracted from gigapixel whole-slide images (WSIs). For example, Lu et al. [10] developed an attention-based deep-learning approach, CLAM, which employs weakly-supervised learning to autonomously recognize subregions with high diagnostic value, this is followed by introducing instance-level clustering in the identified regions to achieve improved performance. Chen et al. [11] proposed a HIPT, an enhanced vision transformer (ViT) architecture that utilizes hierarchical patch structures and employs two stages of DINO-based [12] self-supervised learning to capture high-resolution image representations.

Moreover, some deep learning-based integration methods that incorporate histopathological images and multi-omics data such as RNA-Seq, miRNA, and DNA methylation, have been proposed for predicting survival outcomes [4]. Due to differences among image feature extraction techniques, deep

979-8-3503-3748-8/23/$31.00 ©2023 IEEE

learning methods are primarily categorized into two subtypes: hand-crafted and deep learning-based approaches. The former methods typically utilize the CellProfiler tool to extract image features such as sizes, shapes, brightness levels, textures, and intensity distributions [7]. For instance, Wang et al. [5] and Wu et al. [6] extracted image features through the CellProfiler tool first and then feed the image features to the deep learning-based framework for integrating with genomics data. In comparison with handcrafted methods, deep learning-based methods are garnering increasing attention. For example, Yao et al. [8] proposed a deep correlational survival model, DeepCorrSurv, which consists of two sub-networks for integrating multi-omics and pathological images, where image patches with 1024 × 1024 pixels are fed into a CNN for learning the representations. Chen et al. [9] developed a more efficient framework named MCAT, which divides WSIs into small patches of size 256 × 256 and attains initial representations via a pre-trained ResNet50 model, subsequently, the genomic-guided co-attention and attention pooling mechanisms are adopted for the integration of multimodal data.

Most of the current deep learning-based integration approaches primarily focus on a single perspective within gigapixel WSIs, failing to make full use of the information contained in complete WSIs. Aiming at this drawback, we propose a hierarchical cross-attention masked autoencoder (HC-MAE) to integrate histopathological images and multi-omics data for cancer survival prediction. Specifically, the HC-MAE hierarchically aggregates the representations obtained from various fields of view, thereby capturing the fine-grained details and spatial relationships within histopathological images. During the pre-training phase, we independently pre-train two stages of MAEs, namely MAE-Patch and MAE-Region. This involves using image patches with varying fields of view as inputs and hierarchically merging them from the bottom to the top. For survival prediction, we employ an omics-guided cross-attention mechanism to fuse the representations aggregated from MAE-Region, along with slide-level image patch embeddings and multi-omics data for computing survival risk scores. We conduct experiments to compare the performance of our proposed method, the HC-MAE, with that of the state-of-the-art methods on six cancer datasets. The experimental results demonstrate that the HC-MAE achieves the highest C-index values among the compared methods on five cancer cohorts, indicating significant improvement.

In summary, our contributions are as follows:

- We propose a hierarchical cross-attention masked autoencoder (HC-MAE) to integrate histopathological images and multi-omics for accurate cancer survival prediction.
- To perform survival prediction, we present an omics-guided cross-modal attention mechanism that simultaneously attends to the representations aggregated from lower levels, the current slide-level image patch embeddings, and multi-omics data.
- We carry out complete ablation experiments to analyze the survival prediction performance achieved with different pathological image fields of view, multi-omics, and multimodal data, proving the effectiveness of our proposed HC-MAE in terms of capturing image representations and integrating multimodal data.

II. Method

The overview of our proposed HC-MAE is shown in Fig. 1. We start by pre-training two stages of MAEs to learn high-quality image representations with sizes of 256 × 256 and 4096 × 4096. To perform survival prediction, we integrate the representations aggregated from lower levels, along with slide-level image patch embeddings and multi-omics data, to compute a survival risk score for survival prediction. The following subsections describe the HC-MAE in detail.

A. Problem formulation

Given a gigapixel WSI, CLAM [10] is utilized to automatically filter out the background area, segment the tissue region of the WSI, and then crop it into many sub-regions with sizes of 4096×4096. A WSI with sub-regions possessing high diagnostic value can be expressed as $\mathbf{I}_{Slide}$. $\mathbf{I}_{Slide}$ is cropped in a layer-by-layer manner into patches of various scales, including region-level (4096×4096), patch-level (256×256), and cell-level (16×16) patches, this is written as:

$$\begin{aligned} \mathbf{I}_{Slide} &= \left\{\mathbf{I}_{4096}^{(i)}\right\}_{i=1}^{M}, \quad \mathbf{I}_{4096}^{(i)} \in \mathbb{R}^{4096\times4096\times3} \\ \mathbf{I}_{4096}^{(i)} &= \left\{\mathbf{I}_{256}^{(j)}\right\}_{j=1}^{256}, \quad \mathbf{I}_{256}^{(j)} \in \mathbb{R}^{256\times256\times3} \\ \mathbf{I}_{256}^{(j)} &= \left\{\mathbf{I}_{16}^{(k)}\right\}_{k=1}^{256}, \quad \mathbf{I}_{16}^{(k)} \in \mathbb{R}^{16\times16\times3} \end{aligned} \tag{1}$$

where M is the number of region-level patches in $\mathbf{I}_{Slide}$. $\mathbf{I}_{4096}^{(i)}$, $\mathbf{I}_{256}^{(j)}$, and $\mathbf{I}_{16}^{(k)}$ represent the image patches at the region-level, patch-level, and cell-level, respectively. The sequence lengths of both $\mathbf{I}_{4096}^{(i)}$ and $\mathbf{I}_{256}^{(j)}$ are 256.

Furthermore, we employ a lightweight CNN to embed the image patches of $\mathbf{I}_{16}^{(k)}$, $\mathbf{I}_{256}^{(j)}$ and $\mathbf{I}_{4096}^{(i)}$. The kernel size of the lightweight CNN is set to the corresponding patch size, and the number of channels is equal to the dimensionality of the embedding, which is set to 384 in our case. Due to the computational consumption incurred when embedding $\mathbf{I}_{4096}^{(i)}$, we downsample $\mathbf{I}_{4096}^{(i)}$ by a factor of 16 to $\mathbf{I}_{Down-4096}^{(i)} \in \mathbb{R}^{256\times256\times3}$ before feeding it into the lightweight CNN. Despite the reduced resolution of $\mathbf{I}_{P-4096}^{(i)}$, it still captures coarse-grained spatial information in this field of view. Therefore, the patch embeddings with added position embeddings corresponding to various fields of view, $\mathbf{I}_{16}$, $\mathbf{I}_{256}$, and $\mathbf{I}_{Down-4096}$, are written as $\mathbf{x}_{P-16} \in \mathbb{R}^{256\times384}$, $\mathbf{x}_{P-256} \in \mathbb{R}^{256\times384}$, and $\mathbf{x}_{P-4096} \in \mathbb{R}^{M\times384}$, respectively. In addition, $\mathbf{x}_{omics} \in \mathbb{R}^{3\times300}$ is used to denote the multi-omics data after preprocessing.

B. Hierarchical cross-attention pre-training

In this study, we pre-train two stages of MAEs, including MAE-Patch and MAE-Region, where MAE-Patch has the same model structure as the standard MAE [13] and MAE-Region is a variant of the base MAE.

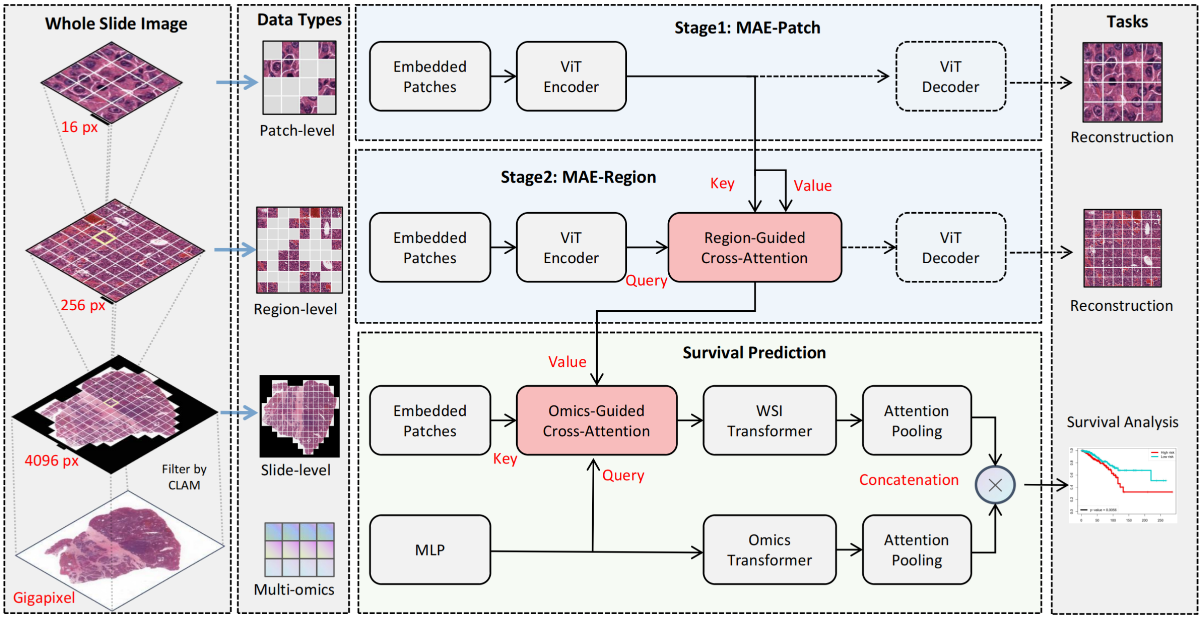

Fig. 1. The framework of the hierarchical cross-attention masked autoencoder integrating histopathological images and multi-omics.

1) Stage 1 with MAE-Patch: A given patch-level image patch $\mathbf{I}_{256}^{(j)}$ is divided into 256 cell-level patches $\mathbf{I}_{16}$ and further embedded as $\mathbf{x}_{P-16}$. Moreover, the cell-level patches are used to pre-train the MAE-Patch module by randomly masking 75% of the tokens and reconstructing the original image. Specifically, MAE-Patch contains two asymmetrical sub-modules, a ViT encoder, and a ViT decoder, both of which consist of a series of Transformer blocks. After pre-training, the ViT encoder in MAE-Patch generates a representation for each patch-level image $\mathbf{I}_{256}^{(j)}$, which is calculated as follows:

$$\mathbf{x}_{A-256}^{(j)} = \mathit{AvgPool}\left(\mathit{ViT\text{-}P}\left(\mathbf{x}_{P-16}\right)\right) \tag{2}$$

where $\mathit{ViT\text{-}P}\left(\cdot\right)$ is the ViT encoder in MAE-Patch, and $\mathit{AvgPool}\left(\cdot\right)$ signifies the average pooling operation.

2) Stage 2 with MAE-Region: Similarly, a region-level image $\mathbf{I}_{4096}^{(i)}$ is divided into 256 patch-level patches $\mathbf{I}_{256}$ and embedded as $\mathbf{x}_{P-256}$. Different from MAE-Patch, the MAE-Region module introduces a region-guided cross-attention behind the ViT encoder to attend to the representations aggregated at the patch-level $\mathbf{x}_{A-256}$, along with the current region-level unmasked patch embeddings $\mathbf{x}_{P-256}$. Essentially, the fine-grained information obtained from MAE-Patch plays a complementary role in reconstructing the masked image patches. As a result, this process enhances the ability of the ViT encoder in MAE-Region to learn a more effective representation. The formulations of the encoding and region-guided cross-attention steps are written as follows:

$$\mathbf{x}_{E-256} = \mathit{ViT\text{-}R}\left(\mathbf{x}_{P-256}\right) \tag{3}$$

$$\begin{aligned}\mathbf{x}_{C-256} &= \mathit{CrossAtt}\left(\mathbf{w}_q\mathbf{x}_{P-256}, \mathbf{w}_k\mathbf{x}_{E-256}, \mathbf{w}_v\mathbf{x}_{E-256}\right)\\ &= \mathit{Softmax}\left(\frac{\mathbf{w}_q\mathbf{x}_{P-256}\mathbf{x}_{E-256}^T\mathbf{w}_k^T}{\sqrt{d_k}}\right)\mathbf{w}_v\mathbf{x}_{E-256}\end{aligned} \tag{4}$$

where $\mathit{ViT\text{-}R}\left(\cdot\right)$ is the ViT encoder in MAE-Region. $\mathbf{w}_q, \mathbf{w}_k, \mathbf{w}_v \in \mathbb{R}^{d_k \times d_k}$ are trainable weight matrices multiplied by the queries $\mathbf{x}_{P-256}$ and key-value pair ($\mathbf{x}_{E-256}$, $\mathbf{x}_{E-256}$), where d_k is the dimensionality of $\mathbf{x}_{E-256}$, 384.

After finishing pre-training, the aggregated representation of image $\mathbf{I}_{4096}^{(i)}$ is produced via the operations of the ViT encoder in MAE-Region, cross-attention, and average pooling. The average pooling process is shown as follows:

$$\mathbf{x}_{A-4096}^{(i)} = \mathit{AvgPool}\left(\mathbf{x}_{C-256}\right), \mathbf{x}_{A-4096} \in \mathbb{R}^{M \times 384} \tag{5}$$

C. Survival prediction

Initially, we use a fully-connected layer to map the given multi-omics data $\mathbf{x}_{omics}$ to a dimension that aligns with the dimensions utilized in the ViT encoder and Transformer, as shown below:

$$\mathbf{x}_O = \mathit{MLP}\left(\mathbf{x}_{omics}\right) \tag{6}$$

For each whole-slide image $\mathbf{I}_{Slide}$, we utilize an omics-guided cross-attention mechanism to integrate the multi-omics data $\mathbf{x}_O$, along with the current slide-level patch embeddings $\mathbf{x}_{P-4096}$ and the representations aggregated from the region-level $\mathbf{x}_{A-4096}$, which is written as:

$$\begin{aligned}\mathbf{x}_C &= \mathit{CrossAtt}\left(\mathbf{w}_q\mathbf{x}_O, \mathbf{w}_k\mathbf{x}_{P-4096}, \mathbf{x}_v\mathbf{x}_{A-4096}\right)\\ &= \mathit{Softmax}\left(\frac{\mathbf{w}_q\mathbf{x}_O\mathbf{x}_{P-4096}^T\mathbf{w}_k^T}{\sqrt{d_k}}\right)\mathbf{w}_v\mathbf{x}_{A-4096}\end{aligned} \tag{7}$$

where $\mathbf{w}_q, \mathbf{w}_k, \mathbf{w}_v \in \mathbb{R}^{d_k \times d_k}$ are trainable weight matrices multiplied by the queries $\mathbf{x}_O$, keys $\mathbf{x}_{P-4096}$, and values $\mathbf{x}_{A-4096}$, respectively. d_k is set to the dimensionality of $\mathbf{x}_{P-4096}$, which is 384. The integrated result $\mathbf{x}_C \in \mathbb{R}^{3 \times 384}$.

To learn the multimodal integrated features between WSIs and multi-omics data in depth, we input $\mathbf{x}_C$ into the WSI

Transformer and $\mathbf{x}_O$ into the omics Transformer, both consisting of 12 blocks, as follows:

$$\mathbf{x}_{CT} = \mathit{Transformer}_{WSI}\left(\mathbf{x}_C\right) \tag{8}$$

$$\mathbf{x}_{OT} = \mathit{Transformer}_{Omics}\left(\mathbf{x}_O\right) \tag{9}$$

Following the Transformer, we employ a global attention pooling layer to aggregate the embedding vectors in $\mathbf{x}_{CT}$ and $\mathbf{x}_{OT}$ respectively. In detail, $\mathbf{x}_{CT}$ and $\mathbf{x}_{OT}$ are fed to a fully connected layer and a softmax operator, respectively, adaptively calculating a weighted sum of all embedding vectors to form bag-level representations:

$$\mathbf{x}_{CT-global} = \mathit{Softmax}\left(\mathit{MLP}\left(\mathbf{x}_{CT}\right)\right) \cdot \mathbf{x}_{CT} \tag{10}$$

$$\mathbf{x}_{OT-global} = \mathit{Softmax}\left(\mathit{MLP}\left(\mathbf{x}_{OT}\right)\right) \cdot \mathbf{x}_{OT} \tag{11}$$

In the last integration step, we concatenate the bag-level representations $\mathbf{x}_{CT-global}$ and $\mathbf{x}_{OT-global}$ to produce the final representation that represents the overall integration features of the input multimodel data, including histopathological images and multi-omics data, this is shown as follows:

$$\mathbf{x} = \mathbf{x}_{CT-global} \oplus \mathbf{x}_{OT-global} \tag{12}$$

where $\oplus$ signifies the concatenation operator.

When learning the final multimodal representation $\mathbf{x}^{(i)}$ for the *i-th* patient sample, we send $\mathbf{x}^{(i)}$ to a sigmoid-activated fully-connected layer, generating a survival risk score $\boldsymbol{\beta}\mathbf{x}^{(i)}$ for survival analysis, where $\boldsymbol{\beta}$ is a trainable weight in the fully-connected layer. Furthermore, we adopt the average negative log partial likelihood as the objective function of our proposed HC-MAE, which is written as:

$$\mathbf{loss} = -\frac{1}{n_E} \sum_{i:E_i=1} \left(\boldsymbol{\beta}\mathbf{x}^{(i)} - \log \sum_{j:T_j>T_i} e^{\boldsymbol{\beta}\mathbf{x}^{(j)}} \right) \tag{13}$$

where the values of E_i and T_i are survival status and survival time for each patient, respectively, while $\boldsymbol{\beta}\mathbf{x}^{(i)}$ is the fully-connected layer used to predict the survival risk score. n_E represents the total number of uncensored samples.

III. Experiments and results analysis

TABLE I
The number of patients, region-level images (4096×4096 size), and patch-level image patches (pixel of 256×256).

Dataset	# Patients	# Region-level	# Patch-level (million)
LIHC	323	65,549	16.8
BRCA	724	92,634	23.7
LUAD	374	59,148	15.1
COAD	250	37,582	9.6
LGG	451	60,804	15.6
STAD	298	50,047	12.8
Total	2,420	365,764	93.6

A. Data description and implementation details

We utilize cancer datasets sourced from The Cancer Genome Atlas (TCGA) [14], a public consortium for cancer data that provides high-resolution WSIs and multi-omics data, with annotated censorship statuses and survival durations. In this work, we conduct experiments on the following six cancer types: hepatocellular carcinoma (LIHC), breast invasive carcinoma (BRCA), lung adenocarcinoma (LUAD), colon adenocarcinoma (COAD), lower grade glioma (LGG), and stomach adenocarcinoma (STAD). When dealing with patient samples, we only consider a sample if it contains both a WSI and corresponding multi-omics data. In addition, patients with survival times of less than 30 days and those with missing follow-up data are excluded. For each patient sample, if there is more than one WSI, we randomly select one for our work. For all the collected gigapixel WSIs, we start by cropping each WSI into patches with 4096×4096 pixels, subsequently, we employ the CLAM model to filter meaningless background regions and identify sub-regions with high diagnostic value. We divide each identified sub-region with 4096×4096 pixels (region-level image patches) into 256 small patches with 256×256 pixels (patch-level image patches). After performing preprocessing, the numbers of patients, identified sub-regions, and small patches are shown in Table I.

Additionally, when working with multi-omics data including RNA-Seq, miRNA, and DNA methylation, for each omics type, we first employ the K-nearest neighbor interpolation method [15] to fill in missing values. Afterward, we calculate the variance of each gene across all patient samples and remove the genes with zero variance. Subsequently, we perform differential gene expression analysis using the pydeseq2 package [16] to select the genes that exhibit significant changes. Finally, we use the random survival forest (RSF) to calculate the feature importance of each gene across all patient samples and retrain only the top 300 genes for survival analysis.

The HC-MAE is implemented in PyTorch and runs on a workstation equipped with 2 80-GB NVIDIA A100 GPUs. The ViT encoders used in the MAE-Patch and MAE-Region modules, along with the Transformers employed for survival prediction, consist of 12 transformer blocks, 12 heads, and 384-dimensional embeddings. During pre-training, we adopt Adam optimization with a weight decay of 1e-2 and a learning rate of 1e-3. The batch size in MAE-Patch is 500, while in MAE-Region, the batch size is 10 with 50 gradient accumulation steps. The epochs in MAE-Patch and MAE-Region are set to 10. During survival prediction, the learning rate and the epoch are set to 6e-4 and 80, respectively. Since the samples have different bag sizes, we set the batch size to 1, with 60 gradient accumulation steps.

B. Evaluation metric

We use the concordance index (C-index) as the metric and train all methods with 5-fold cross-validation on each cancer dataset [5] [6]. A higher C-index indicates better predictive performance, while a C-index of 0.5 suggests that predictions are equivalent to those produced by random chance.

TABLE II
PERFORMANCE COMPARISON OF OUR PROPOSED HC-MAE WITH STATE-OF-THE-ART METHODS USING THE C-INDEX VALUE ON SIX CANCER DATASETS

Feature extraction of WSI	Method	LIHC	BRCA	LUAD	COAD	LGG	STAD
Hand-craft	GPDBN	0.643±0.019	0.636±0.047	0.615±0.053	0.593±0.081	0.844±0.025	0.587±0.025
	CAMR	0.691±0.052	0.656±0.072	0.647±0.059	0.606±0.074	0.803±0.044	0.587±0.029
Deep Learning	DeepCorrSurv	0.700±0.048	0.659±0.018	0.662±0.032	0.549±0.026	0.828±0.034	0.609±0.049
	MCAT	0.711±0.029	0.663±0.041	0.664±0.026	0.691±0.132	0.844±0.032	0.622±0.034
	HIPT (WSI only)	0.668±0.031	0.640±0.055	0.635±0.045	0.612±0.033	0.810±0.031	0.612±0.053
	HIPT (Integration)	0.694±0.042	0.651±0.075	0.653±0.042	0.635±0.026	0.842±0.028	**0.631±0.048**
	HC-MAE (Ours)	**0.737±0.031**	**0.695±0.026**	**0.691±0.044**	**0.703±0.083**	**0.851±0.033**	0.623±0.045

C. Comparison with the state-of-the-art methods

To assess the performance of our proposed HC-MAE, we conducted a comparative analysis with two types of state-of-the-art methods: GPDBN and CAMR, which are based on hand-crafted image feature extraction methods, and DeepCorrSurv, MCAT, and HIPT, which employ deep learning-based extraction methods. Additionally, the original HIPT only supports WSIs as input, so we add a few operations, such as omics-guided cross-attention, a Transformer, and attention pooling, to integrate WSIs with multi-omics data.

The C-index values produced by all methods on all six cancer datasets are shown in Table II. In terms of the C-index values, the integration methods that use deep learning for image feature extraction generally perform better than those using hand-crafted extraction. Comparing HC-MAE with the state-of-the-art models, HC-MAE demonstrates superior performance with the highest C-index values for 5 out of 6 cancer datasets. Especially on the BRCA dataset, HC-MAE attains the top C-index value of 0.695 ± 0.026, outperforming the second-best method, MCAT, by 3.2% and exceeding the lowest-performing method, GPDBN, by 5.9%.

Furthermore, the BRCA patients are stratified into low-risk and high-risk groups using the median of the predicted risk scores as a risk indicator. The Kaplan-Meier curves produced by all investigated methods, along with their corresponding log-rank test p-values, are presented in Fig. 2. Among the methods utilizing hand-crafted extraction, CAMR yields a superior log-rank test p-value of 0.0035 compared to that of GPDBN (0.0048). However, the low-risk and high-risk curves of the CAMR method intersect and are not clearly distinguishable. Among the integration methods utilizing deep learning-based image feature extraction, a similar intersection phenomenon is observed for DeepCorrSurv. In addition, MCAT, HIPT, and our proposed HC-MAE obtain log-rank test P-values of 0.0056, 0.0079, and 2.1e-7, respectively. Notably, HC-MAE achieves superior performance to that of all other investigated methods, exhibiting the greatest capability to distinguish between the low-risk and high-risk groups.

D. Ablation study of the HC-MAE

To evaluate the effectiveness of our proposed HC-MAE, we conduct a complete ablation study to compare the performance achieved for various module configurations. The module configurations are described below:

- MAE-Pat: This is the ViT encoder used in MAE-Patch.
- MAE-Reg: The ViT encoder used in MAE-Region uses only region-level images for pre-training.
- MAE-PatReg: The ViT encoder used in MAE-Region, aggregates region-level patch embeddings with the representations gathered from the MAE-Patch module.
- HC-RegSli: Hierarchically cross-attending the slide-level patch embeddings with the representations aggregated from MAE-Reg.
- HC-PatRegSli: Hierarchically cross-attending the slide-level patch embedding with the representations aggregated from MAE-PatReg.
- Trans-MO: The Transformer uses only multi-omics data.
- HC-MAE: This is our proposed complete model.

TABLE III
ABLATION EXPERIMENTS OF THE HC-MAE ON LIHC DATASET

Data type	Method	C-index
Image	MAE-Pat	0.649±0.055
	MAE-Reg	0.614±0.025
	MAE-PatReg	0.674±0.042
	HC-RegSli	0.641±0.066
	HC-PatRegSli	0.711±0.032
Multi-omics	Trans-MO	0.700±0.045
Multimodal	HC-MAE	**0.737±0.031**

Table III shows the C-index values produced by the module configurations on the LIHC dataset. In comparison with MAE-Pat (0.649±0.055) and MAE-Reg (0.614±0.025), MAE-PatReg (0.674±0.042) attains a superior performance, which means that the cross-attention mechanism we introduce to the MAE can adequately capture the spatial relationship across different fields of view. Comparing MAE-Pat with MAE-Reg, MAE-Pat has a higher C-index value. Additionally, a similar observation is also obvious when contrasting HC-PatRegSli with HC-RegSli, indicating that patch-level images can provide fine-grained cell-level information and that the corresponding MAE-Patch method can produce high-quality image patch representations. Additionally, the Trans-MO method (0.700±0.045), which utilizes multi-omics data, produces a result close to that of the HC-PatRegSli method which is only based on entire WSIs. Notably, the complete model, the HC-MAE, achieves the best C-index value of 0.737±0.031, demonstrating the effectiveness of our proposed method fusing histopathological images and multi-omics data.

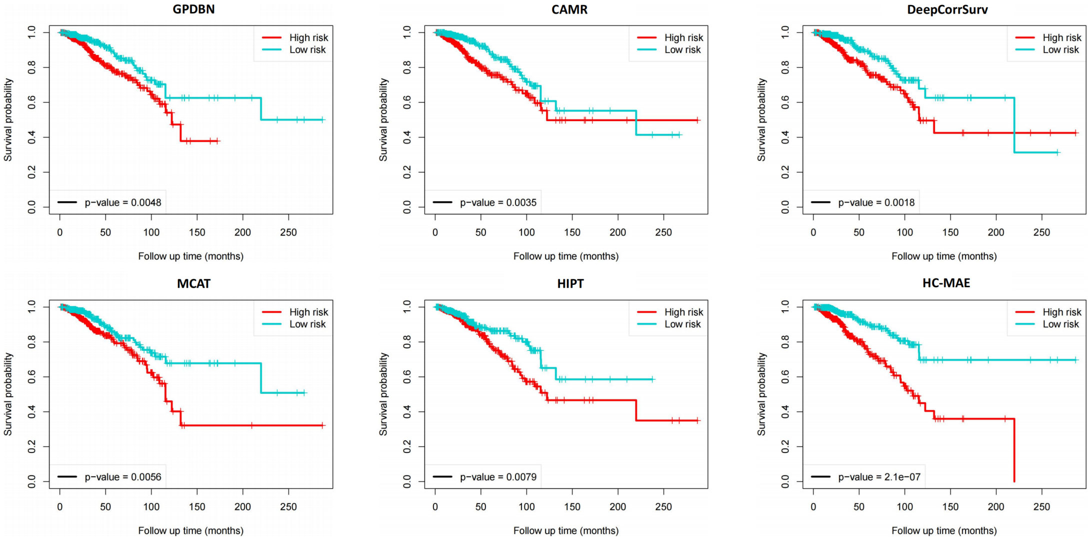

Fig. 2. Performance comparison of HC-MAE and other methods using Kaplan–Meier curve on BRCA dataset.

IV. Conclusions

In this work, we propose a hierarchical cross-attention masked autoencoder to integrate histopathological images and multi-omics for cancer survival prediction. The notable property of our proposed HC-MAE is that it hierarchically aggregates the representations learned from different fields of view, effectively capturing the fine-grained details and spatial relationships within histopathological images. During pre-training, we pre-train two stages of MAEs, namely MAE-Patch and MAE-Region. Especially in MAE-Region, we adopt a region-guided cross-attention mechanism that leverages two levels of representations for reconstructing the masked images, allowing the ViT encoder in the MAE to capture high-quality region-level representations. During survival prediction, the representations aggregated from MAE-Region are fused with slide-level image patch embeddings and multi-omics data by using an omics-guided cross-attention mechanism, to produce the survival risk scores. The comparison experiment proves that our proposed HC-MAE model outperforms the state-of-the-art methods in terms of cancer survival prediction, while the ablation experiment validates the effectiveness of HC-MAE with regard to capturing image representations and integrating multimodal data.

Acknowledgment

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

References

[1] Sung H, Ferlay J, Siegel R L, Laversanne M, Soerjomataram I, 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.

[2] Yagawa Y, Tanigawa K, Kobayashi Y, Yamamoto M, “Cancer immunity and therapy using hyperthermia with immunotherapy, radiotherapy, chemotherapy, and surgery,” J. Cancer Metastasis Treat 3(10):218, 2017.

[3] Beck, Andrew H, “Open access to large scale datasets is needed to translate knowledge of cancer heterogeneity into better patient outcomes,” PLoS Medicine, 12(2):e1001794, 2015.

[4] Cui C, Yang H, Wang Y, Zhao S, Asad Zet, et al., “Deep multi-modal fusion of image and non-image data in disease diagnosis and prognosis: a review,” Progress in Biomedical Engineering, 2023.

[5] Wang Z, Li R, Wang M, Li A, “GPDBN: deep bilinear network integrating both genomic data and pathological images for breast cancer prognosis prediction,” Bioinformatics, 37(18):2963-2970, 2021.

[6] Wu X, Shi Y, Wang M, Li A, “CAMR: cross-aligned multimodal representation learning for cancer survival prediction,” Bioinformatics, 39(1):btad025, 2023.

[7] McQuin C, Goodman A, Chernyshev V, Kamentsky L, Cimini B A, et al., “CellProfiler 3.0: Next-generation image processing for biology,” PLoS Biology, 16(7):e2005970, 2018.

[8] Yao J, Zhu X, Zhu F, Huang J, “Deep correlational learning for survival prediction from multi-modality data,” International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI), 406-414, 2017.

[9] Chen R J, Lu M Y, Weng W H, Chen T Y, Williamson D F, et al., “Multimodal co-attention transformer for survival prediction in gigapixel whole slide images,” Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 4015-4025, 2021.

[10] Lu M Y, Williamson D F K, Chen T Y, Chen R J, Barbieri M, et al., “Data-efficient and weakly supervised computational pathology on whole-slide images,” Nature Biomedical Engineering, 5(6):555-570, 2021.

[11] Chen R J, Chen C, Li Y, Chen T Y, Trister A D, et al., “Scaling vision transformers to gigapixel images via hierarchical self-supervised learning,” Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 16144-16155, 2022.

[12] Caron M, Touvron H, Misra I, Jégou H, Mairal J, et al., “Emerging properties in self-supervised vision transformers,” Proceedings of the IEEE/CVF international conference on computer vision (ICCV), 9650-9660, 2021.

[13] He K, Chen X, Xie S, Li Y, Dollár P, et al., “Masked autoencoders are scalable vision learners,” Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR), 16000-16009, 2022.

[14] Tomczak K, Czerwińska P, Wiznerowicz M, “Review The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge,” Contemporary Oncology/Współczesna Onkologia, 2015(1):68-77, 2015.

[15] Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, et al., “Missing value estimation methods for DNA microarrays,” Bioinformatics, 17(6):520-525, 2001.

[16] Muzellec B, Telenczuk M, Cabeli V, Andreux M, “PyDESeq2: a python package for bulk RNA-seq differential expression analysis,” BioRxiv, 2022.