Multimodal Optimal Transport-based Co-Attention Transformer with Global Structure Consistency for Survival Prediction

Yingxue Xu Hao Chen*
The Hong Kong University of Science and Technology
{yxueb, jhc}@cse.ust.hk

Abstract

Survival prediction is a complicated ordinal regression task that aims to predict the ranking risk of death, which generally benefits from the integration of histology and genomic data. Despite the progress in joint learning from pathology and genomics, existing methods still suffer from challenging issues: 1) Due to the large size of pathological images, it is difficult to effectively represent the gigapixel whole slide images (WSIs). 2) Interactions within tumor microenvironment (TME) in histology are essential for survival analysis. Although current approaches attempt to model these interactions via co-attention between histology and genomic data, they focus on only dense local similarity across modalities, which fails to capture global consistency between potential structures, i.e. TME-related interactions of histology and co-expression of genomic data. To address these challenges, we propose a Multimodal Optimal Transport-based Co-Attention Transformer framework with global structure consistency, in which optimal transport (OT) is applied to match patches of a WSI and genes embeddings for selecting informative patches to represent the gigapixel WSI. More importantly, OT-based co-attention provides a global awareness to effectively capture structural interactions within TME for survival prediction. To overcome high computational complexity of OT, we propose a robust and efficient implementation over micro-batch of WSI patches by approximating the original OT with unbalanced mini-batch OT. Extensive experiments show the superiority of our method on five benchmark datasets compared to the state-of-the-art methods. The code is released.

1. Introduction

Survival prediction is a complex ordinal regression task that aims to estimate the relative risk of death in cancer prognosis, which generally integrates the manual assessment of qualitative morphological information from pathology data and quantitative molecular profiles from genomic data in clinical settings [6, 55, 4]. Despite recent advances in multimodal learning for histology and genomics [5, 55, 27, 34], there still exists several open issues. Among others, one daunting challenge is to capture key information from heterogeneous modalities for effective fusion. Particularly, due to the large size of about 500,000 × 500,000 pixels, it is challenging to effectively represent gigapixel whole slide images without losing key information. Furthermore, visual concepts of tumor microenvironment (TME) within pathological images are verified to have significant associations with survival analysis in various cancer types [42, 1, 25], e.g. cellular components including fibroblast cells and various immune cells that can alter cancer cell behaviors [33]. However, the TME-related patches occupy only a tiny proportion of the entire WSI, leading to a fine-grained visual recognition problem [54] that is indiscernible...
by conventional multimodal learning.

Recently attention-based multiple instance learning (MIL) [19, 26] has provided a typical solution to identify informative instances, in which a WSI is formulated as a bag of patches (instances) and an attention score is assigned for each instance as a weight for selection. In multimodal learning, genomic data has been applied to guide the selection of TME-related instances by co-attention mechanism across modalities [5], as genes expression might correspond to some morphological characteristics revealed in pathological TME [29, 38]. They [5] densely calculate similarity scores for each pair of pathology and genomic instances as weights of selection to capture fine-grained visual information in WSIs, as shown in Fig. 1 (a). However, this type of approaches with a local view may not thoroughly learn information about TME, since they ignore global potential structure [55] within modality, e.g. interactions within TME for histology and co-expression for genomics [31, 51].

Fruitful works [31, 33, 5] demonstrate that the interactions within TME are important indicators affecting survival outcomes, e.g. co-occurrence of tumor cells with tumor-infiltrating lymphocytes (TILs) of the immune system is a positive prognostic indicator. However, these collaborative components in TME might be spatially dispersed throughout the entire WSI, which indicates long-range structural associations in WSI, as shown in Fig. 1 (b). On the other hand, genes co-expression [49, 55] also suggests a potential structure. There might be intrinsic consistency between these two potential structures [55, 31], as some studies have verified that genomic mutations can alter normal functions and biological processes of TILs within TME [51, 47]. By leveraging the global consistency between their potential structures to match histology and genomic data, it is more likely to identify TME-related patches from WSIs. However, existing co-attention-based multimodal learning focuses on only dense local similarity, neglecting the global coherence of potential structures.

To address these challenges, we propose a Multimodal Optimal Transport-based Co-Attention Transformer (MOTCat) framework with global structure consistency, in which optimal transport-based co-attention is applied to match instances between histology and genomics from a global perspective, as shown in Fig. 1 (b). Optimal transport (OT) [11, 3, 41], as a structural matching approach, is able to generate an optimal matching solution with the overall minimum matching cost, based on local pairwise cost between instances of histology and genomics. As a result, instances of a WSI that have high global structure consistency with genes co-expressions can be identified to represent the WSI. These instances might be more strongly associated with TME that contributes to survival prediction. In this way, the aforementioned two challenges can be addressed by OT-based Co-Attention simultaneously.

Compared with the conventional co-attention mechanism, the advantages of the proposed OT-based co-attention are three-fold. First, optimal transport provides the instances matching process with global awareness. Marginal constraints of total mass equality enforce a trade-off of instances within modality for transporting during optimization, instead of only considering the local similarity of pairwise instances as conventional co-attention does. It enables the modeling of structural interactions of WSIs and co-expressions of genomics, which is beneficial to survival prediction. Second, the learned patch-level attention score is not a rigorous metric for selecting informative instances under weak supervision [48], i.e. slide-level survival time, especially with only a small number of WSI samples. As an alternative, the optimal matching flow is the rigorous closed-form solution to OT problem without accessing any label, achieved by solving the Linear Programming problem. Last, the optimal matching flow provides a transformation between two modalities under preserving the potential structure, which reduces cross-modal heterogeneity gap.

Nevertheless, due to a massive number of patches from gigapixel WSIs, OT-based co-attention might suffer from high computational complexity, preventing from applying OT to pathology data in practice. To address this, we propose a robust and efficient implementation of OT-based co-attention for matching histology and genomics. Specifically, we split all instances of a WSI into a subset termed Micro-Batch and get the averaged result as a proxy to approximate the solution to the original OT problem over all instances by unbalanced optimal transport [13], which can provide a more robust approximation to the subset sampling with the theoretical and experimental guarantee.

It is worth noting that the proposed method is a generalized multimodal learning framework that captures potential structure across modalities with global structure consistency. The contributions of this work are summarized as follows: (1) we propose a novel multimodal OT-based Co-Attention Transformer with global structure consistency, where OT-based co-attention is used to identify informative instances by global optimal matching, which allows modeling interactions of histology and co-expression of genomics. (2) We propose a robust and efficient implementation of OT-based co-attention over Micro-Batch. (3) Extensive experiments on five benchmark datasets show significant improvement over state-of-the-art methods.

2. Related Work

2.1. Multimodal Learning in Healthcare

In clinical settings, the patient state is often characterized by a spectrum of various modalities [30], such as pathology [54, 32], radiology [39, 50], genomics [28, 47, 31], etc. Multimodal learning has shown superiority by leveraging
the complementary information from multimodal data [30], in which the key issue is to overcome data heterogeneity for better feature fusion. Most existing methods can be roughly classified into three categories: early fusion [18, 43], late fusion [20, 4], and intermediate fusion [5, 24]. Early fusion is to aggregate all modalities at the input level. The most straightforward solution is fusion operators such as concatenation [18], Kronecker Product [35], etc. However, it neglects intra-modality dynamics [30]. On the contrary, late fusion [15] integrates the predictions from individually separated models at the decision level for the final decision, which cannot fully explore cross-modal interactions. Recently, intermediate fusion has attracted much interest by capturing cross-modal interconnections at different levels, where the typical one is attention-based multimodal fusion [3]. For example, Zhu et al. [52] proposed a triplet attention to integrate MRI and diffusion tensor imaging (DTI) for epilepsy diagnosis. IMGFN [55] designed a graph attention network for survival prediction. MCAT [5] proposed a co-attention to identify informative instances of gigapixel WSI with genomic features as queries. HFBSurv [27] progressively integrated multimodal information from the low level to the high level for survival analysis. Note that our method MOTCat belongs to one of the intermediate fusion approaches, which explores the global consistency of potential structure by optimal transport.

2.2. Multiple Instance Learning in Pathology

Since it is difficult to represent WSIIs of large size, MIL algorithms have achieved remarkable progress in pathology, in which WSIIs are often formulated as a bag of pathology patches. MIL in pathology generally can be divided into two categories: instance-based algorithms [2, 44, 7] and embedding-based algorithms [19, 16, 32]. The former aims to select a small number of instances within each WSI for training an aggregation model. The latter one is to map each instance into a low-dimension fixed-length embedding offline, and then to learn a bag-level representation based on instance-level embeddings. There are several strategies for aggregating instance-level embeddings. One is clustering-based methods [37, 40], where all the instance-level embeddings are clustered to several centroids that are integrated for the final prediction. The most common strategy is attention-based MIL (AB-MIL) [19, 26] that assigns a weight for each instance, in which various approaches are developed differing in the ways to generate the weighting values. For example, one of the first AB-MIL [19] used a side-branch network for generating attention weights, followed by DS-MIL [26] adopted cosine distance to measure the similarity of instances as weights. Recently, transformer-based MIL like TransMIL [36] has leveraged the self-attention mechanism to explore long-range interactions in WSI. Furthermore, DTFD-MIL [48] proposed the double-tier MIL framework for more rigorous instance weights. Unlike AB-MIL, we use optimal match flow generated by OT between pathological and genomic instances to identify instances of WSI that have global potential structure consistency with that of genomics.

2.3. Survival Prediction

Survival outcome prediction, also known as time-to-event analysis [23], aims to predict the probability of experiencing an event of interest, e.g. death event in the clinical setting, before a specific time under both right-censored and uncensored data. Right-censored data refers to cases where the outcome event cannot be observed within the study period [23]. Before the deep learning era, statistical models dominated survival analysis. The typical one is Cox Proportional Hazards (CoxPH) model [9] that characterizes the hazard function, the risk of death, by an exponential linear model. DeepSurv [12] is the first work incorporating deep learning into survival analysis, which models the risk function by several fully connected layers and feeds it to the CoxPH model for the final hazard function. Furthermore, DeepConvSurv [53] first attempted to use pathological images for a deep survival model. Recent works [27, 6, 34, 4, 5] tend to directly estimate the hazard function by deep learning models without any statistical assumption. Histology and genomics data are often integrated as the gold standard for predicting survival outcomes with promising performance [4, 6, 55]. Similarly, our method incorporates these two modalities for better survival prediction by learning their global structural coherence.

3. Method

3.1. Overview and Problem Formulation

The overall framework of the proposed method is shown in Fig. 2. In this work, we aim to estimate the hazard functions $h_n(t)$ considered as the risk probability of experiencing death event in a small interval after time point $t$ for $n$-th patient, based on pathology data $X_{np}$ and genomic data $X_{ng}$.

For pathology, we formulate each WSI as a "bag" structure in a conventional multiple instance learning (MIL) setting, where a bag consists of a collection of instances that are patched from the tissue-containing image regions of a WSI, and instance-level features are extracted from patch images. Genomic data is also organized as a "bag" based on the biological functional impact of genes. For each instance of the genomic bag, we adopt an encoder to extract genes embeddings. The bags formulations for histology and genomics are illustrated in Sec. 3.1.1 and 3.1.2, respectively.

Then, Optimal Transport-based Co-Attention is applied to match instances of a WSI and the corresponding genomic instances to model pathological interactions and genomic co-expressions, so that the instances with high structure
consistency can be identified to represent the whole modality. This part will be demonstrated in Sec. 3.2.

After that, a Transformer encoder will integrate all instances after selection for each modality into a bag-level representation, in which each instance is treated as a token. The bag-level representation will be concatenated into the multimodal features for predicting a hazard function to get the ordinal survival risk. The details will be described in Sec. 3.1.3. Finally, optimization process over Micro-Batch and the corresponding computational complexity analysis will be presented in Sec. 3.3.

### 3.1.1 WSI Bags Formulation

We formulate the learning process of pathology images (WSIs) as a weakly-supervised MIL task, in which each WSI $X^p_n$ is represented as a bag with access to only bag-level (slide-level) labels. Given a bag $X^p_n = \{x^p_{n,i}\}_{i=1}^{M_p}$ of $M_p$ permutation-invariant instances, the goal is to encode instance-level features of a WSI into a compact bag-level representation, in which each instance is treated as a token. After identifying informative instances with global structure consistency. After aggregating selected instances, features from both modalities are concatenated for survival prediction.

The optimal parameters of model; the bag-level representation $H^p_n = \mathcal{T}_p(\hat{B}^p_n)$ for the subsequent features fusion.

**Algorithm 1 The MOTCat Algorithm**

**Input:** A WSI bag as $X^p_n = \{x^p_{n,i}\}_{i=1}^{M_p}$; A genomic bag as $X^g = \{x^g_{n,j}\}_{j=1}^{M_g}$. $m$ refers to the size for a micro-batch of a WSI.

**Output:** The optimal parameters of model;

1. // Sample micro-batch from the WSI bag $X^{p,m}_n = \{x^{p,1}_{n,i}\}_{i=1}^{m}$ each time.
2. for $X^{p,m}_n$ in $X^p_n$
3. Fix model parameters $f_p$, $f_g$, $T_p$ and $T_g$.
4. // To extract a bag of instance-level features
5. // Eq. 1 over micro-batch of histology
6. $B^{p,m}_n = \{f_p(x^{p,1}_{n,i}) : x^{p,1}_{n,i} \in X^{p,m}_n\}$
7. // Eq. 2 of genomics
8. $B^{g}_n = \{f_g(x^{g}_n) : x^{g}_n \in X^{g}\}$
9. Optimize $P^{n}_m$ via Sinkhorn-Knopp matrix scaling algorithm [8, 14]:
10. $P^{n}_m \leftarrow \arg\min_{P^{n}_m} \mathcal{W}(B^{p,m}, B^{g})$ // Eq. 5
11. Fix $P^{n}_m$ and select instances from $B^{p,m}_n$ by:
12. $\hat{B}^{p,m}_n = P^{n}_m \odot B^{p,m}_n$
13. // To get bag-level representation
14. $H^{p,m}_n = \mathcal{T}_p(\hat{B}^{p,m}_n)$ // for histology
15. $H^{g}_n = \mathcal{T}_g(\hat{B}^{g}_n)$ // for genomics
16. $H^{n}_n = H^{p,m}_n \bowtie H^{g}_n$ // Concatenation
17. Calculate the overall loss function $\mathcal{L}$ of Eq. 6.
18. Update parameters of $f_p$, $T_p$, $T_g$ and the last fully connected layers of $f_p$ by Adam.
19. end for

---

Figure 2. Overview of Multimodal Optimal Transport-based Co-Attention Transformer (MOTCat) architecture, in which both modalities are formulated as bags representations and a micro-batch of the WSI bag is sampled as pathological input. Then OT-based Co-attention is to calculate optimal matching flow for identifying informative instances with global structure consistency. After aggregating selected instances, features from both modalities are concatenated for survival prediction.
3.1.2 Genomic Bags Formulation

Genomic data consists of $1 \times 1$ attributes, such as mutation status, transcript abundance (RNA-Seq), copy number variation (CNV) and other molecular characterizations. Following the organization of genomic data in the previous work [5, 28] that categories them according to the biological functional impact, the expressive genomic bag can be organized as $X^g_{n,j} = \{x_{n,j}^g\}_{j=1}^{M_g} \in \mathbb{R}^{d_g \times d_j}$ with $M_g$ instances of unique functional categories. For j-th category, the corresponding genomic embedding consists of $d_j$-dimensional attributes encoded by a separated network $d$. Responding genomic embedding consists of a bag $B^g_n \in \mathbb{R}^{M_g \times d}$ of genomic data with $M_g$ instances can be constructed as follows:

$$B^g_n = \{f^g_j(x_{n,j}^g) : x_{n,j}^g \in X^g_n\} = \{b^g_{n,j}\}_{j=1}^{M_g},$$

which is aggregated to the bag-level embedding by a transformer encoder $H^g_n = T_g(B^g_n)$ as well.

3.1.3 Multimodal Survival Prediction Formulation

Survival outcome prediction aims to predict the risk probability of an outcome event occurring before a specific time, in which the outcome event is not always observed, leading to a right-censored event. In this task, let $c \in \{0, 1\}$ be censor status indicating whether the outcome event is right-censored ($c = 1$) or not ($c = 0$), and $t \in \mathbb{R}^+$ refers to the overall survival time (in months).

Given a training set of $N$ pathology-genomics pairs $D = \{(X^p_n, X^g_n, c_n, t^p_n)\}_{n=1}^{N}$, we can acquire the bag-level features $H^p_n$ of WSIs and $H^g_n$ of genomic data for $n$-th patient data $X_n = (X^p_n, X^g_n, c_n, t^p_n)$, as described in Section 3.1.1 and 3.1.2. After integrating $H^p_n$ and $H^g_n$ into the multimodal features $H_n$ by concatenation, we estimate the hazard function $h_n(t|H_n) = h_n(T = t|T \geq t, H_n) \in [0, 1]$ which is viewed as the probability of death event occurring at around time point $t$. Instead of predicting the overall survival time $t^n$, the prediction task is to estimate the ordinal risk of death via the cumulative survival function:

$$S_n(t|H_n) = \prod_{z=1}^{t} (1 - h_n(z|H_n))$$

3.2. Optimal Transport-based Co-Attention

To identify TME-related instances of WSIs, we utilize the global potential structure consistency between TME interactions of pathology and genes co-expressions as the evidence of selecting instances, in which optimal transport is used to learn the optimal matching flow between a WSI feature bag $B^p_n \in \mathbb{R}^{M_p \times d}$ and a genomic bag $B^g_n \in \mathbb{R}^{M_g \times d}$. Formally, an optimal transport from the WSI feature bag $B^p_n = \{b^p_{n,1}, b^p_{n,2}, \ldots, b^p_{n,M_p}\}$ to the genomic embedding bag $B^g_n = \{b^g_{n,1}, b^g_{n,2}, \ldots, b^g_{n,M_g}\}$ can be defined by the discrete Kantorovich formulation [21] to search the overall optimal matching flow $P_n$ between $B^p_n$ and $B^g_n$:

$$\mathcal{W}(B^p_n, B^g_n) = \min_{P_n \in \Pi(\mu_p, \mu_g)} <P_n, C_n>_{F}$$

where $C_n \geq 0 \in \mathbb{R}^{M_p \times M_g}$ is a cost matrix given by $C_{n,u,v} = c(b^p_{n,u}, b^g_{n,v})$ with a ground distance metric $c(\cdot)$, such as $l_2$-distance, that measures the distance of local pairwise instances $b^p_{n,u} \in B^p_n$ in the WSI bag and the genomic one of a unique functional category $b^g_{n,v} \in B^g_n$.

Here $\Pi(\mu_p, \mu_g) = \{P_n \in \mathbb{R}^{M_p \times M_g} | P_n 1_{M_p} = \mu_p, P_n 1_{M_g} = \mu_g\}$ involves the marginal constraints of total mass equality between marginal distributions, $\mu_p$ and $\mu_g$, for the WSI bag and the genomic bag, and $1_k$ is a $k$-dimensional vector of ones. Specifically, $\Pi(\mu_p, \mu_g)$ refers to the set of joint probabilistic couplings between the two marginal empirical distributions of pathology data and genomic data. Intuitively, it describes how to distribute instances of a WSI bag $B^p_n$ to the genomic embedding of a genomic bag $B^g_n$ based on the cost matrix $C_n$, under the marginal constraints of total mass equality between distributions $\mu_p$ and $\mu_g$. Note that $< \cdot, >_{F}$ refers to the Frobenius dot product, and thus Eq. 4 encourages to achieve the overall minimum matching cost by finding optimal matching flow based on local pairwise similarity. In this way, the optimal transport of Eq. 4 is able to enforce a trade-off of instances within modality, which allows the model to capture the potential structure of visual interactions for histology and co-expressions for genomic data.

Once acquiring the optimal matching flow $P^*_n$, informative instances of a WSI are identified by $P^*_n = P^*_n 1_{M_p}$ to represent the WSI, which also aligns pathology distribution to genomics distribution under preserving the potential structure across modalities for alleviating heterogeneity.

As such, the proposed OT-based Co-Attention simultaneously addresses two issues: not only does it achieve the effective representation for gigapixel WSIs by identifying informative instances via OT matching, but marginal constraints of OT enable the selected instances to have global structure consistency with genomic data.

3.3. Optimization over Micro-Batch

Due to the large size of the WSI bag, it is difficult to apply optimal transport for matching histology data and genomic data. Recent work [48] has validated that MIL on WSI can benefit from multiple subsets of a bag randomly sampled from the original WSI bag. It provides a way to train the model over Micro-Batch of a WSI, where micro-batch is defined as a subset sampled from a bag of WSI instances. Furthermore, inspired by a variant [13] of OT that offers a solution to approximate the original OT over mini-batches with a theoretical guarantee of convergence, we propose to use the variant UMBOT formulation [13] over micro-batch of WSI instead of the original OT over
all instances of a WSI. Then the Eq. 4 becomes:

\[
\mathcal{W}^m(B_n^{p,m}, B_n^0) = \min_{\mathbf{P}_n^{m} \in \Pi(\mathbf{P}_n^{m} \circ \mu_p)} <\mathbf{P}_n^{m}, \mathbf{C}_n^{m}>_F \\
+ \epsilon KL(\mathbf{P}_n^{m}||\mu_p) + \tau (D_{\phi}(\mathbf{P}_n^{m,\gamma}||\mu_p) + D_{\phi}(\mathbf{P}_n^{m,\gamma}||\mu_\gamma))
\]

(5)

where \(B_n^{p,m}\) is micro-batch of size \(m\) sampled from \(B_n^p\). Similarly, \(\mathbf{C}_n^{m} \in \mathbb{R}^{M_n \times M_p}\) is the cost matrix over micro-batch of WSI \(B_n^{p,m}\) and \(B_n^0\), and the optimal matching flow \(\mathbf{P}_n^{m} \in \mathbb{R}^{M_n \times M_p}\) is optimized over micro-batch as well. Here \(\mathbf{P}_n^{m,\gamma}\) and \(\mathbf{P}_n^{m,\gamma}\) represent the marginals of \(\mathbf{P}_n^{m}\), and \(D_{\phi}\) is Csisz`ar divergences, \(\tau\) and \(\epsilon \geq 0\) are the coefficients of marginal penalization and entropic regularization, respectively. \(KL\) refers to Kullback-Leiblle divergence.

**Computational complexity.** With the benefit of UMBOT for approximating the original OT, the computational complexity of optimization is reduced from \(O(M^3 \log(M))\) to \(O((M_n \times M_p)^2) = O(M \times m)\), where \(M = \max(M_p, M_s)\) and \(m \ll M\). The whole optimization procedure of the proposed MOTCat is presented in Alg. 1.

**Loss function.** Following the previous work [5], the overall loss function is formulated by NLL-loss [46]:

\[
\mathcal{L} = - \frac{m}{M_p} \sum_{n=1}^{N} \sum_{X_n^p \in X_n^p} c_n \cdot \log (S_n^m(t_n|H_n^m)) \\
- \frac{m}{M_p} \sum_{n=1}^{N} \sum_{X_n^p \in X_n^p} \{(1 - c_n) \cdot \log (S_n^m(t_n - 1|H_n^m)) \\
+ (1 - c_n) \cdot \log (h_n^m(t_n|H_n^m))\}
\]

(6)

where \(H_n^m\) refers to the multimodal features over micro-batch formulated in line 16 of Alg. 1. \(h_n^m\) and \(S_n^m\) are the hazard function and cumulative survival function over micro-batch, respectively.

4. Experiment

In this section, we first present datasets and settings following previous experimental protocols [5, 6] for fair comparisons. Next, we demonstrate the performance results of the proposed method compared with the state-of-the-art (SOTA) methods, including unimodal and multimodal approaches. After that, we investigate the effectiveness of each component in our method and the effect on the size of micro-batch strategy. Finally, from a statistical point of view, we present Kaplan-Meier survival curves and Logrank test to show the performance of survival analysis. Interpretable visualization for histology data and genomic data are presented in supplementary materials.

4.1. Datasets and Settings

**Datasets.** To demonstrate the performance of the proposed method, we conducted various experiments over five public cancer datasets from The Cancer Genome Atlas (TCGA) that contains paired diagnostic WSIs and genomic data with ground-truth survival outcome: Bladder Urothelial Carcinoma (BLCA), Breast Invasive Carcinoma (BRCA), Uterine Corpus Endometrial Carcinoma (UCEC), Glioblastoma & Lower Grade Glioma (GBMLGG), and Lung Adenocarcinoma (LUAD). The number of cases for each cancer type is shown by \(N\) in Tab. 1. Note that cases of these cancer datasets used for all compared methods are not less than cases used in the proposed method. For genomic data, the number of unique functional categories \(M_p\) is set as 6 following [28, 5], including 1) Tumor Supression, 2) Oncogenesis, 3) Protein Kinases, 4) Cellular Differentiation, 5) Transcription, and 6) Cytokines and Growth.

**Evaluation.** For each cancer dataset, we perform 5-fold cross-validation with a 4:1 ratio of training-validation sets and report the cross-validated concordance index (C-Index) and its standard deviation (std) to quantify the performance of correctly ranking the predicted patient risk scores with respect to overall survival.

**Implementation.** For each WSI, we first apply the OTSU’s threshold method to segment tissue regions, and then non-overlapping 256 \(\times\) 256 patches are extracted from the tissue region over 20\(\times\) magnification. Then, we use ImageNet-pretrained [10] ResNet-50 [17] and a 256-d fully-connected layer as the feature encoder \(f_p\) to extract the 1024-d embedding for each patch, where parameters of ResNet-50 are frozen. For genomic data, we adopt SNN [22] as the feature encoder \(f_p\) following the setting of [5].

During training, we follow the setting of [5] for a fair comparison. Specifically, we adopt Adam optimizer with the initial learning rate of 2 \(\times\) 10\(^{-4}\) and weight decay of 1 \(\times\) 10\(^{-5}\). Due to the large size of WSIs, the batch size for WSIs is 1 with 32 gradient accumulation steps, and all experiments are trained for 20 epochs. The size of Micro-Batch \(m\) is set as 256. Regarding the hyper-parameters of OT in Eq. 5, the coefficient of marginal penalization \(\tau\) is 0.5 for all cancer datasets, and the coefficient of entropic regularization \(\epsilon\) is 0.05 for BLCA and LUAD, as well as 0.1 for BRCA, UCEC and GBMLGG.

4.2. Results

We compare our method against the unimodal baselines and the multimodal SOTA methods as follows:

**Unimodal Baseline.** For genomic data, we adopt SNN [22] that has been used previously for survival outcome prediction in the TCGA [4, 5], and SNNTrans [22, 36] which incorporates SNN as the feature extractor and TransMIL [36] as the global aggregation model for MIL. For histology, we compare the SOTA MIL methods AttnMIL [19], DeepAttnMISL [45], CLAM [32], TransMIL [36], and DTFD-MIL [48].

**Multimodal SOTA.** We compare four SOTA methods for multimodal survival outcome prediction including
Table 1. C-Index (mean ± std) performance over five cancer datasets. Patho. and Geno. refer to pathology modality and genomic modality, respectively. The best results and the second-best results are highlighted in **bold** and in *underline*, respectively.

<table>
<thead>
<tr>
<th>Model</th>
<th>Patho. Geno.</th>
<th>BLCA (N = 373)</th>
<th>BRCA (N = 956)</th>
<th>UCEC (N = 480)</th>
<th>GBMLGG (N = 569)</th>
<th>LUAD (N = 453)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SNN* [22]</td>
<td>✓</td>
<td>0.618 ± 0.022</td>
<td>0.624 ± 0.060</td>
<td>0.679 ± 0.040</td>
<td>0.834 ± 0.012</td>
<td>0.611 ± 0.047</td>
</tr>
<tr>
<td>SNTrans* [22, 36]</td>
<td>✓</td>
<td>0.659 ± 0.032</td>
<td>0.647 ± 0.063</td>
<td>0.656 ± 0.038</td>
<td>0.839 ± 0.014</td>
<td>0.638 ± 0.022</td>
</tr>
<tr>
<td>AttnMIL* [19]</td>
<td>✓</td>
<td>0.599 ± 0.048</td>
<td>0.609 ± 0.065</td>
<td>0.658 ± 0.036</td>
<td>0.818 ± 0.025</td>
<td>0.620 ± 0.061</td>
</tr>
<tr>
<td>DeepAttnMISL [45]</td>
<td>✓</td>
<td>0.504 ± 0.042</td>
<td>0.524 ± 0.043</td>
<td>0.597 ± 0.059</td>
<td>0.734 ± 0.029</td>
<td>0.548 ± 0.050</td>
</tr>
<tr>
<td>CLAM-SB* [32]</td>
<td>✓</td>
<td>0.559 ± 0.034</td>
<td>0.573 ± 0.044</td>
<td>0.644 ± 0.061</td>
<td>0.779 ± 0.031</td>
<td>0.594 ± 0.063</td>
</tr>
<tr>
<td>CLAM-MB* [32]</td>
<td>✓</td>
<td>0.565 ± 0.027</td>
<td>0.578 ± 0.032</td>
<td>0.609 ± 0.082</td>
<td>0.776 ± 0.034</td>
<td>0.582 ± 0.072</td>
</tr>
<tr>
<td>TransMIL* [36]</td>
<td>✓</td>
<td>0.575 ± 0.034</td>
<td>0.666 ± 0.029</td>
<td>0.655 ± 0.046</td>
<td>0.798 ± 0.043</td>
<td>0.642 ± 0.046</td>
</tr>
<tr>
<td>DTFD-MIL* [48]</td>
<td>✓</td>
<td>0.546 ± 0.021</td>
<td>0.609 ± 0.059</td>
<td>0.656 ± 0.045</td>
<td>0.792 ± 0.023</td>
<td>0.585 ± 0.066</td>
</tr>
<tr>
<td>Pathomic [4]</td>
<td>✓ ✓</td>
<td>0.586 ± 0.062</td>
<td>-</td>
<td>0.826 ± 0.009</td>
<td>0.543 ± 0.065</td>
<td></td>
</tr>
<tr>
<td>PONET [34]</td>
<td>✓ ✓</td>
<td>0.643 ± 0.037</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.646 ± 0.047</td>
</tr>
<tr>
<td>Porpoise* [6]</td>
<td>✓ ✓</td>
<td>0.636 ± 0.024</td>
<td>0.652 ± 0.042</td>
<td><strong>0.695 ± 0.032</strong></td>
<td>0.834 ± 0.017</td>
<td>0.647 ± 0.031</td>
</tr>
<tr>
<td>MCAT* [5]</td>
<td>✓ ✓</td>
<td>0.672 ± 0.032</td>
<td>0.659 ± 0.031</td>
<td>0.649 ± 0.043</td>
<td>0.835 ± 0.024</td>
<td>0.659 ± 0.027</td>
</tr>
<tr>
<td>MOTCat (Ours)</td>
<td>✓ ✓</td>
<td><strong>0.683 ± 0.026</strong></td>
<td><strong>0.673 ± 0.006</strong></td>
<td>0.675 ± 0.040</td>
<td><strong>0.849 ± 0.028</strong></td>
<td><strong>0.670 ± 0.038</strong></td>
</tr>
</tbody>
</table>

MCAT [5], Pathomic [4], Porpoise [6] and PONET [34].

Results are shown in Tab. 1, where the methods marked * are re-implemented and the best results are reported, and other results are obtained from their papers. Note that due to varying cancer datasets used in their works, not all results of the five datasets used in this work can be reported, and hence we use ‘-’ to indicate the results are not reported in their papers.

**Unimodal vs. Multimodal.** Compared with all unimodal approaches, the proposed method achieves the highest performance in 4 out of 5 cancer datasets, indicating the effective fusion of multimodal data in our method. Note that the overall performance of genomic data is better than that of histology, which validates the reasonability of using genomic data for guiding the selection of instances in WSIs. In particular, most multimodal methods are inferior to the unimodal model of genomics in UCEC dataset, suggesting the serious challenge in multimodal fusion. Nevertheless, the proposed method can achieve the comparative performance with genomic model on UCEC.

**Multimodal SOTA vs. MOTCat.** In one-versus-all comparisons of multimodal models, the proposed method achieves superior performance on all benchmarks with 1.0%-2.6% performance gains except UCEC. In comparing the proposed method with the most similar work MCAT in multimodal fusion, our method gets better results on all datasets, indicating the effectiveness of learning the intramodal potential structure consistency from a global perspective between TME-related interactions of histology and genomic co-expression.

### 4.3. Ablation Study

We first investigate the effectiveness of OT-based co-attention component (denoted by OT) and the micro-batch strategy (denoted by MB). Furthermore, the effect on the size of micro-batch and the actual computational speed of our method are explored.

**Component validation.** As shown in Tab. 2, we observe that multimodal fusion can benefit from micro-batch strategy by comparing (a) and (b), and OT-based co-attention further improves the performance by approximating the original results with the averaged results over micro-batches via UMBOT. As a result, our method achieves the best overall performance. These improvements both demonstrate the effectiveness of each component for the proposed method.

Additionally, recent work [48] has claimed that MIL on WSIs can benefit from learning over sub-bags in a WSI, which is also validated by performance increases in variant (b). The reason why it can profit from micro-batch might be that it increases the number of bags, so that more features can be used for survival analysis.

**Size of Micro-batch.** To show the robustness of approximation to the size of Micro-batch strategy, we compare several variants of the proposed method over the varying sizes of micro-batch, as shown in the quantitative analysis of Fig. 3. The qualitative analysis of the effect on size of Micro-Batch can be found in the supplementary materials, which also achieves the consistent conclusion.

For the quantitative analysis in Fig. 3, two variants of the proposed method are compared: 1) **UMBOT → EMD:** To demonstrate the robustness of different OT variants, we replace UMBOT used in our method with the original OT (i.e. EMD) for comparison, which is equivalent to the Eq. (4) done over Micro-batch. 2) **UMBOT → CoAtt:** To compare the robustness of the proposed OT-based co-attention and the original co-attention, we replace UMBOT with the co-attention used in MCAT [5] and train the model over Micro-batch for comparison.
Table 2. Ablation study assessing C-Index (mean ± std) performance of MOTCat over five datasets, in which version (c) in bold is the proposed method with full components, and (a) and (b) refer to its variants. The best results are marked in bold.

<table>
<thead>
<tr>
<th>Variants</th>
<th>OT MB</th>
<th>BLCA</th>
<th>BRCA</th>
<th>UCEC</th>
<th>GBMLGG</th>
<th>LUAD</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a) MCAT</td>
<td>✓</td>
<td>0.672 ± 0.032</td>
<td>0.659 ± 0.031</td>
<td>0.649 ± 0.043</td>
<td>0.835 ± 0.024</td>
<td>0.659 ± 0.027</td>
</tr>
<tr>
<td>(b) MOTCat w/o OT</td>
<td>✓ ✓</td>
<td>✓ (✓ 0.683 ± 0.026)</td>
<td>0.673 ± 0.006</td>
<td>0.675 ± 0.040</td>
<td>0.849 ± 0.028</td>
<td>0.670 ± 0.038</td>
</tr>
</tbody>
</table>

From the results of Fig. 3, we can see that 1) our method achieves the best averaged performance of various sizes compared with the two variants, especially on BRCA with the most samples of 956. 2) Further, on UCEC and LUAD datasets, the proposed method gets the most robust results. 3) Although the second variant UMBOT → CoAttn shows slightly better robustness on BLCA and BRCA, the performance of our method surpasses it by a large margin. In a nutshell, the proposed method achieves a better trade-off between performance and robustness, compared with different OT variants or the co-attention of MCAT.

Computational Speed. We measure the actual averaged computational speed for 10 WSIs (about 150k patches in total) on one GPU of NVIDIA GeForce RTX 3090. The proposed method is compared with a variant ‘org-OT’ that replaces UMBOT by the original OT. However, the original OT with the high complexity takes extreme long time for one WSI (about 20k patches), and thus we fail to measure the actual computational time for the original OT in a short time. Here we present the speed of our method, which suggests our method makes it practicable to apply OT to histology and genomics. Specifically, the training speed of the proposed method is 6540 p/s (p/s refers to the number of processing patches per second) and the inference speed is 11885 p/s.

4.4. Statistical Analysis

To show a statistical difference in patient stratification performance, we visualize the Kaplan-Meier survival curves for different methods, in which patients are separated into two groups of low-risk and high-risk based on predicted risk scores, and then the statistics on ground-truth survival time are presented for each group in Fig. 4. The Logrank test is conducted to measure the statistically significant difference in two groups of patients, where a lower P-value indicates better performance of patient stratification. From the results of Fig. 4, intuitively our method separates patients of low and high risk more clearly on all datasets. In the Logrank test, our method achieves a lower P-value on all datasets in comparison to MCAT, especially on BLCA, BRCA and UCEC with a large margin of magnitude.

5. Conclusion

In this paper, we present a novel Multimodal Optimal Transport-based Co-Attention Transformer with global structure consistency to tackle two important issues in mul-
timodal learning of histology and genomics for survival prediction. First, we utilize a new OT-based Co-Attention to match pairwise instances between histology and genomics for selecting informative instances strongly related to tumor microenvironment, which provides a way to effectively represent gigapixel WSI s. Second, optimal transport offers a global awareness for modeling potential structure within modality, i.e. pathological interactions and genomic co-expression. Furthermore, to enable applying OT-based Co-Attention in practice, we propose a robust and efficient implementation over micro-batch of WSI bags, in which the approximation over micro-batch might provide a solution to update the parameters of extractor for patch images in an end-to-end way instead of extracting feature embeddings offline, which will be further explored in the future.

Acknowledgement

This work was supported by National Natural Science Foundation of China (No. 62202403), Shenzhen Science and Technology Innovation Committee (Project No. SGDX20210823103201011) and Hong Kong Innovation and Technology Fund (No. PRP/034/22FX).

References


