This ICCV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Graph Domain Adaptation with Dual-branch Encoder and Two-level Alignment for Whole Slide Image-based Survival Prediction

Yuntao Shou[1,2], Xiangyong Cao[1,2,*], PeiqiangYan[3], Qiaohui[4], Qian Zhao[3,†], Deyu Meng[2,3,5,6]
[1]School of Computer Science and Technology, Xi'an Jiaotong University, Xi'an, 710049, China
[2]Ministry of Education Key Laboratory of Intelligent Networks and Network Security, Xi'an Jiaotong University, Xi'an, 710049, China
[3]School of Mathematics and Statistics, Xi'an Jiaotong University, Xi'an, 710049, China
[4]China Telecom Shaanxi Branch, Xi'an, 710049, China
[5]Macao Institute of Systems Engineering, Macau University of Science and Technology, Macao, China.
[6]Pazhou Lab (Huangpu), Guangzhou, China.

Abstract

In recent years, whole slide image (WSI)-based survival analysis has attracted much attention. In practice, WSIs usually come from different hospitals (or domains) and may have significant differences. These differences generally result in large gaps in distribution between different WSI domains and thus, the survival analysis models trained on one domain may fail to transfer to another. To address this issue, we propose a Dual-branch Encoder and Two-level Alignment (DETA) framework to explore both feature and category-level alignment between different WSI domains. Specifically, we first formulate the concerned problem as graph domain adaptation (GDA) using the graph representation of WSIs. Then, we construct a dual-branch graph encoder, including the message passing (MP) and the shortest path (SP) branches, to explicitly and implicitly extract semantic information from the graph-represented WSIs. To realize GDA, we propose a two-level alignment approach: at the category level, we develop a coupling technique by virtue of the dual-branch structure, leading to reduced divergence between the category distributions of the two domains; at the feature level, we introduce an adversarial perturbation strategy to better augment source domain feature, resulting in improved alignment in feature distribution. Extensive experiments have demonstrated the effectiveness of our proposed DETA framework in WSI-based survival analysis under the domain shift scenario.

1. Introduction

In clinic, survival analysis aims to predict the time from the start of treatment to death or disease recurrence, which plays an important role in clinical prognosis [18]. Traditional survival analysis methods usually rely on short-term clinical indicators [19, 52] and long-term follow-up reports [4, 40], which are often very time-consuming in practical applications. In recent years, with the development of deep learning technology, significant progress has been made in medical image analysis, allowing survival analysis to be more effectively integrated with more informative pathology images [16, 36]. By virtue of deep neural networks (DNNs), features from pathological images can be automatically extracted, thereby helping researchers to establish powerful survival models. In recent years, the whole slide histopathological images (WSIs) are of particular interest for this task [2, 31, 51], since they can provide rich information of diseased tissue samples, e.g., the microscopic changes in tumor cells and their microenvironment, allowing the model to have a deeper understanding of the charac-

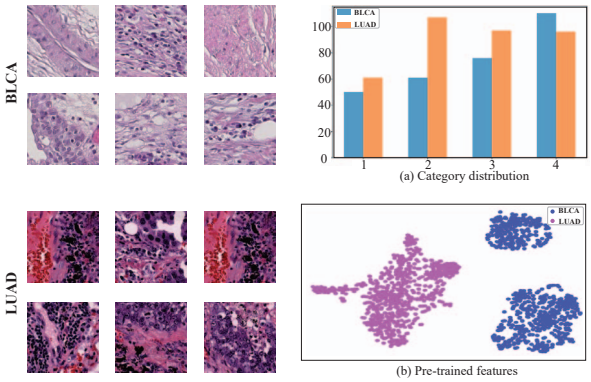


Figure 1. An example of pathology images with distribution shift. The left shows image from two WSI datasets, i.e., BLCA and LUAD. The right shows the category distribution and the t-SNE visualization of the extracted features by CLAM [24] method.

*Corresponding author: caoxiangyong@xjtu.edu.cn, timmy.zhaoqian@xjtu.edu.cn

teristics and behavior of the tumor.

Since WSIs are generally very large, e.g., can be with 1 million by 1 million pixels, a common strategy in WSI-based survival analysis is to divide a WSI into patches sharing a same label, which can thus be formulated as a multi-instance learning (MIL) problem [1, 6, 29, 37]. By exploiting the similarities between those patches, a graph with patches being its nodes can be constructed as a representation of the WSI, and thus the graph-based methods relying on the graph convolutional networks (GCNs) have been developed to solve such a MIL problem [22, 53]. Since GCNs can effectively capture complex spatial structures and contextual information to reflect potential pathological features and their interactions, a good global representation of the WSI is expected to be learned, thereby achieving accurate survival prediction among the state-of-the-art ones [8, 41].

Though achieving promising performance, existing WSI-based survival analysis methods are generally constructed without considering the domain shift issue which is often encountered in real scenarios. Specifically, WSIs collected by different hospitals or laboratories may have significant differences in distribution due to imaging equipment, processing procedures, and sample sources. We show an example in Fig. 1 by visualizing the category distributions and the distributions of features extracted by CLAM [24] of two WSI datasets. It can be seen that, in addition to the apparent visual differences in the original patches shown on the left, both the category and feature distributions of the two datasets significantly deviate from each other. Due to such a domain shift, the survival analysis models trained on one domain could hardly transfer to another, resulting in degraded performance in the target domain, which limits their availability in real applications.

To address the above issue, we propose an innovative Dual-branch Encoder and Two-level Alignment (DETA) framework to realize the graph domain adaptation (GDA) for the WSI-based survival analysis by fully exploring both category and feature-level distribution alignment between different WSI domains. Specifically, using the graph representation of WSIs, we first formulate the concerned problem as GDA. Then, we design a dual-branch graph encoder, containing a message passing (MP) and a shortest path (SP) branch to fully utilize the features of both the source and target WSI graphs. Within the dual structure, the MP branch can implicitly learn the topological semantics in the graph by neighborhood aggregation, while the SP branch complementarily provides explicit high-order structural semantics. Next, we develop several techniques to align the source and target domains at both category and feature levels. Specifically, at the category level, we define coupled evidence lower bounds, by virtue of the dual-branch structure of the graph encoder to jointly train the graph encoder and risk predictor, so that the divergence between the category distributions of the two domains is expected to be reduced; at the feature level, we adversarially train a domain classifier and learn the perturbations to the graph features of the source domain, so that the feature distributions of both domains can be better aligned. With all these designs, the final survival prediction performance is expected to be enhanced.

Our main contributions can be summarized as follows:

- To the best of our knowledge, we make the first attempt to alleviate the domain shift issue in the WSI-based survival analysis task, thereby enhancing its availability in real scenarios.
- We propose a Dual-branch Encoder and Two-level Alignment (DETA) framework to explicitly and implicitly extract semantic information from WSIs, as well as align both the category and feature distributions of source and target domains, such that the risk predictor can perform well in the target domain.
- We apply the proposed method to four public TCGA datasets with different adaptation directions and demonstrate its superior performance for the WSI-based survival analysis under the domain shift scenario.

2. Related Work

2.1. WSIs-based Survival Analysis

Survival prediction plays a vital role in clinical medicine, providing doctors with valuable information to evaluate disease progression and treatment effects. Traditionally, survival prediction usually relies on rich clinical data, including short-term clinical indicators [19], [52], long-term follow-up reports [4, 40] and various radiological images [12, 30]. Recently, with the rapid development of deep learning, WSI-based survival analysis has attracted increasing attention, since rich information about the disease can be provided by the gigapixel pathological images [5, 25]. Since the WSI is generally with very high resolution, a common strategy is to divide it into patches sharing the same labels and treat the WSI-based survival analysis as MIL problem [1, 54]. For example, CMIB [56] captures the private and public features of different modalities through a co-attention mechanism and introduces a multimodal information bottleneck to obtain a robust representation of fused features. TANGLE [17] adopts a modality-specific encoder and contrastive learning to aligns the outputs and learn a task-agnostic slice embeddings. However, since WSIs come from different hospitals or laboratories, the domain shift issue can occur in real scenarios, limiting the availability of WSI-based survival analysis in practical applications.

2.2. Unsupervised Domain Adaptation

Unsupervised domain adaptation (UDA) aims to learn domain-invariant representations so that the model can be transferred from a source domain with rich labels to a tar-

get domain with scarce labels, which is crucial to improving the generalization ability of the model in different application scenarios [27], [35], [11], [28]. Existing UDA methods can be roughly divided into two categories: domain difference-based methods and adversarial methods. Domain difference-based methods usually combine multiple distribution indicators (e.g., maximum mean difference [32] and Wasserstein distance [34]) to measure the difference between the source domain and the target domain. By quantifying the distribution difference between the source domain and the target domain, the model is adjusted to better adapt to the target domain. In contrast, adversarial methods [7, 50] adopt a more implicit strategy to reduce inter-domain differences by introducing domain discriminators.

2.3. Graph Domain Adaption

Graph domain adaptation (GDA) focuses on how to effectively transfer information from source graphs to unlabeled target graphs to learn effective node-level representations [23], [45], [26]. Existing methods usually focus on domain alignment through graph neural networks (GNNs) to promote knowledge transfer between source and target graphs [44], [57], [7]. However, these methods often ignore the importance of category distribution alignment in the case of label scarcity and domain shift [14], [50], [49], [48]. Particularly, when the number of samples in the target domain is scarce, it is difficult for a model to accurately capture the feature distribution without sufficient category information.

3. Problem Formulation

WSI-based survival analysis. Suppose $Z = \{z_1, z_2, \ldots, z_N\}$ is the WSI-based survival data of N patients, where z_i is represented by a triplet $z_i = \{x_i, c_i, y_i\}$, x_i is the WSI, $c_i \in \{0, 1\}$ denotes the right censorship status, with $c_i = 1$ indicating that the i-th sample is uncensored and $c_i = 0$ censored, and y_i represents the observed time. The goal of survival analysis is to estimate the following hazard function $h(y|x)$:

$$h(y|x) = \lim_{\Delta y \to 0} \frac{P(y \le O \le y + \Delta y | O \ge y, x)}{\Delta y}, \tag{1}$$

where O is the survival time, and consequently obtain the survival function $S(y|x) = P(O \ge y|x)$.

In the deep learning practice for the WSI-based survival analysis, the time is often discretized into intervals and y_i takes the integer value indicating which time interval (or the risk level) the i-th patient belongs to [54, 55]. In this case, the hazard function becomes

$$h(y|x) = P(O = y | O \ge y, x), \quad y = 1, 2, \cdots, K, \tag{2}$$

where K denotes the total risk levels, and then the survival function can be obtained as

$$S(y|x) = \prod_{j=1}^{y} (1 - h(j|x)). \tag{3}$$

By virtue of DNNs, one can explicitly parameterize $h(y|x)$:

$$h(y|x) = \phi_y(g(x)), \quad y = 1, 2, \cdots, K, \tag{4}$$

where $g(\cdot)$ is the feature extractor and $\phi_y(\cdot)$ maps the feature vector to the conditional probability $P(O = y | O \ge y, x)$ so that $\sum_{y=1}^{K} P(O = y | O \ge y, x) = 1$, which can be realized by a multilayer perception (MLP) with the softmax output. Then h can be learned by minimizing the following loss derived from the maximum likelihood estimation:

$$\begin{aligned} \mathcal{L}_{surv} = &- \sum_{i=1}^{N} c_i \left[\log S(y_i|x_i) + \log h(y_i|x_i) \right] \\ &- \sum_{i=1}^{N} (1 - c_i) \log S(y_i + 1 | x_i) \end{aligned}. \tag{5}$$

It should be mentioned that we follow the discrete-time setting, since the category distribution can be easier to explore for domain adaptation.

Graph representation for WSI-based survival analysis. As mentioned in the Introduction, graph-based methods have achieved promising performance for WSI-based survival analysis [8, 22, 41, 53], since the graph-representation is helpful to capture complex structures and contextual information to reflect pathological features and their interaction, leading to good global representations for WSIs. The main idea is to first divide the large WSI into a series non-overlap patches and extract features for them, and then construct a graph with each node being the feature of one patch and each edge represents the neighborhood relationship between two patches. Such a WSI graph can be formally represented as $G = (V, E, F)$, where V denotes the node set, E represents the edge set, and $F \in \mathbb{R}^{|V| \times d}$ is the node attribute matrix, with $|V|$ is the total number of nodes, and d is the dimension of the attribute (or feature) of each node. With the constructed WSI graphs, the original WSI-based survival analysis now can be regarded as a graph learning problem using data triplets $\{G_i, c_i, y_i\}_{i=1}^{N}$, where G_i is the graph constructed by WSI x_i. In this work, following [18, 46], we adopt CLAM [24] to crop each WSI into 512×512 non-overlap patches and extract 1024-dimensional feature for each patch, and use the K-nearest neighbor (KNN) algorithm to construct the graph according to the Euclidean distance between patch features.

GDA for survival analysis. As mentioned in the Introduction, the above-formulated WSI-based survival analysis may suffer from the domain shift issue in real applications. Therefore, our work is to develop a GDA method to alleviate this issue. Before that, we first formally describe the problem we are trying to solve. Suppose we have a set of training data from the so-called source domain, with known labels of the risk level, denoted as $\mathcal{D}^s = \{G_i^s, c_i^s, y_i^s\}_{i=1}^{N_s}$, where N_s is the number of data from the source domain. Then we can train a model to define the hazard function h using $\mathcal{D}^s$. Now a new dataset comes from the so-called target domain with unknown labels, denoted

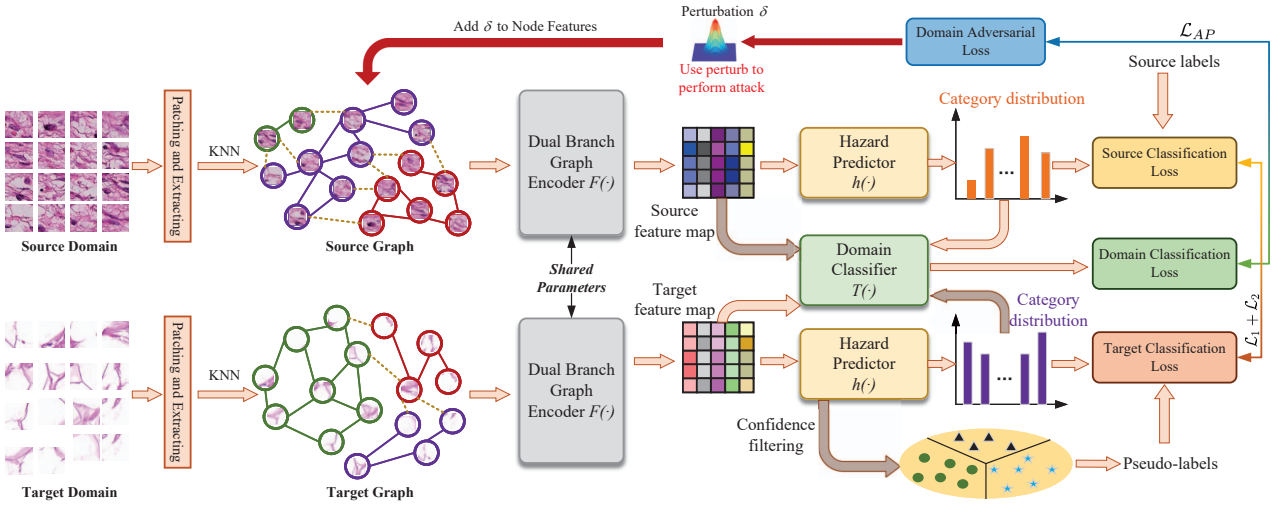


Figure 2. An overview of our proposed archicture DETA. The dual branch graph encoder utilizes the MP branch and the SP branch to explicitly and implicitly extract semantic information. To realize GDA, we propose a two-level alignment strategy to train the graph encoder and the hazard predictor, using both the labeled WSI graphs from the source domain and the unlabeled ones from the target domain.

as $\mathcal{D}^t = \{G_j^t\}_{j=1}^{N_t}$, where N_t is the number of data from the target domain, and we want to predict the risk level for each G_j^t using the trained h. However, since there is no guarantee that $\mathcal{D}^s$ and $\mathcal{D}^t$ have the same distribution, directly using h could have poor prediction performance. Therefore, we try to fine-tune h by aligning the distributions of the two domains to enable it to perform well on the target domain.

4. Proposed Method

In this section, we first briefly overview the proposed DETA framework, and then describe the details of its components, as well as the training strategy.

4.1. Framework Overview

Fig. 2 shows the overview of the proposed DETA framework. In the specific, the main network consists of a dual-branch graph encoder $F(\cdot)$ that intends to both explicitly and implicitly extract semantic information, respectively using the SP branch and the MP branch, which will be discussed in Section 4.2; and a hazard predictor $h(\cdot)$ that takes the feature output by F and predicts the risk level of the corresponding WSI, which is constructed with MLP. To realize GDA, we propose a two-level alignment strategy to train the graph encoder and the hazard predictor, using both the labeled WSI graphs from the source domain and the unlabeled ones from the target domain. Specifically, at the category level, we design the coupling dual evidence lower bounds (ELBOs), by virtue of the dual-branch structure of the graph encoder, to align the category distributions of the two domains; while at the feature level, we propose an adversarial perturbation strategy to add perturbations to the features extracted from the source domain in an adversarial way against that from the target one with the assist of the domain classifier $T(\cdot)$, such that the features distributions of the two domains are expected to be better aligned.

4.2. Design of the Dual-branch Graph Encoder

In existing studies, message passing neural networks (MPNNs) are often used to implicitly capture the topological semantics in the graph through neighborhood aggregation and use these representations to transfer knowledge between different domains [38, 49], which have shown their effectiveness. However, since both the graph structures and feature distributions of different domains can largely deviate from each other, only relying on the neighborhood information could not be sufficient enough to capture the graph semantics for GDA. Therefore, we adopt the graph kernel method to explicitly capture the high-order semantic relation in the graph [3, 13] using the SP branch. Combining the MP branch, the dual-branch graph encoder is constructed, as shown in Fig. 3. In the following, we discuss the detail of the two branches.

The MP Branch. The MP branch in our graph encoder consists of multiple stages of message passing process. One single message passing stage can be typically divided into two steps, i.e., message passing and node update. In the message passing step, the node v receives messages, i.e., features, from its neighbor nodes $\mathcal{N}(v)$ and aggregates them; and in the node update step, the feature of each node is updated by combining its current feature and the aggregation of neighbor features. This process can be written as

$$h_v^{(t+1)} = \text{UPD}\left(h_v^{(t)}, \text{AGG}\left(\left\{h_u^{(t)} : u \in \mathcal{N}(v)\right\}\right)\right) \quad (6)$$

where $h_v^{(t)}$ denotes the feature of node v at the t-th stage, UPD is the update function, and AGG is an aggregation function, following the standard GCN process [50].

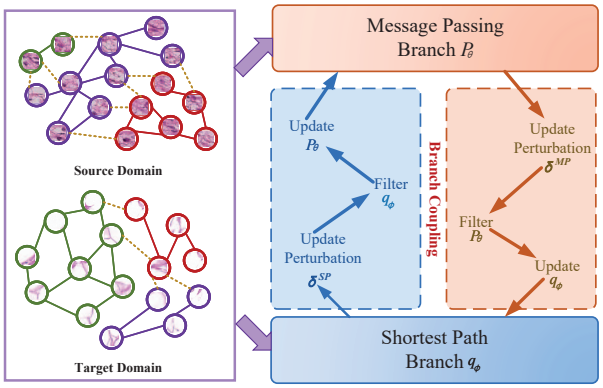


Figure 3. An overview of the proposed branch coupling.

The SP Branch. As mentioned before, the SP branch aims to explicitly capture the high-order semantic relation in the graph using the graph kernel method [3, 13]. Specifically, we generate a series of shortest paths from local substructures and extract higher-order and global semantic representations from the local information.

Let $\mathcal{N}_k(u)$ denote the set of all nodes reachable from node u via the shortest path of length k. In each layer of the SP branch, the representation of node u is updated by aggregating features all nodes from $\{\mathcal{N}_k(u)\}_{k=1}^K$, where K is a hyperparameter representing the maximum length of the shortest path. Furthermore, we introduce position encoding for the node representation $m^l_{u\in\mathcal{N}_k(u)}$ in each layer l to identify the node's position as follows:

$$\begin{aligned}\hat{m}^l_{u\in\mathcal{N}_k(u)} &= \sigma(m^l_{v\in\mathcal{N}_k(u)} + TE(k)),\\ TE(k)_{2i} = \sin\left(\frac{k}{10000^{2i/d}}\right), &TE(k)_{2i+1} = \cos\left(\frac{k}{10000^{2i/d}}\right),\end{aligned} \tag{7}$$

where σ is the activation function, $TE(k)$ is the position embeddings and $i \in [1, d]$ is the index of embeddings.

Then the update of the feature representation of node v according to different path lengths k can be summarized as

$$m^l_{u,\mathcal{N}_k(u)} = \mathcal{C}^l_{SP}\left(\hat{m}^{l-1}_u, \mathcal{A}^l_{SP}\left(\sum_{k=1}^{K}\left\{\hat{m}^{l-1}_u\right\}_{u\in\mathcal{N}_k(u)}\right)\right) \tag{8}$$

where $\mathcal{C}_{SP}$ and $\mathcal{A}_{SP}$ is the combination and aggregation operators, respectively.

4.3. Coupling Dual Branches for Category-level Alignment

Generating pseudo labels for the target domain is helpful for domain adaptation [10, 49]. Specifically, adding target domain samples with pseudo labels to the training process of the source domain can prevent the excessive shift of certain categories. However, due to the domain shift issue, the pseudo labels could be inaccurate and have a category distribution different from the source domain. Therefore, the category-level alignment is important to ensure the consistency of the category distribution of pseudo labels to that of the source domain. Notably, "category" means risk groups (e.g., low risk, high risk, etc.) based on patient survival time. To this end, by exploiting the characteristics of the dual-branch encoder structure, we proposed coupling the two branches by deriving ELBOs from two directions, to better align the category distributions of the two domains and enhance the quality of the pseudo label. In addition, we also use a filtering mechanism to retain high-confidence pseudo labels. It should be mentioned that, since the losses guide the training of the encoder, such a category-level alignment can also help align the features of two domains.

Now we introduce the idea of our coupling strategy. Let's first consider the GDA problem we face: given the source graph G^s, its label y^s and target graph G^t, we want to predict the label for G^t, denoted as $\hat{y}^t$, which can be treated as a latent variable here. Then we can write the following log-likelihood and its ELBO:

$$\begin{aligned}&\log p\left(y^s \mid G^s, G^t\right)\\ &\geq \mathbb{E}_{q(\hat{y}^t|G^t)}\left[\log p\left(y^s, \hat{y}^t \mid G^s, G^t\right) - \log q\left(\hat{y}^t \mid G^t\right)\right],\end{aligned} \tag{9}$$

where the p, q are seen as two parameterizations for the label distribution given a graph G. Traditionally, an ELBO is used to learn the approximation q to the target p. However, note that there are two branches in the graph encoder, and thus we can use the two branches to parameterize the two distributions, respectively. Furthermore, we can also exchange the positions of p and q, and thus the two branches can mutually act as targets and approximations to couple their training, thereby promoting the training efficacy.

To be specific, with p being fixed, the ELBO in Eq. (9) can be equivalently write as:

$$\begin{aligned}&KL\left(q\left(y^t \mid G^t\right) || p\left(y^t \mid G^s, G^t, y^s\right)\right)\\ &= -\mathbb{E}_{q(\hat{y}^t|G^t)}\left[\log p(y^s, \hat{y}^t|G^s, G^t) - \log q(\hat{y}^t|G^t)\right]\\ &+ \mathbb{E}_{q(\hat{y}^t, G^t)}\left[p\left(y^s_i \mid G^s_i\right)\right]\end{aligned} \tag{10}$$

Then denoting the distributions parameterized by MP and SP as p_θ and q_ϕ, using the coupling strategy introduced before, we get the following losses for updating MP and SP, respectively, with the other fixed:

$$\begin{aligned}\mathcal{L}_1 =&\mathbb{E}_{p_\theta(\hat{y}^t_i|G^t)>\zeta}\left[\log q_\phi\left(y^s_i, \hat{y}^t_i \mid G^s_i, G^t_i\right) - \log q_\phi\left(\hat{y}^t_i \mid G^t_i\right)\right]\\ &- \mathbb{E}_{q_\phi(y^s, G^s)} \log p_\theta\left(y^s_i \mid G^s_i\right)\\ \mathcal{L}_2 =&\mathbb{E}_{p_\phi(\hat{y}^t_i|G^t)>\zeta}\left[\log q_\theta\left(y^s_i, \hat{y}^t_i \mid G^s_i, G^t_i\right) - \log q_\theta\left(\hat{y}^t_i \mid G^t_i\right)\right]\\ &- \mathbb{E}_{q_\theta(y^s, G^s)} \log p_\phi\left(y^s_i \mid G^s_i\right)\end{aligned} \tag{11}$$

where $\hat{y}^t_i$ is the target pseudo-label filtered by the MP or SP branch. Note that we have introduced a threshold ζ when calculating the expectations, which acts as a confidence filter to select samples with high quality pseudo labels.

Table 1. The experimental results of survival analysis in one TCGA datasets as the training set and the other three datasets as the test sets. We highlight the top two best performing scores in red and blue, respectively.

Methods	Year	BLCA→LGG	BLCA→UCEC	BLCA→LUAD	LGG→BLCA	LGG→UCEC	LGG→LUAD
AttMIL [15]	2018	0.5445±0.0014	0.5561±0.0033	0.5582±0.0095	0.5147±0.0011	0.5159±0.0011	0.5280±0.0008
CLAM [24]	2021	0.5429±0.0076	0.5623±0.0019	0.5240±0.0048	0.5321±0.0031	0.5540±0.0065	0.5127±0.0005
TransMIL [33]	2021	0.5890±0.0012	0.5667±0.0074	0.5100±0.0005	0.5316±0.0015	0.5942±0.0011	0.5046±0.0023
DSMIL [20]	2021	0.5816±0.0037	0.5752±0.0094	0.5411±0.0013	0.5428±0.0021	0.5714±0.0067	0.5631±0.0018
STRAP [47]	2021	0.6134±0.0374	0.6218±0.0241	0.5799±0.0213	0.5826±0.0319	0.6077±0.0412	0.6152±0.0183
PathOmics [9]	2023	0.5861±0.0072	0.5758±0.0073	0.5732±0.0030	0.5490±0.0021	0.5513±0.0080	0.5672±0.0010
CMTA [55]	2023	0.5900±0.0076	0.5806±0.0041	0.5886±0.0028	0.5699±0.0014	0.5655±0.0024	0.5367±0.0005
RRTMIL [39]	2024	0.5714±0.0031	0.5738±0.0033	0.5712±0.0048	0.5511±0.0006	0.5738±0.0082	0.5719±0.0014
MoME [46]	2024	0.5974±0.0013	0.5769±0.0018	0.5862±0.0047	0.5794±0.0006	0.5628±0.0065	0.5872±0.0031
WiKG [21]	2024	0.5918±0.0021	0.5889±0.0110	0.5772±0.0007	0.5882±0.0024	0.5758±0.0064	0.5593±0.0018
SurvPath [18]	2024	0.6075±0.0045	0.5905±0.0077	0.5799±0.0019	0.5772±0.0027	0.5794±0.0017	0.5954±0.0019
CTL [43]	2024	0.6215±0.0216	0.6431±0.0192	0.5956±0.0243	0.5915±0.0372	0.5891±0.0280	0.6033±0.0147
ECB [43]	2024	0.6059±0.0491	0.6526±0.0372	0.5749±0.0184	0.5884±0,0591	0.5993±0.0374	0.6058±0.0417
DETA (Ours)	-	0.6566±0.0030	0.7026±0.0015	0.6259±0.0083	0.6227±0.0054	0.6528±0.0039	0.6452±0.0033

Methods	Year	UCEC→LGG	UCEC→BLCA	UCEC→LUAD	LUAD→LGG	LUAD→UCEC	LUAD→BLCA
AttMIL [15]	2018	0.5327±0.0034	0.5238±0.0047	0.5100±0.0032	0.5281±0.0051	0.5409±0.0120	0.5220±0.0008
CLAM [24]	2021	0.5178±0.0021	0.5150±0.0045	0.5442±0.0019	0.5525±0.0083	0.5387±0.0065	0.5118±0.0005
TransMIL [33]	2021	0.6137±0.0021	0.5660±0.0016	0.5180±0.0013	0.5345±0.0027	0.5461±0.0014	0.5263±0.0012
DSMIL [20]	2021	0.5507±0.0043	0.5524±0.0036	0.5079±0.0011	0.5553±0.0103	0.5535±0.0067	0.5387±0.0039
STRAP [47]	2021	0.6176±0.0025	0.6054±0.0073	0.5920±0.0022	0.6117±0.0044	0.6135±0.0034	0.5994±0.0077
PathOmics [9]	2023	0.5649±0.0069	0.5780±0.0013	0.5653±0.0015	0.5406±0.0074	0.5609±0.0023	0.5514±0.0039
CMTA [55]	2023	0.5992±0.0018	0.5589±0.0008	0.5655±0.0045	0.5510±0.0048	0.5650±0.0094	0.5827±0.0021
RRTMIL [39]	2024	0.5875±0.0017	0.5524±0.0008	0.5742±0.0018	0.5443±0.0045	0.5611±0.0008	0.5688±0.0026
MoME [46]	2024	0.5821±0.0006	0.5719±0.0020	0.5711±0.0014	0.5540±0.0023	0.6026±0.0034	0.5769±0.0036
WiKG [21]	2024	0.5160±0.0028	0.5404±0.0025	0.5658±0.0044	0.6077±0.0044	0.5641±0.0113	0.5653±0.0005
SurvPath [18]	2024	0.5973±0.0014	0.5917±0.0036	0.5769±0.0025	0.6123±0.0077	0.5714±0.0037	0.5625±0.0013
CTL [43]	2024	0.6094±0.0085	0.6113±0.0079	0.6145±0.0068	0.6146±0.0057	0.6231±0.0049	0.6038±0.0066
ECB [43]	2024	0.6059±0.0046	0..6051±0.0073	0.6110±0.0059	0.6127±0.0033	0.6286±0.0049	0.5847±0.0031
DETA (Ours)	-	0.6227±0.0049	0.6245±0.0069	0.6347±0.0036	0.6198±0.0061	0.6426±0.0018	0.6127±0.0108

4.4. Adversarial Perturbation for Feature-level Alignment

Adding perturbation in the source domain is an effective data augmentation strategy for better domain alignment [1]. An often adopted approach is to add random perturbation to the features of the source domain [26, 50]. However, although such randomness is expected to improve the generalization ability of the model, it could ignore key semantic information for alignment. Therefore, we introduce adaptively learnable perturbations to more effectively identify meaningful perturbation directions, such that the features of two domains can be more consistent in distribution.

Specifically, we add perturbations δ^{MP} and δ^{SP} to the node features of the source graph and input them to the MP and SP branch, respectively, and learn the perturbations in an adversarial way with the aid of a domain classifier $D(\cdot)$:

$$\min_{\|\delta^{MP}\|\leq\epsilon,\|\delta^{SP}\|\leq\epsilon} \max_{D} \mathcal{L}_{AP} = \mathbb{E}_{G^t\in\mathcal{D}^t} \log\left(1 - D(H^t, \hat{p}^t)\right) + \mathbb{E}_{G^s\in\mathcal{D}^s} \log D(H^s, \hat{p}^s), \tag{12}$$

where H^s is the output of the graph encoder with the perturbed input in source domain, H^t is the output without perturbation in target domain, and $\hat{p}_i^s$ and $\hat{p}_j^t$ are the category distributions of the two domains predicted by the hazard function. With such a adversarial training process, the feature distributions of both domains can be better aligned.

To optimize Eq. 12, we first randomly initialize the perturbation $\delta_t^{MP/SP}$ in the range $(-\epsilon, \epsilon)$. We then update $\delta^{MP/SP}$ using the gradient descent direction:

$$\delta_{t+1}^{MP/SP} = \delta_t^{MP/SP} - \epsilon\xi\left(\delta_t^{MP/SP}\right) / \left\|\xi\left(\delta_t^{MP/SP}\right)\right\|_F, \quad \xi\left(\delta_t^{MP/SP}\right) = \nabla_{\delta^{MP/SP}} \log D(\mathbf{H}^s), \tag{13}$$

where $\xi\left(\delta_t^{MP/SP}\right)$ is the gradient with respect to the perturbation $\delta^{MP/SP}$ and $t \in [1, T]$ is the update step size.

4.5. Model Training

The proposed method is implemented with a two-stage training. In the pre-training stage, we train the graph encoder $F(\cdot)$ and hazard predictor $h(\cdot)$ on the source domain dataset $\mathcal{D}$ using the survival loss defined in Eq. (5). Then in the GDA stage, fine-tune $F(\cdot)$ and $h(\cdot)$ with the combination of the survival loss and the losses proposed in Section 4.3 and 4.4. The overall training objective in the GDA stage can be summarized as follows:

$$\min_{\delta^{MP},\delta^{SP},h,S} \left\{\mathcal{L}_{surv} + \mathcal{L}_1 + \mathcal{L}_2 + \max_{D}\{\mathcal{L}_{AP}\}\right\}. \tag{14}$$

5. Experiments

In this section, we conduct extensive experiments on four TCGA datasets to evaluate the effectiveness of our proposed model. First, we describe the datasets and evaluation metrics used in the experiments. Then, we compare our model with some of the current state-of-the-art (SOTA) methods. Next, we conduct ablation studies to show the effectiveness of each key component. To further show the effectiveness

Table 2. The results of ablation studies in one TCGA datasets as the training set and the other three datasets as the test sets. We highlight the best performing scores in red.

Methods	BLCA→LGG	BLCA→UCEC	BLCA→LUAD	LGG→BLCA	LGG→UCEC	LGG→LUAD
w/o MP	0.5489$_{\downarrow 0.1077}$	0.6181$_{\downarrow 0.0854}$	0.5790$_{\downarrow 0.0469}$	0.5782$_{\downarrow 0.0445}$	0.6101$_{\downarrow 0.0427}$	0.5699$_{\downarrow 0.0753}$
w/o SP	0.5415$_{\downarrow 0.1151}$	0.6646$_{\downarrow 0.038}$	0.5687$_{\downarrow 0.0842}$	0.5687$_{\downarrow 0.0540}$	0.5976$_{\downarrow 0.0561}$	0.5735$_{\downarrow 0.0717}$
w/o δ^{MP}	0.5368$_{\downarrow 0.1198}$	0.6255$_{\downarrow 0.0771}$	0.5734$_{\downarrow 0.0525}$	0.5584$_{\downarrow 0.0643}$	0.6077$_{\downarrow 0.0451}$	0.5691$_{\downarrow 0.0761}$
w/o δ^{SP}	0.5249$_{\downarrow 0.1317}$	0.6134$_{\downarrow 0.0892}$	0.5669$_{\downarrow 0.0590}$	0.5531$_{\downarrow 0.0696}$	0.5929$_{\downarrow 0.0599}$	0.5617$_{\downarrow 0.0835}$
w/o AP	0.5201$_{\downarrow 0.1365}$	0.6079$_{\downarrow 0.0947}$	0.5618$_{\downarrow 0.0641}$	0.5490$_{\downarrow 0.0737}$	0.5905$_{\downarrow 0.0675}$	0.5600$_{\downarrow 0.0852}$
w/o BC	0.5347$_{\downarrow 0.1219}$	0.6357$_{\downarrow 0.0489}$	0.5721$_{\downarrow 0.0538}$	0.5544$_{\downarrow 0.0683}$	0.6031$_{\downarrow 0.0497}$	0.5684$_{\downarrow 0.0768}$
DETA (Ours)	0.6566	0.7026	0.6259	0.6227	0.6528	0.6452
Methods	UCEC→LGG	UCEC→BLCA	UCEC→LUAD	LUAD→LGG	LUAD→UCEC	LUAD→BLCA
w/o MP	0.5598$_{\downarrow 0.0629}$	0.5565$_{\downarrow 0.0680}$	0.5721$_{\downarrow 0.0626}$	0.5373$_{\downarrow 0.0825}$	0.5966$_{\downarrow 0.0460}$	0.5864$_{\downarrow 0.0263}$
w/o SP	0.5595$_{\downarrow 0.0632}$	0.5491$_{\downarrow 0.0754}$	0.5690$_{\downarrow 0.0657}$	0.5499$_{\downarrow 0.0699}$	0.5993$_{\downarrow 0.0433}$	0.5689$_{\downarrow 0.0429}$
w/o δ^{MP}	0.5517$_{\downarrow 0.0710}$	0.5556$_{\downarrow 0.0689}$	0.5718$_{\downarrow 0.0656}$	0.5441$_{\downarrow 0.0757}$	0.5964$_{\downarrow 0.0462}$	0.5882$_{\downarrow 0.0245}$
w/o δ^{SP}	0.5488$_{\downarrow 0.0739}$	0.5536$_{\downarrow 0.0709}$	0.5711$_{\downarrow 0.0636}$	0.5452$_{\downarrow 0.0774}$	0.5913$_{\downarrow 0.0513}$	0.5689$_{\downarrow 0.0438}$
w/o AP	0.5412$_{\downarrow 0.0815}$	0.5487$_{\downarrow 0.0758}$	0.5702$_{\downarrow 0.0645}$	0.5359$_{\downarrow 0.0839}$	0.5813$_{\downarrow 0.0613}$	0.5622$_{\downarrow 0.0505}$
w/o BC	0.5459$_{\downarrow 0.0768}$	0.5510$_{\downarrow 0.0735}$	0.5683$_{\downarrow 0.0664}$	0.5397$_{\downarrow 0.0801}$	0.5826$_{\downarrow 0.0600}$	0.5772$_{\downarrow 0.0355}$
DETA (Ours)	0.6227	0.6245	0.6347	0.6198	0.6426	0.6127

of our method, we also provide a survival analysis example, and visualize the feature and category alignment effects.

5.1. Datasets, Evaluation Metrics and Baselines

Datasets. The TCGA is a large-scale public database that collects genomic and clinical data of thousands of cancer patients, covering 33 common cancer types [42]. TCGA has been widely used in survival analysis, providing rich data support for identifying genetic variants. We selected prognostic data from TCGA to evaluate the performance of our proposed model. The datasets include: BLCA, LGG, LUAD, and UCEC. We used a 5-fold cross-validation in each data set, dividing the data into a training set and a validation set to ensure the generalization ability of the model.
Evaluation Metrics. C-index is a metric used to evaluate the model's ability to predict survival time ranking. It measures the consistency between the survival time order predicted by the model and the actual order. The value range of C-index is $[0, 1]$. The definition of C-index is as follows:

$$\text{C-index} = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} I(T_i < T_j)(1 - c_j) \tag{15}$$

where n represents the total number of samples, T_i and T_j represent the survival time of the i-th patient and the j-th patient, respectively, c_j is the censorship status.
Baselines. We compare the proposed DETA with representative methods, including the GNN-based methods: WiKG [21], Attention-based methods: AttMIL [15], CLAM [24], CMTA [55], MoME[46], Transformer-based methods: TransMIL [33], PathOmics [9], RRTMIL [39], SurvPath [18], CNN-based methods: DSMIL [20], and domain adaptation methods: STRAP [47], CTL [43], ECB [43].

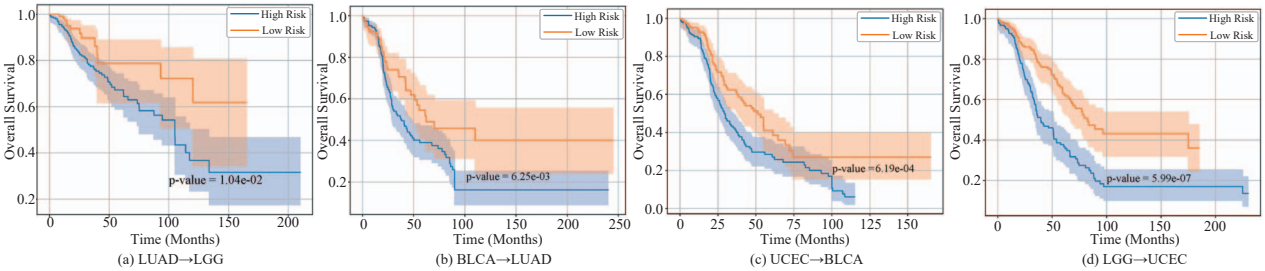
5.2. Comparisons with SOTA methods

As shown in Table 1, our proposed DETA method achieves the best performance on all TCGA datasets. This performance improvement is mainly due to the following two key factors: (1) Using MP branch and SP branch to extract key semantic information from an implicit and explicit perspective respectively significantly enhances the ability to capture global dependencies. (2) DETA further enhances cross-domain category and feature alignment through branch coupling and adversarial perturbation mechanisms. The branch coupling achieves efficient category alignment by fusing the output features of the MP and SP branch. By introducing adversarial perturbations into the source domain features, the model can better align the feature distribution in the target domain, effectively alleviating the distribution difference between the source domain and the target domain.

5.3. Ablation Studies

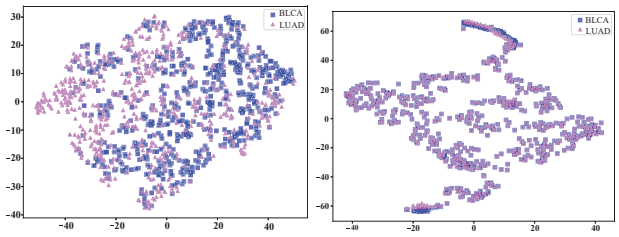
To comprehensively evaluate the impact of each module on the performance of DETA, we conducted ablation experiments. Specifically, we removed different modules one by one and evaluated the model performance. (1) w/o MP: both branches all using the SP branch. (2) w/o SP: both branches all using the MP branch. (3) w/o δ^{MP}: MP branch without perturbation. (4) w/o δ^{SP}: SP branch without perturbation. (5) w/o AP: The perturbation was removed from both the MP branch and the SP branch. (6) w/o BC: The branch coupling was removed.

From Table 2, several conclusions can be made. (1) DETA outperforms the variants w/o MP or w/o SP, which highlights the importance of extracting graph semantic information from both implicit and explicit perspectives. The dual-branch design can effectively make up for the shortcomings of a single branch and achieve comprehensive capture of multi-level graph semantics. (2) DETA performs significantly better than the models without perturbation modules (i.e., w/o δ^{MP}, w/o δ^{SP}, and w/o AP). The results show the important role of learnable perturbation modules in cross-domain tasks. (3) w/o BC performs better than w/o AP but worse than DETA. We attribute the fact that although the branch coupling is removed in the w/o BC variant, the perturbation module is still retained, which can alleviate the



(a) LUAD→LGG (b) BLCA→LUAD (c) UCEC→BLCA (d) LGG→UCEC

Figure 4. According to the predicted risks, all patients are stratified into low-risk and high-risk group, and Kaplan-Meier analysis is used to display the changes in the survival probability over time. We further use the Log-rank test to perform statistical significance testing.



(a) CMTA (b) Ours

Figure 5. t-SNE visualizations of the feature distributions of source and target domains by CMTA and our method.

domain distribution difference. However, due to the lack of category alignment, the model's performance is limited.

5.4. Survival Analysis

To further verify the effectiveness of the DETA model, we divided all patients into low-risk and high-risk groups based on the risk score predicted by DETA. Specifically, we used the median of the risk score as the boundary of the grouping and used the Kaplan-Meier survival curve analysis method to visualize the survival of all patients. As shown in Fig. 4, the analysis results intuitively show the survival differences between different risk groups. To evaluate the statistical significance of the survival difference, we also performed a Log-rank test. This test measures whether the survival difference between the low-risk group and the high-risk group is significant by calculating the p-value. Generally, a p-value less than or equal to 0.05 is considered statistically significant, which means that there is a significant difference in the survival curves between the two groups. It can be clearly seen from the results of Fig. 4 that the p-value is significantly less than 0.05 on all data sets, indicating the potential availability of the proposed method in survival analysis.

5.5. Visualization of the Feature and Category Alignment

To better show the mechanism of the proposed method for its promising performance, we visualize the feature and category alignment results in comparison with CMTA, which is one of the SOTA WSI-based survival analysis methods. Specifically, Fig. 5 shows the t-SNE embedding of the features extracted from the source and target domains by our model and CMTA, respectively, and we can see a clear feature distribution shift between the source domain and the target domain by CMTA. In contrast, our model minimizes the distribution divergence between the source and target domains. Furthermore, we also counted the empirical category distributions by both methods in Fig 6, and the result shows that the category distributions of the source and target domain by our method look very similar to each other, showing its promising category alignment effect.

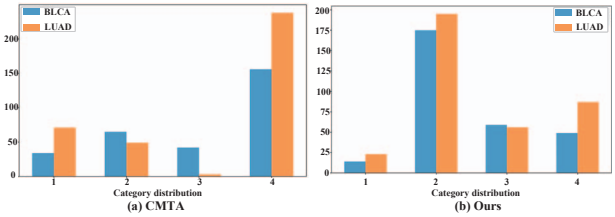


Figure 6. Visualizations of the category distributions of source and target domains by CMTA and our method.

6. Conclusions

In this paper, we propose a DETA framework to explore both feature and category-level alignment between different WSI domains. Specifically, we first formulate the concerned problem as GDA by virtue of the graph representation of WSIs. Then we construct a dual-branch graph encoder, including the MP branch and the SP branch, to explicitly and implicitly extract semantic information from the graph-represented WSIs. To realize GDA, we propose a two-level alignment approach: at the category level, we develop a coupling technique by virtue of the dual-branch structure, leading to reduced divergence between the category distributions of the two domains; at the feature level, we introduce an adversarial perturbation strategy to better augment source domain feature, resulting in improved alignment in feature distribution. Extensive experiments on four TCGA datasets have validated the effectiveness of our proposed DETA framework and demonstrated its superior performance in WSI-based survival analysis.

7. Acknowledgements

This work is supported by the National Nature Science Foundation of China (No. 62272375, No. 62272374, No. 62192781, No. 62250009, No. 62137002, No. 12471485, 62331028), Natural Science Foundation of Shaanxi Province (No.2024JC-JCQN-62), Project of China Knowledge Center for Engineering Science and Technology, Project of Chinese academy of engineering "The Online and Offline Mixed Educational Service System for 'The Belt and Road' Training in MOOC China", Tianyuan Fund for Mathematics of the National Natural Science Foundation of China (Grant No. 12426105), and the Major Key Project of Pengcheng Laboratory under Grant PCL2024A06.

References

[1] Faruk Ahmed, Andrew Sellergren, Lin Yang, Jinhua Xu, Boris Babenko, Abbi Ward, Niels Olson, Arash Mohtashamian, Yossi Matias, Greg Corrado, et al. Pathalign: A vision–language model for whole slide images in histopathology. In *MICCAI Workshop on Computational Pathology with Multimodal Data (COMPAYL)*. 2, 6

[2] Damiano Baroncini, Mauro Zaffaroni, Lucia Moiola, Lorena Lorefice, Giuseppe Fenu, Pietro Iaffaldano, Marta Simone, Fulvia Fanelli, Francesco Patti, Emanuele D'Amico, et al. Long-term follow-up of pediatric ms patients starting treatment with injectable first-line agents: a multicentre, italian, retrospective, observational study. *Multiple Sclerosis Journal*, 25(3):399–407, 2019. 1

[3] Karsten M Borgwardt and Hans-Peter Kriegel. Shortest-path kernels on graphs. In *Fifth IEEE International Conference on Data Mining (ICDM'05)*, pages 8–pp. IEEE, 2005. 4, 5

[4] Ruggero Capra, Cinzia Cordioli, Sarah Rasia, Fabio Gallo, Alessio Signori, and Maria Pia Sormani. Assessing long-term prognosis improvement as a consequence of treatment pattern changes in ms. *Multiple Sclerosis Journal*, 23(13): 1757–1761, 2017. 1, 2

[5] Richard J Chen, Chengkuan Chen, Yicong Li, Tiffany Y Chen, Andrew D Trister, Rahul G Krishnan, and Faisal Mahmood. 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

[6] Richard J Chen, Tong Ding, Ming Y Lu, Drew FK Williamson, Guillaume Jaume, Andrew H Song, Bowen Chen, Andrew Zhang, Daniel Shao, Muhammad Shaban, et al. Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 30(3):850–862, 2024. 2

[7] Quanyu Dai, Xiao-Ming Wu, Jiaren Xiao, Xiao Shen, and Dan Wang. Graph transfer learning via adversarial domain adaptation with graph convolution. *IEEE Transactions on Knowledge and Data Engineering*, 35(5):4908–4922, 2022. 3

[8] Donglin Di, Changqing Zou, Yifan Feng, Haiyan Zhou, Rongrong Ji, Qionghai Dai, and Yue Gao. Generating hypergraph-based high-order representations of whole-slide histopathological images for survival prediction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(5):5800–5815, 2022. 2, 3

[9] Kexin Ding, Mu Zhou, Dimitris N Metaxas, and Shaoting Zhang. Pathology-and-genomics multimodal transformer for survival outcome prediction. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 622–631. Springer, 2023. 6, 7

[10] Mucong Ding, Kezhi Kong, Jiuhai Chen, John Kirchenbauer, Micah Goldblum, David Wipf, Furong Huang, and Tom Goldstein. A closer look at distribution shifts and out-of-distribution generalization on graphs. In *NeurIPS 2021 Workshop on Distribution Shifts: Connecting Methods and Applications*. 5

[11] Kaituo Feng, Changsheng Li, Xiaolu Zhang, and JUN ZHOU. Towards open temporal graph neural networks. In *The Eleventh International Conference on Learning Representations*, 2023. 3

[12] Marco Francone, Franco Iafrate, Giorgio Maria Masci, Simona Coco, Francesco Cilia, Lucia Manganaro, Valeria Panebianco, Chiara Andreoli, Maria Chiara Colaiacomo, Maria Antonella Zingaropoli, et al. Chest ct score in covid-19 patients: correlation with disease severity and short-term prognosis. *European radiology*, 30:6808–6817, 2020. 2

[13] Hongyang Gao, Yi Liu, and Shuiwang Ji. Topology-aware graph pooling networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(12):4512–4518, 2021. 4, 5

[14] Gaoyang Guo, Chaokun Wang, Bencheng Yan, Yunkai Lou, Hao Feng, Junchao Zhu, Jun Chen, Fei He, and S Yu Philip. Learning adaptive node embeddings across graphs. *IEEE Transactions on Knowledge and Data Engineering*, 35(6): 6028–6042, 2022. 3

[15] Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *International conference on machine learning*, pages 2127–2136. PMLR, 2018. 6, 7

[16] Seidu Inusah, Maria P Sormani, Stacey S Cofield, Inmaculada B Aban, Solomon K Musani, Vinodh Srinivasasainagendra, and Gary R Cutter. Assessing changes in relapse rates in multiple sclerosis. *Multiple sclerosis journal*, 16(12): 1414–1421, 2010. 1

[17] Guillaume Jaume, Lukas Oldenburg, Anurag Vaidya, Richard J Chen, Drew FK Williamson, Thomas Peeters, Andrew H Song, and Faisal Mahmood. Transcriptomics-guided slide representation learning in computational pathology. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9632–9644, 2024. 2

[18] Guillaume Jaume, Anurag Vaidya, Richard J Chen, Drew FK Williamson, Paul Pu Liang, and Faisal Mahmood. Modeling dense multimodal interactions between biological pathways and histology for survival prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11579–11590, 2024. 1, 3, 6, 7

[19] EY Kalafi, NAM Nor, NA Taib, MD Ganggayah, C Town, and SK Dhillon. Machine learning and deep learning ap-

proaches in breast cancer survival prediction using clinical data. *Folia biologica*, 65(5-6):212–220, 2019. 1, 2

[20] Bin Li, Yin Li, and Kevin W Eliceiri. Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14318–14328, 2021. 6, 7

[21] Jiawen Li, Yuxuan Chen, Hongbo Chu, Qiehe Sun, Tian Guan, Anjia Han, and Yonghong He. Dynamic graph representation with knowledge-aware attention for histopathology whole slide image analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11323–11332, 2024. 6, 7

[22] Ruoyu Li, Jiawen Yao, Xinliang Zhu, Yeqing Li, and Junzhou Huang. Graph cnn for survival analysis on whole slide pathological images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 174–182. Springer, 2018. 2, 3

[23] Mingkai Lin, Wenzhong Li, Ding Li, Yizhou Chen, Guohao Li, and Sanglu Lu. Multi-domain generalized graph meta learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4479–4487, 2023. 3

[24] Ming Y Lu, Drew FK Williamson, Tiffany Y Chen, Richard J Chen, Matteo Barbieri, and Faisal Mahmood. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature biomedical engineering*, 5(6):555–570, 2021. 1, 2, 3, 6, 7

[25] Ming Y Lu, Bowen Chen, Andrew Zhang, Drew FK Williamson, Richard J Chen, Tong Ding, Long Phi Le, Yung-Sung Chuang, and Faisal Mahmood. Visual language pretrained multiple instance zero-shot transfer for histopathology images. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pages 19764–19775, 2023. 2

[26] Yadan Luo, Zijian Wang, Zhuoxiao Chen, Zi Huang, and Mahsa Baktashmotlagh. Source-free progressive graph learning for open-set domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):11240–11255, 2023. 3, 6

[27] Xinhong Ma, Junyu Gao, and Changsheng Xu. Active universal domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8968–8977, 2021. 3

[28] Massimiliano Mancini, Lorenzo Porzi, Samuel Rota Bulo, Barbara Caputo, and Elisa Ricci. Inferring latent domains for unsupervised deep domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(2):485–498, 2019. 3

[29] Ramin Nakhli, Allen Zhang, Ali Mirabadi, Katherine Rich, Maryam Asadi, Blake Gilks, Hossein Farahani, and Ali Bashashati. Co-pilot: Dynamic top-down point cloud with conditional neighborhood aggregation for multi-gigapixel histopathology image representation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21063–21073, 2023. 2

[30] Elke Platz, Allison A Merz, Pardeep S Jhund, Ali Vazir, Ross Campbell, and John J McMurray. Dynamic changes and prognostic value of pulmonary congestion by lung ultrasound in acute and chronic heart failure: a systematic review. *European journal of heart failure*, 19(9):1154–1163, 2017. 2

[31] Serena Ruggieri, Luca Prosperini, Sarmad Al-Araji, Pietro Osvaldo Annovazzi, Alvino Bisecco, Olga Ciccarelli, Nicola De Stefano, Massimo Filippi, Vinzenz Fleischer, Nikos Evangelou, et al. Assessing treatment response to oral drugs for multiple sclerosis in real-world setting: a magnims study. *Journal of Neurology, Neurosurgery & Psychiatry*, 95 (2):142–150, 2024. 1

[32] Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada. Maximum classifier discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3723–3732, 2018. 3

[33] Zhuchen Shao, Hao Bian, Yang Chen, Yifeng Wang, Jian Zhang, Xiangyang Ji, et al. Transmil: Transformer based correlated multiple instance learning for whole slide image classification. *Advances in neural information processing systems*, 34:2136–2147, 2021. 6, 7

[34] Jian Shen, Yanru Qu, Weinan Zhang, and Yong Yu. Wasserstein distance guided representation learning for domain adaptation. In *Proceedings of the AAAI conference on artificial intelligence*, 2018. 3

[35] Ankit Singh. Clda: Contrastive learning for semi-supervised domain adaptation. *Advances in Neural Information Processing Systems*, 34:5089–5101, 2021. 3

[36] Maria Pia Sormani and Paolo Bruzzi. Can we measure long-term treatment effects in multiple sclerosis? *Nature Reviews Neurology*, 11(3):176–182, 2015. 1

[37] Chetan L Srinidhi, Ozan Ciga, and Anne L Martel. Deep neural network models for computational histopathology: A survey. *Medical image analysis*, 67:101813, 2021. 2

[38] Mingchen Sun, Kaixiong Zhou, Xin He, Ying Wang, and Xin Wang. Gppt: Graph pre-training and prompt tuning to generalize graph neural networks. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1717–1727, 2022. 4

[39] Wenhao Tang, Fengtao Zhou, Sheng Huang, Xiang Zhu, Yi Zhang, and Bo Liu. Feature re-embedding: Towards foundation model-level performance in computational pathology. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11343–11352, 2024. 6, 7

[40] Luís A Vale-Silva and Karl Rohr. Long-term cancer survival prediction using multimodal deep learning. *Scientific Reports*, 11(1):13505, 2021. 1, 2

[41] Zhikang Wang, Jiani Ma, Qian Gao, Chris Bain, Seiya Imoto, Pietro Liò, Hongmin Cai, Hao Chen, and Jiangning Song. Dual-stream multi-dependency graph neural network enables precise cancer survival analysis. *Medical Image Analysis*, 97:103252, 2024. 2, 3

[42] John N Weinstein, Eric A Collisson, Gordon B Mills, Kenna R Shaw, Brad A Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, and Joshua M Stuart. The cancer genome atlas pan-cancer analysis project. *Nature genetics*, 45(10):1113–1120, 2013. 7

[43] Ruby Wood, Enric Domingo, Viktor Hendrik Koelzer, Timothy S Maughan, and Jens Rittscher. Cluster triplet loss for unsupervised domain adaptation on histology images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5122–5131, 2024. 6, 7

[44] Man Wu, Shirui Pan, and Xingquan Zhu. Attraction and repulsion: Unsupervised domain adaptive graph contrastive learning network. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 6(5):1079–1091, 2022. 3

[45] Qitian Wu, Hengrui Zhang, Junchi Yan, and David Wipf. Handling distribution shifts on graphs: An invariance perspective. In *International Conference on Learning Representations*, 2022. 3

[46] Conghao Xiong, Hao Chen, Hao Zheng, Dong Wei, Yefeng Zheng, Joseph JY Sung, and Irwin King. Mome: Mixture of multimodal experts for cancer survival prediction. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 318–328. Springer, 2024. 3, 6, 7

[47] Rikiya Yamashita, Jin Long, Snikitha Banda, Jeanne Shen, and Daniel L Rubin. Learning domain-agnostic visual representation for computational pathology using medically-irrelevant style transfer augmentation. *IEEE Transactions on Medical Imaging*, 40(12):3945–3954, 2021. 6, 7

[48] Xu Yang, Cheng Deng, Tongliang Liu, and Dacheng Tao. Heterogeneous graph attention network for unsupervised multiple-target domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(4):1992–2003, 2020. 3

[49] Gilad Yehudai, Ethan Fetaya, Eli Meirom, Gal Chechik, and Haggai Maron. From local structures to size generalization in graph neural networks. In *International Conference on Machine Learning*, pages 11975–11986. PMLR, 2021. 3, 4, 5

[50] Nan Yin, Li Shen, Mengzhu Wang, Long Lan, Zeyu Ma, Chong Chen, Xian-Sheng Hua, and Xiao Luo. Coco: A coupled contrastive framework for unsupervised domain adaptive graph classification. In *International Conference on Machine Learning*, pages 40040–40053. PMLR, 2023. 3, 4, 6

[51] Jan Yperman, Veronica Popescu, Bart Van Wijmeersch, Thijs Becker, and Liesbet M Peeters. Motor evoked potentials for multiple sclerosis, a multiyear follow-up dataset. *Scientific Data*, 9(1):207, 2022. 1

[52] Zhenjun Yu, Yu Zhang, Yingying Cao, Manman Xu, Shaoli You, Yu Chen, Bing Zhu, Ming Kong, Fangjiao Song, Shaojie Xin, et al. A dynamic prediction model for prognosis of acute-on-chronic liver failure based on the trend of clinical indicators. *Scientific reports*, 11(1):1810, 2021. 1, 2

[53] Yu Zhao, Fan Yang, Yuqi Fang, Hailing Liu, Niyun Zhou, Jun Zhang, Jiarui Sun, Sen Yang, Bjoern Menze, Xinjuan Fan, et al. Predicting lymph node metastasis using histopathological images based on multiple instance learning with deep graph convolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4837–4846, 2020. 2, 3

[54] Ziqi Zhao, Wentao Li, Ping Liu, Aili Zhang, Jianqi Sun, and Lisa X Xu. Survival analysis for multimode ablation using self-adapted deep learning network based on multisource features. *IEEE Journal of Biomedical and Health Informatics*, 28(1):19–30, 2023. 2, 3

[55] 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. 3, 6, 7

[56] Lai Zhouyou, Long Lifan, Cui Jiaqi, Xiong Deng, Liu Yuanjun, Yin Hao, and Wang Yan. Cmib: A cross modal interaction with information bottleneck framework for multimodal survival analysis. In *2024 IEEE International Conference on Cybernetics and Intelligent Systems (CIS) and IEEE International Conference on Robotics, Automation and Mechatronics (RAM)*, pages 526–530. IEEE, 2024. 2

[57] Qi Zhu, Carl Yang, Yidan Xu, Haonan Wang, Chao Zhang, and Jiawei Han. Transfer learning of graph neural networks with ego-graph information maximization. *Advances in Neural Information Processing Systems*, 34:1766–1779, 2021. 3