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

Supplementary Material

Abstract

In this supplementary material, we provide the mathematical derivations of the important equations used in the main text, complexity analysis, more implementation details, more experimental results and future work.

1. Derivations of Equations (9) and (10)

Following the nations used in the main text, the evidence lower bound (ELBO) in Eq. (9) of the main text can be derived as follows:

$$\begin{aligned}\log p(y^s|G^s,G^t) &= \log \int p(y^s,\hat{y}^t|G^s,G^t)d\hat{y}^t \\ &= \log \int \frac{p(y^s,\hat{y}^t|G^s,G^t)}{q(\hat{y}^t|G^t)} q(\hat{y}^t|G^t)d\hat{y}^t \\ &\geq \int q(\hat{y}^t|G^t) \log \frac{p(y^s,\hat{y}^t|G^s,G^t)}{q(\hat{y}^t|G^t)} d\hat{y}^t \\ &\geq \mathbb{E}_{q(\hat{y}^t|G^t)} \left[\log p(y^s,\hat{y}^t|G^s,G^t) - \log q(\hat{y}^t|G^t)\right] \\ &\triangleq \mathcal{L}(q)\end{aligned} \quad . \tag{1}$$

The right-hand side of the above equation, i.e., the ELBO can be further written as

$$\begin{aligned}\mathcal{L}(q) &= \mathbb{E}_{q(\hat{y}^t|G^t)} \left[\log p(\hat{y}^t|G^s,G^t,y^s)p(y^s|G^s) - \log q(\hat{y}^t|G^t)\right] \\ &= \mathbb{E}_{q(\hat{y}^t|G^t)} \left[\log \frac{p(\hat{y}^t|G^s,G^t,y^s)}{q(\hat{y}^t|G^t)}\right] + \mathbb{E}_{q(\hat{y}^t|G^t)}[p(y^s|G^s)] \\ &= -KL(q(y^t|G^t) \parallel p(y^t|G^s,G^t,y^s)) + \mathbb{E}_{q(\hat{y}^t|G^t)}[p(y^s|G^s)]\end{aligned} \quad . \tag{2}$$

The second term in the right-hand side of the above equation is a constant with respect to $q(\hat{y}^t|G^t)$ by fixing $p(y^s|G^s)$, and thus maximizing the ELBO is equivalent to minimizing the KL divergence between $q(\hat{y}^t|G^t)$ and $p(\hat{y}^t|G^s,G^t,y^s))$.

1.1. Complexity

Assuming a graph $G = (V,E)$ with N nodes and M edges, and a hidden size of d. In the MP branch, the time complexity for L-layer GCN is $O\left(L\left(Md+Nd^2\right)\right)$, the space complexity is $O(Ld^2+LNd)$. In the SP branch, the time complexity is $O(KN^2+NSLd^2+NSd)$, where S is the number of shortest paths. The space complexity is $O(Ld^2+N^2+NSd+NSLd)$. The time complexity of the domain discriminator is $O(NQd^2)$, and the space complexity is $O\left((2L+Q)d^2+NQd\right)$, where Q is the number of discriminator layers.

Algorithm 1 Learning Algorithm of DETA

Require: Source data $\mathcal{D}^s$; Target data $\mathcal{D}^t$.
Ensure: Parameters θ and ϕ for SP and MP branches.

// Dual Graph Branch for Semantics Mining
Initialize θ and ϕ.
Warm up the SP and MP branch to update θ and ϕ.
while not convergence **do**
  // Adaptive Perturbation for Domain Alignment
  Warm up $D(\cdot)$ by Eq. (12).
  Initialize each δ^{MP} and δ^{SP} in the range of $(-\epsilon,\epsilon)$.
  for $t=1,2,\ldots,T$ **do**
    Update SP branch perturbation δ^{SP} by Eq. (13).
  end for
  // Branch Coupling for Category Alignment
  Filter target pseudo-labels with the MP branch.
  Optimize parameters ϕ with fixed θ by Eq. (11).
  for $t=1,2,\ldots,T$ **do**
    Update MP branch perturbation δ^{MP} by Eq. (13).
  end for
  Filter target pseudo-labels with the SP branch.
  Optimize parameters θ with fixed ϕ by Eq. (11).
end while

2. Datasets

The Cancer Genome Atlas (TCGA) is a public, widely used database that contains genomic and clinical data from thousands of cancer patients, covering 33 common types of cancer, including but not limited to breast cancer, lung cancer, gastric cancer, liver cancer, etc. In this paper, we used prognostic data from five different cancer datasets in TCGA to evaluate our model. Specifically, the five selected cancer types include: Bladder Urothelial Carcinoma (BLCA), containing data from 373 patients; Breast Invasive Carcinoma (BRCA), containing data from 956 patients; Glioblastoma Multiforme and Lower Grade Glioma (GBMLGG), containing data from 569 patients; Lung Adenocarcinoma (LUAD), containing data from 453 patients; and Uterine Corpus Endometrial Carcinoma (UCEC), containing data from 480 patients. To comprehensively evaluate the performance of our model, we used a 5-fold cross-validation strategy for model training and validation on each dataset. In addition, we also compared our model with other existing comparison methods to further verify its effectiveness and advantages.

3. Implementation Details

In our DETA method, we designed two branches to conduct experiments: in the MP branch, we adopted GCN [5], and in the SP branch, we used the shortest path model [1]. Specifically, in the SP branch, we set the maximum path length K of all datasets to 5. For the adversarial perturbation module, we set the number of steps T of perturbation learning to 5. Meanwhile, we pre-trained the dual-branch model for 15 epochs and updated the branch coupling module 10 times on this baseline. In the pseudo-label generation process of the target dataset, we used the mean of the predicted probability as the pseudo-label filtering threshold ς to ensure the quality of the pseudo-label. For performance comparison, we chose to use one of the sub-datasets as the source dataset and the remaining sub-datasets as the target dataset to evaluate the performance of the model in the domain adaptation task. In terms of the setting of model parameters, we set the initial learning rate to 0.0005. All experiments were conducted on the same device equipped with NVIDIA A800 GPU to ensure the fairness and consistency of the experimental results. The whole training process is summarized in Algorithm 1.

4. More Experimental Results

Tables 1 and 2 show the comparison performance of DETA and baselines. From the results, we have a similar observation as we proposed in Section 5.2 of the main text. Furthermore, our method could possibly be extended to general GDA tasks as mentioned in the conclusion part of the paper. We have conducted preliminary experiments on the Office-Home dataset in comparison with a SOTA method in Table 3. Moreover, Fig. 1 shows the t-SNE embedding of the features extracted from the source and target domains by our proposed model, which is also consistent with that presented in Section 5.5 of the main text.

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.

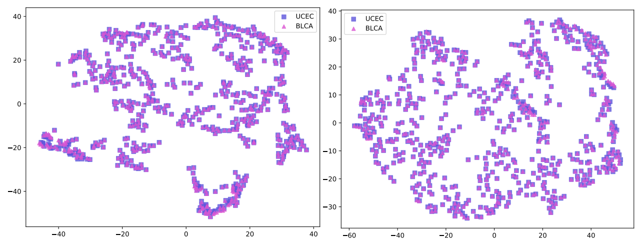
Methods	BRCA→LGG	BRCA→UCEC	BRCA→LUAD	BRCA→BLCA
AttMIL [3]	0.6566±0.0147	0.5728±0.0228	0.5619±0.0162	0.5578±0.0547
CLAM [8]	0.6259±0.0142	0.5436±0.0160	0.5507±0.0145	0.5329±0.0334
TransMIL [9]	0.6241±0.0271	0.5543±0.0179	0.5454±0.0013	0.5465±0.0020
DSMIL [6]	0.6643±0.0149	0.5881±0.0083	0.5726±0.0218	0.5737±0.0254
PathOmics [2]	0.6719±0.0301	0.5957±0.0144	0.5744±0.0258	0.5783±0.0277
CMTA [12]	0.6749±0.0287	0.5917±0.0341	0.5855±0.0192	0.5759±0.0358
RRTMIL [10]	0.6729±0.0043	0.6005±0.0019	0.5840±0.0138	0.5687±0.0212
MoME [11]	0.6824±0.0501	0.5992±0.0448	0.6137±0.0345	0.5886±0.0286
WiKG [7]	0.6915±0.0143	0.5867±0.0157	0.6003±0.0262	0.5828±0.0417
SurvPath [4]	0.7014±0.0270	0.6125±0.0228	0.5994±0.0493	0.6135±0.0341
DETA (Ours)	0.7530±0.0512	0.6744±0.0418	0.6657±0.0132	0.6513±0.0284

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	BRCA→LGG	BRCA→UCEC	BRCA→LUAD	BRCA→BLCA
w/o MP	$0.7087_{\downarrow 0.0443}$	$0.6181_{\downarrow 0.0563}$	$0.6226_{\downarrow 0.0431}$	$0.6143_{\downarrow 0.0388}$
w/o SP	$0.7004_{\downarrow 0.0526}$	$0.6318_{\downarrow 0.0426}$	$0.5963_{\downarrow 0.0694}$	$0.6101_{\downarrow 0.0412}$
w/o δ^{MP}	$0.6966_{\downarrow 0.0564}$	$0.6291_{\downarrow 0.0453}$	$0.6112_{\downarrow 0.0546}$	$0.5942_{\downarrow 0.0571}$
w/o δ^{SP}	$0.6806_{\downarrow 0.0724}$	$0.6157_{\downarrow 0.0587}$	$0.6061_{\downarrow 0.0596}$	$0.5947_{\downarrow 0.0584}$
w/o δ^{MP}/δ^{SP}	$0.6750_{\downarrow 0.0780}$	$0.6089_{\downarrow 0.0655}$	$0.5974_{\downarrow 0.0683}$	$0.5908_{\downarrow 0.0605}$
w/o BC	$0.6878_{\downarrow 0.0652}$	$0.6448_{\downarrow 0.0296}$	$0.6069_{\downarrow 0.0588}$	$0.5934_{\downarrow 0.0579}$
DETA (Ours)	0.7530	0.6744	0.6657	0.6513

Table 3. Accuracy on Office-Home across different domain shifts.

Methods	A→C	A→P	A→R	C→A	C→P	C→R	P→A	P→C	P→R
ECB [1]	68.5	85.4	88.3	79.2	86.8	89.0	79.3	66.4	88.5
DETA	71.9	86.8	88.7	83.4	87.5	89.1	84.0	69.6	89.7



(a) UCEC→BLCA (b) LGG→UCEC

Figure 1. t-SNE visualizations of the feature distributions of source and target domains by our method.

5. Future Work

In the future, we will explore the applicability of DETA to other WSI analysis tasks and more general problems that can be formulated as GDA. However, similar to other domain adaptation methods, DETA requires labeled source domain data. It may not directly apply to settings with limited supervision, e.g., few-shot or fully unsupervised scenarios. We will consider source-free methods in future work.

References

[1] Ralph Abboud, Radoslav Dimitrov, and Ismail Ilkan Ceylan. Shortest path networks for graph property prediction. In *Learning on Graphs Conference*, pages 5–1. PMLR, 2022. 2

[2] 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. 2

[3] Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *International conference on machine learning*, pages 2127–2136. PMLR, 2018. 2

[4] Guillaume Jaume, Anurag Vaidya, Richard J Chen, Drew FK Williamson, Paul Pu Liang, and Faisal Mahmood. Modeling dense multimodal interactions between biological path-

ways and histology for survival prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11579–11590, 2024. 2

[5] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*, 2022. 2

[6] 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. 2

[7] 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. 2

[8] 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. 2

[9] 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. 2

[10] 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. 2

[11] 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. 2

[12] 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. 2