Prototype-Guided Cross-Modal Knowledge Enhancement for Adaptive Survival Prediction

Fengchun Liu[1], Linghan Cai[1], Zhikang Wang[2], Zhiyuan Fan[1], Jin-gang Yu[3], Hao Chen[4], and Yongbing Zhang[1](✉)

[1] Harbin Institute of Technology(Shenzhen), Shenzhen, China
ybzhang08@hit.edu.cn
[2] Monash University, Melbourne, Australia
[3] South China University of Technology, Guangzhou, China
[4] The Hong Kong University of Science and Technology, Hong Kong, China

Abstract. Histo-genomic multimodal survival prediction has garnered growing attention for its remarkable model performance and potential contributions to precision medicine. However, a significant challenge in clinical practice arises when only unimodal data is available, limiting the usability of these advanced multimodal methods. To address this issue, this study proposes a prototype-guided cross-modal knowledge enhancement (ProSurv) framework, which eliminates the dependency on paired data and enables robust learning and adaptive survival prediction. Specifically, we first introduce an intra-modal updating mechanism to construct modality-specific prototype banks that encapsulate the statistics of the whole training set and preserve the modality-specific risk-relevant features/prototypes across intervals. Subsequently, the proposed cross-modal translation module utilizes the learned prototypes to enhance knowledge representation for multimodal inputs and generate features for missing modalities, ensuring robust and adaptive survival prediction across diverse scenarios. Extensive experiments on four public datasets demonstrate the superiority of ProSurv over state-of-the-art methods using either unimodal or multimodal input, and the ablation study underscores its feasibility for broad applicability. Overall, this study addresses a critical practical challenge in computational pathology, offering substantial significance and potential impact in the field. Codes are available at https://github.com/cyclexfy/ProSurv.

Keywords: Survival Analysis · Missing-Modality Learning · Prototype Learning · Cross-Modal Translation.

1 Introduction

Survival analysis seeks to estimate the risk of specific events, e.g. death or disease recurrence; consequently, it is crucial for disease progression estimation and treatment strategy selection in clinical practice [11,25]. Recently, whole-slide

F. Liu, L. Cai, and Z. Wang contributed equally to this work.

pathological images (WSIs) and genomic data have been widely used to model patient characteristics, with the first and second types of data capturing microscopic morphology features and quantitative molecular information, respectively.

Existing computational pathology methods can be roughly categorized as: WSI-based unimodal, genome-based unimodal, and histo-genomic multimodal methods. The unimodal methods, either WSI-based [1,6,9,16,20,22,23] or genome-based [13] ones, have demonstrated significant success in model performance and interpretation. However, the intrinsic complexity and heterogeneity of tumors urge the use of multimodal data to provide a more comprehensive characterization of patients. Born out of necessity, numerous multimodal approaches [4,10,27,30] have been developed to leverage the complementary and shared information provided by different data modalities. These approaches achieve better prediction accuracy by effectively combining the strengths of both modalities.

Although multimodal approaches have an advantage over unimodal ones in performance, they face a significant challenge in clinical practice [24], i.e., these developed methods do not apply to patients without complete data. This scenario is particularly common in current clinical systems, where patients often transfer between hospitals during the diagnostic phase or may not undergo specific examinations due to suboptimal conditions [18]. From the methodology perspective, there are two potential solutions. First, using knowledge distillation [2,8] to transfer multimodal knowledge to unimodal networks. However, this is constrained by low training efficiency and flexibility. Second, direct cross-modal reconstruction for multimodal knowledge completion is infeasible due to inherent heterogeneity between pathology and genomic modality [15].

To address the challenges, this paper proposes a prototype-guided cross-modal knowledge enhancement method (ProSurv). Specifically, through an intra-modal updating mechanism, we construct modality-specific prototype banks that capture risk-relevant features across intervals and encapsulate statistics of the whole training set. Afterward, the proposed cross-modal translation module utilizes prototypes as guidance to enhance knowledge representation for multimodal inputs and generate missing modality features for unimodal input. Consequently, our method eliminates the dependency on paired data and enables robust learning and adaptive survival prediction in diverse scenarios. We extensively evaluate ProSurv on four public datasets and demonstrate the superiority of our method over state-of-the-art methods using either unimodal or multimodal input. The ablation study underscores its feasibility for broad applicability.

2 Methodology

Fig. 1 illustrates the framework of ProSurv, which comprises three steps including data preprocessing and feature extraction, prototype bank update and cross-modal translation, and knowledge-enhanced learning. The following subsections introduce them in detail.

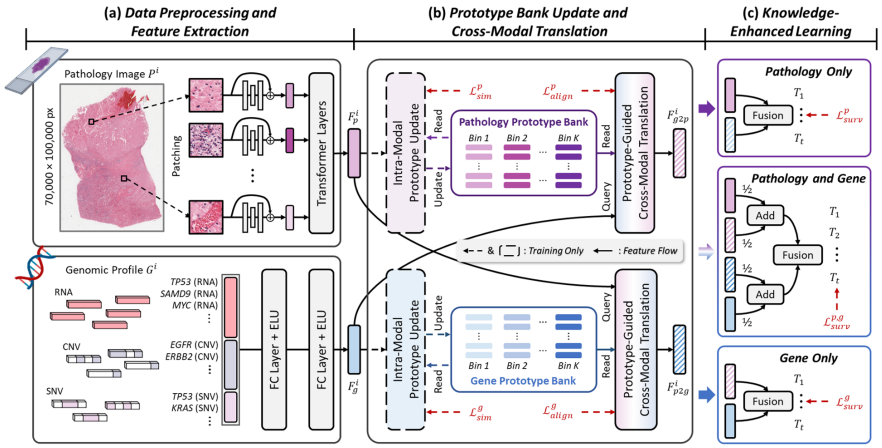

Fig. 1. Overview of the ProSurv. We first proceed with data preprocessing and feature extraction for WSIs and genomic data (a). Afterward, prototype banks are established corresponding to the input modalities and prototype-guided cross-modal translation modules enable the cross-modal feature translation (b). Eventually, the features from input data and translated modules jointly achieve the adaptive survival analysis (c).

2.1 Problem Formulation

Survival analysis [5] aims to predict time-to-event outcomes, where the event may not be fully observed. For the i-th patient, we model the survival and hazard functions as $f^i_{\text{surv}}(T \geq t \mid X^i)$ and $f^i_{\text{hazard}}(T = t \mid T \geq t, X^i)$, respectively. Here, $X^i = \{P^i, G^i, c^i, t^i\}$ represents the patient's data, where P^i, G^i, and $c^i \in \{0, 1\}$ are the WSI, genomic profiles, and censorship status, respectively, and $t^i \in \mathbb{R}^+$ indicates overall survival (in months). Here, survival time t^i is discretized into K intervals $\{bin_1, ..., bin_K\}$, allowing the model to estimate the hazard function for each bin. The survival function is approximated as: $f^i_{\text{surv}}(t_k) = \prod_{u=1}^{k}(1 - f^i_{\text{hazard}}(t_u))$, where t_k corresponds to the k-th bin. We learn the representation $f(X^i)$ to compute the survival loss $\mathcal{L}_{\text{surv}}(f(X^i), t^i, c^i)$ using the negative log-likelihood (NLL) loss function [29].

2.2 Data Preprocessing and Feature Extraction

Pathological Image Feature. Given an input WSI P^i from the i-th patient, we remove the non-tissue regions and then crop the rest into non-overlapping patches. The preprocessed data can be denoted as $\{p^i_j\}_{j=1}^{N^i}$, where N^i represents the total patch number. The patches are then fed into a pre-trained feature extractor $E_p(\cdot)$ to obtain a bag of patch features. Afterward, L Transformer layers [21,26] are employed to model relations between patches, followed by a global average pooling (GAP) to derive the WSI-level representation. The whole process can be formulated as: $F^i_p = \text{GAP}(\text{Transformer}^{(L)}(\{E_p(p^i_j)\}_{j=1}^{N^i})) \in \mathbb{R}^{1\times d}$.

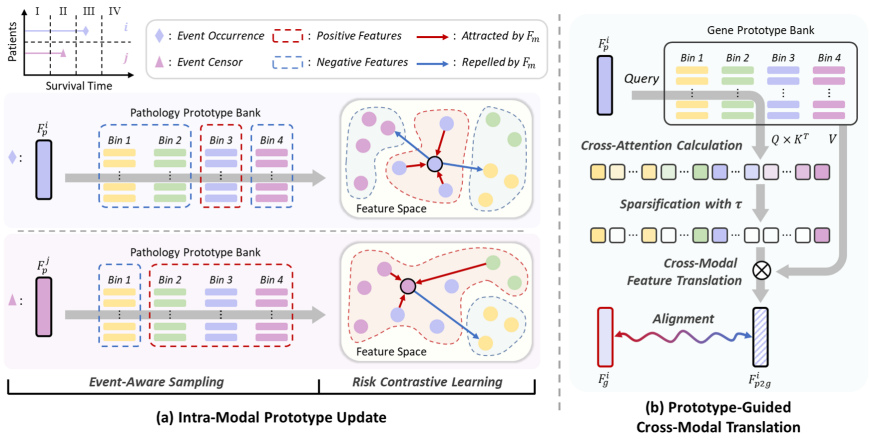

Fig. 2. Details of the intra-modal prototype update (a) and prototype-guided cross-modal translation (b). Here, we employ pathological features as an input for demonstration.

Genomic Feature. The input genomic data G^i consists of numerous 1×1 measurements $\{g_j^i\}_{j=1}^M$, where M means gene panel range. A Self-Normalizing Neural Network [13] $E_g(\cdot)$ is applied for feature extraction. This process can be formulated as $F_g^i = E_g(\{g_j^i\}_{j=1}^M) \in \mathbb{R}^{1\times d}$, where d is the feature dimension.

2.3 Intra-Modal Prototype Update

To preserve modality-specific knowledge, we construct pathological prototype bank $B_p \in \mathbb{R}^{K\times n\times d}$ and genomics prototype bank $B_g \in \mathbb{R}^{K\times n\times d}$ corresponding to the two modalities. Here, K is identical to the number of intervals and n is the number of prototypes in each bank. To enable prototypes with more risk-relevant knowledge, we develop an intra-modal prototype updating mechanism (Fig. 2(a)).

Given the extracted features from the i-th patient, we calculate the cosine similarity between the features and the corresponding prototypes recorded in the bank. Taking the pathological feature F_p^i as an example, we can obtain:

$$w_p^{i,k} = \frac{1}{K}\sum_{j=1}^{n} w_p^{i,k,j}, \text{ where } w_p^{i,k,j} = \frac{\text{Norm}(F_p^i)\cdot\text{Norm}(B_p^{k,j})}{\|\text{Norm}(F_p^i)\|\|\text{Norm}(B_p^{k,j})\|}. \quad (1)$$

In Eq. 1, Norm$(\cdot)$ is a min-max normalization, $\|\cdot\|$ denotes the Euclidean Norm, $B_p^{k,j}$ indicates the j-th prototype in the k-th bin of the pathology prototype bank, $w_p^{i,k}$ measures the average value between the input data and the prototypes of the k-th bin. To enhance prototype representations, we employ risk contrastive learning, which pulls prototypes from the same time interval closer to the input feature and pushes those from different intervals apart. Considering

the complexity of survival data, we introduce event-aware sampling. Specifically, for uncensored data, prototypes from the bin corresponding to the label are treated as positive samples, while others are negative. For censored data, the censoring time is treated as a key node: prototypes before the censoring time are negative samples, while those after it are positive, as the patient is alive before the censoring time. Hence, the entire procedure can be summarized as follows:

$$\mathcal{L}_{\text{sim}}^{p} = (1-c^i)(-w_p^{t^i} + \sum_{t^k \neq t^i} \frac{w_p^{t^k}}{K-1}) + c^i(-\sum_{t^k \geq t^i} \frac{w^{t^k}}{K-t^i+1} + \sum_{t^k < t^i} \frac{w^{t^k}}{t^i-1}), \quad (2)$$

where t^i represents the the bin corresponding to the event occurrence.

In the training process, the intra-modal prototype update mechanisms are applied to each modality, which can be formulated as: $\mathcal{L}_{\text{sim}}^{p,g} = \mathcal{L}_{\text{sim}}^{p} + \mathcal{L}_{\text{sim}}^{g}$.

2.4 Prototype-Guided Cross-Modal Translation

The prototype-guided cross-modal translation aims to utilize the input features along with the prototype banks to reconstruct cross-modal features, as shown in Fig. 2(b). The function is mathematically achieved through a cross-attention mechanism [21]. Here, we take the i-th patient as an example. When reconstructing genomic features, pathological features F_p^i serves as the query Q_p to interact with keys K_g, the genomic prototype bank, producing attention scores reflecting histo-genomics associations. Afterward, the values V_g derived from the genomic bank are multiplexed with the attention scores, generating the final reconstructed cross-modal features. The whole process can be formulated as follows:

$$F_{p2g}^i = \sigma\left(\frac{(W_p^q F_p^i)(W_g^k B_g)^T}{\tau\sqrt{d}}\right)(W_g^v B_g), \quad (3)$$

where $\sigma(\cdot)$ is a softmax function, T and τ represent a transpose operation and a temperature coefficient, respectively. Similarly, translating genomic features F_g^i to the pathological ones can be formulated as: $F_{g2p}^i = \sigma\left(\frac{(W_g^q F_g^i)(W_p^k B_p)^T}{\tau\sqrt{d}}\right)(W_p^v B_p)$.

To enhance the learning process, we selectively align reconstructed features with their original counterparts when the complete modality pairs are available. The corresponding loss $\mathcal{L}_{\text{align}}$ is defined as:

$$\mathcal{L}_{\text{align}}^{p,g} = \mathcal{L}_{\text{align}}^{p} + \mathcal{L}_{\text{align}}^{g} = \|F_p^i - F_{g2p}^i\|^2 + \|F_g^i - F_{p2g}^i\|^2. \quad (4)$$

2.5 Knowledge-Enhanced Learning and Prediction

Knowledge-enhanced learning (Fig. 1(c)) is designed to conduct adaptive survival prediction regardless of the input data conditions. Here, we introduce two cases: the patient with complete data modalities and incomplete data modality.

Complete Modalities. With the complete modalities, we can obtain the original features F_p^i, F_g^i, and the translated features F_{g2p}^i, F_{p2g}^i. Then, we fuse the

Table 1. Performance comparison using C-index (mean ± standard deviation) on four cancer datasets. The best result is shown in red, and the second-best one is in blue. "P" and "G" are abbreviations of pathology and gene, respectively. "∗" represents the multimodal training and unimodal testing scenario.

Methods	Modal		Datasets				Overall
	P	G	BRCA	BLCA	STAD	CRAD	
ABMIL	✓	×	0.613±0.118	0.575±0.050	0.561±0.059	0.586±0.115	0.584
TransMIL	✓	×	0.625±0.080	0.614±0.046	0.522±0.050	0.525±0.118	0.572
WIKG	✓	×	0.613±0.086	0.585±0.047	0.551±0.049	0.517±0.127	0.569
MambaMIL	✓	×	0.617±0.078	0.612±0.058	0.574±0.041	0.588±0.095	0.598
SNN	×	✓	0.576±0.112	0.539±0.052	0.539±0.032	0.616±0.090	0.568
SNNTrans	×	✓	0.551±0.097	0.553±0.043	0.555±0.057	0.614±0.125	0.568
MCAT	✓	✓	0.654±0.077	0.616±0.055	0.572±0.098	0.615±0.141	0.614
MOTCAT	✓	✓	0.665±0.111	0.600±0.051	0.563±0.084	0.593±0.169	0.605
CMTA	✓	✓	0.604±0.045	0.642±0.072	0.577±0.098	0.586±0.075	0.602
SurvPath	✓	✓	0.675±0.069	0.584±0.040	0.551±0.040	0.585±0.100	0.599
G-HANet	✓	∗	0.677±0.071	0.603±0.054	0.588±0.050	0.598±0.169	0.617
ProSurv	✓	∗	0.701±0.101	0.611±0.032	0.620±0.069	0.626±0.090	0.640
	∗	✓	0.578±0.110	0.646±0.046	0.540±0.079	0.595±0.056	0.590
	✓	✓	0.675±0.070	0.655±0.047	0.609±0.073	0.641±0.088	0.645

original features with the corresponding translated features using an averaging function and then utilize the concatenated features as the final patient representation for survival prediction. The process can be formulated as follows:

$$H_{p,g} = \phi(\text{Concat}(F^i_{ep}, F^i_{eg})), \text{ where } F^i_{ep} = \frac{F^i_p + F^i_{g2p}}{2}, \ F^i_{eg} = \frac{F^i_g + F^i_{p2g}}{2}. \quad (5)$$

Here, $\phi(\cdot)$ is a full-connected (FC) layer followed by a sigmoid function, Concat means concatenation. Under this scenario, the training loss is expressed as: $\mathcal{L}^{p,g}_{\text{total}} = \mathcal{L}^{p,g}_{\text{surv}} + \alpha\mathcal{L}^{p,g}_{\text{sim}} + \beta\mathcal{L}^{p,g}_{\text{align}}$, where α and β are hyperparameters for balance.

Incomplete Modality. When the patient only has one kind of data, ProSurv also can perform robust learning and prediction. For example, with only WSI input, ProSurv can generate the translated genomic feature F^i_{p2g} from pathological features F^i_p. Afterward, we fuse these features for the final prediction H_p:

$$H_p = \phi(\text{Concat}(F^i_p, F^i_{p2g})). \quad (6)$$

Vice versa as: $H_g = \phi(\text{Concat}(F^i_g, F^i_{g2p}))$. The training loss under the incomplete data input is considered: $\mathcal{L}^p_{\text{total}} = \mathcal{L}^p_{\text{surv}} + \alpha\mathcal{L}^p_{\text{sim}}$ (only pathology input) or $\mathcal{L}^g_{\text{total}} = \mathcal{L}^g_{\text{surv}} + \alpha\mathcal{L}^g_{\text{sim}}$ (only gene input).

3 Experiments and Results

3.1 Datasets & Evaluation Metric & Implementation Details

Experiments were performed on four public datasets from The Cancer Genome Atlas (TCGA). Specifically, we used 868 cases of Breast Invasive Carcinoma

Table 2. Ablation study. For each testing scenario, the best performance is highlighted in red, while the second-best performance is in blue. "w/o" means without.

Methods	Modal		Datasets				Overall
	P	G	BRCA	BLCA	STAD	CRAD	
w/o Prototypes	✓	*	0.660±0.068	0.601±0.041	0.503±0.047	0.617±0.111	0.595
w/o $\mathcal{L}_{\text{sim}}$	✓	*	0.639±0.069	0.618±0.043	0.533±0.104	0.618±0.131	0.602
w/o $\mathcal{L}_{\text{align}}$	✓	*	0.687±0.079	0.616±0.042	0.534±0.100	0.616±0.128	0.613
ProSurv	✓	*	0.701±0.101	0.611±0.032	0.620±0.069	0.626±0.090	0.640
w/o Prototypes	*	✓	0.500±0.085	0.625±0.058	0.517±0.073	0.549±0.054	0.548
w/o $\mathcal{L}_{\text{sim}}$	*	✓	0.580±0.123	0.634±0.048	0.553±0.127	0.565±0.058	0.583
w/o $\mathcal{L}_{\text{align}}$	*	✓	0.554±0.079	0.644±0.044	0.558±0.134	0.556±0.070	0.578
ProSurv	*	✓	0.578±0.110	0.646±0.046	0.540±0.079	0.595±0.056	0.590
w/o Prototypes	✓	✓	0.648±0.060	0.637±0.064	0.495±0.059	0.623±0.108	0.601
w/o $\mathcal{L}_{\text{sim}}$	✓	✓	0.610±0.087	0.648±0.047	0.531±0.112	0.610±0.129	0.600
w/o $\mathcal{L}_{\text{align}}$	✓	✓	0.662±0.058	0.654±0.047	0.528±0.117	0.606±0.124	0.613
ProSurv	✓	✓	0.675±0.070	0.655±0.047	0.609±0.073	0.641±0.088	0.645

(BRCA), 359 cases of Bladder Urothelial Carcinoma (BLCA), 318 cases of Stomach Adenocarcinoma (STAD), and 294 cases of Colon and Rectum Adenocarcinoma (CRAD). All WSIs were processed at 20× magnification using CLAM [16] and genomic data was processed using min-max normalization. For each dataset, we randomly split the data into training, validation, and test sets with a ratio of 6:2:2, and reported the average C-index [7] across 5-fold cross-validation.

ProSurv is implemented on PyTorch 1.12.1 [19] using an NVIDIA RTX 4090 GPU. Patch-level features are extracted using the UNI model [3]. 4,096 patches out of the whole image are randomly selected from each WSI during training, while all patches are used for validation and testing. The Adam optimizer [12] is employed for model optimization with a constant learning rate of 1e-4 and a weight decay of 1e-4. Each model is trained for a maximum of 50 epochs, and the best-performing model on the validation set is used for testing. We set the hyper-parameters as: $K = 4$, $n = 32$, $\tau = 0.5$, $\alpha = 0.2$, and $\beta = 0.2$.

3.2 Model Performance Comparison

We compared the ProSurv with 11 state-of-the-art survival analysis methods, ranging from image-based [9,14,20,28], genome-based [13], multimodal [4,10,27,30], and multimodal training unimodal testing method [24] under various scenarios.

Table 1 illustrates the quantitative results, from which we can derive the following observations: (1) Multimodal methods generally outperform unimodal methods, highlighting that multimodal data contains more survival-related information than unimodal data. (2) ProSurv achieves state-of-the-art performance with multimodal inference, obtaining an overall C-index of 0.645. This indicates the superiority of the knowledge enhancement brought by prototype-guided cross-translation. (3) ProSurv excels in inference using unimodal data. When using only pathology images, ProSurv achieves a promising C-index of 0.640, surpassing G-HANet (0.617) and MCAT (0.614) significantly. This demonstrates

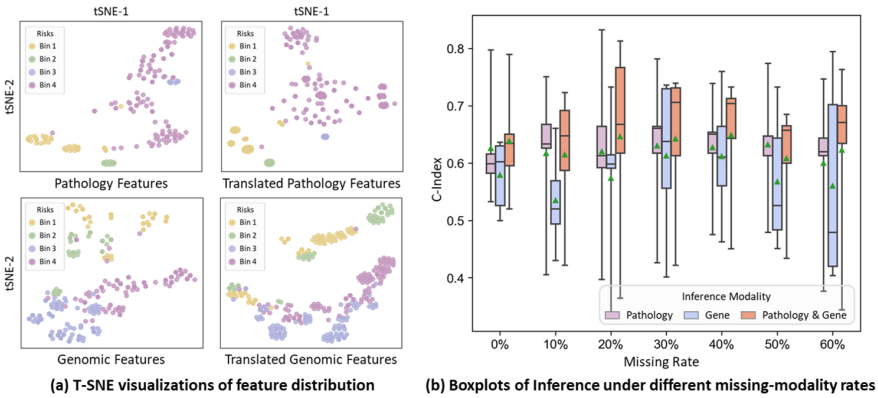

Fig. 3. Distribution visualizations of original and translated features (a). Robustness evaluation using a mixture of uni- and multimodal training data on TCGA-CRAD (b).

that the prototypes capture the multimodal knowledge and can effectively enhance the patients' representations for survival prediction using unimodal data.

3.3 Ablation Study

In this subsection, we perform ablation experiments to validate the effectiveness of each module, with experimental results in Table 2 and Fig. 3.

Effectiveness of Prototypes. We replace the cross-modal translation modules with multi-layer perceptrons. The overall C-index decreases by 4.2% under various scenarios. This is expected, as discarding prototypes means losing modality-specific knowledge, reducing the knowledge throughout the computation process.

Effectiveness of Intra-Modal Prototype Update. The use of intra-modal prototype update makes the prototypes effectively capture risk-relevant knowledge, with 3.8%, 0.7%, and 4.5% C-index improvements under pathology-only, gene-only, and complete modality scenes, respectively.

Effectiveness of Prototype-Guided Cross-Modal Translation. Alignment loss $\mathcal{L}_{\text{align}}$ is used to optimize translated features in ProSurv. As shown in Table 2, abandoning the loss leads to inferior model performance. Additionally, Fig. 3(a) visualizes the feature distributions of original and translated features using t-SNE [17]. The highly identical distributions of the counterpart features reveal the reality of the translated features and the effectiveness of the module.

Model Robustness Evaluation. Different from existing methods, ProSurv can adaptively learn knowledge from incomplete data. Fig. 3(b) illustrates the model performance across varying rates of unimodal data while training the neural network. When the rate is lower than 50%, the performance of ProSurv does not degrade significantly, showing its powerful learning capability and providing an insightful solution for adaptive survival analysis.

4 Conclusion

In this paper, we present ProSurv, a prototype-guided cross-modal knowledge enhancement framework for adaptive survival analysis. The key innovations of ProSurv lie in using prototype banks to generate cross-modal features and achieve an adaptive survival prediction using arbitrary input data. Extensive experiments including performance comparison and ablation study underscore the superiority and robustness of ProSurv. ProSurv addresses a critical practical challenge in computational pathology, offering substantial significance for precision medicine.

Acknowledgments. This work was supported in part by the National Natural Science Foundation of China under 62031023 & 62331011; in part by the Shenzhen Science and Technology Project under GXWD20220818170353009.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Cai, L., Huang, S., Zhang, Y., Lu, J., Zhang, Y.: Attrimil: Revisiting attention-based multiple instance learning for whole-slide pathological image classification from a perspective of instance attributes. Medical Image Analysis p. 103631 (2025)
2. Chen, C., Dou, Q., Jin, Y., Liu, Q., Heng, P.A.: Learning with privileged multimodal knowledge for unimodal segmentation. IEEE transactions on medical imaging **41**(3), 621–632 (2021)
3. Chen, R.J., Ding, T., Lu, M.Y., Williamson, D.F., Jaume, G., Chen, B., Zhang, A., Shao, D., Song, A.H., Shaban, M., et al.: Towards a general-purpose foundation model for computational pathology. Nature Medicine (2024)
4. Chen, R.J., Lu, M.Y., Weng, W.H., Chen, T.Y., Williamson, D.F., Manz, T., Shady, M., Mahmood, F.: Multimodal co-attention transformer for survival prediction in gigapixel whole slide images. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4015–4025 (2021)
5. Clark, T.G., Bradburn, M.J., Love, S.B., Altman, D.G.: Survival analysis part i: basic concepts and first analyses. British journal of cancer **89**(2), 232–238 (2003)
6. Dong, J., Jiang, J., Jiang, K., Li, J., Cai, L., Zhang, Y.: Disentangled pseudo-bag augmentation for whole slide image multiple instance learning. IEEE Transactions on Medical Imaging (2025)
7. Harrell, F.E., Califf, R.M., Pryor, D.B., Lee, K.L., Rosati, R.A.: Evaluating the yield of medical tests. Jama **247**(18), 2543–2546 (1982)
8. Hu, M., Maillard, M., Zhang, Y., Ciceri, T., La Barbera, G., Bloch, I., Gori, P.: Knowledge distillation from multi-modal to mono-modal segmentation networks. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part I 23. pp. 772–781. Springer (2020)
9. Ilse, M., Tomczak, J., Welling, M.: Attention-based deep multiple instance learning. In: International conference on machine learning. pp. 2127–2136. PMLR (2018)

10. Jaume, G., Vaidya, A., Chen, R.J., Williamson, D.F., Liang, P.P., Mahmood, F.: 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. pp. 11579–11590 (2024)
11. Jiang, L., Xu, C., Bai, Y., Liu, A., Gong, Y., Wang, Y.P., Deng, H.W.: Autosurv: interpretable deep learning framework for cancer survival analysis incorporating clinical and multi-omics data. NPJ precision oncology **8**(1), 4 (2024)
12. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
13. Klambauer, G., Unterthiner, T., Mayr, A., Hochreiter, S.: Self-normalizing neural networks. Advances in neural information processing systems **30** (2017)
14. Li, J., Chen, Y., Chu, H., Sun, Q., Guan, T., Han, A., He, Y.: 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. pp. 11323–11332 (2024)
15. Li, R., Wu, X., Li, A., Wang, M.: Hfbsurv: hierarchical multimodal fusion with factorized bilinear models for cancer survival prediction. Bioinformatics **38**(9), 2587–2594 (2022)
16. Lu, M.Y., Williamson, D.F., Chen, T.Y., Chen, R.J., Barbieri, M., Mahmood, F.: Data-efficient and weakly supervised computational pathology on whole-slide images. Nature biomedical engineering **5**(6), 555–570 (2021)
17. Van der Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of machine learning research **9**(11) (2008)
18. Ning, Z., Zhao, Z., Feng, Q., Chen, W., Xiao, Q., Zhang, Y.: Mutual-assistance learning for standalone mono-modality survival analysis of human cancers. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(6), 7577–7594 (2022)
19. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems **32** (2019)
20. Shao, Z., Bian, H., Chen, Y., Wang, Y., Zhang, J., Ji, X., et al.: Transmil: Transformer based correlated multiple instance learning for whole slide image classification. Advances in neural information processing systems **34**, 2136–2147 (2021)
21. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
22. Wang, Z., Gao, Q., Yi, X., Zhang, X., Zhang, Y., Zhang, D., Liò, P., Bain, C., Bassed, R., Li, S., et al.: Surformer: An interpretable pattern-perceptive survival transformer for cancer survival prediction from histopathology whole slide images. Computer Methods and Programs in Biomedicine **241**, 107733 (2023)
23. Wang, Z., Ma, J., Gao, Q., Bain, C., Imoto, S., Liò, P., Cai, H., Chen, H., Song, J.: Dual-stream multi-dependency graph neural network enables precise cancer survival analysis. Medical Image Analysis **97**, 103252 (2024)
24. Wang, Z., Zhang, Y., Xu, Y., Imoto, S., Chen, H., Song, J.: Histo-genomic knowledge association for cancer prognosis from histopathology whole slide images. IEEE Transactions on Medical Imaging (2025)
25. Xiang, J., Wang, X., Zhang, X., Xi, Y., Eweje, F., Chen, Y., Li, Y., Bergstrom, C., Gopaulchan, M., Kim, T., et al.: A vision–language foundation model for precision oncology. Nature pp. 1–10 (2025)

26. Xiong, Y., Zeng, Z., Chakraborty, R., Tan, M., Fung, G., Li, Y., Singh, V.: Nyströmformer: A nyström-based algorithm for approximating self-attention. In: Proceedings of the AAAI conference on artificial intelligence. vol. 35, pp. 14138–14148 (2021)
27. Xu, Y., Chen, H.: Multimodal optimal transport-based co-attention transformer with global structure consistency for survival prediction. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 21241–21251 (2023)
28. Yang, S., Wang, Y., Chen, H.: Mambamil: Enhancing long sequence modeling with sequence reordering in computational pathology. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 296–306. Springer (2024)
29. Zadeh, S.G., Schmid, M.: Bias in cross-entropy-based training of deep survival networks. IEEE transactions on pattern analysis and machine intelligence **43**(9), 3126–3137 (2020)
30. Zhou, F., Chen, H.: Cross-modal translation and alignment for survival analysis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 21485–21494 (2023)