This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

PathM3: A Multimodal Multi-Task Multiple Instance Learning Framework for Whole Slide Image Classification and Captioning

Qifeng Zhou[1], Wenliang Zhong[1], Yuzhi Guo[1], Michael Xiao[1,2], Hehuan Ma[1], and Junzhou Huang[1⋆]

Department of Computer Science and Engineering, The University of Texas at Arlington, Arlington, TX 76019, USA
Colleyville Heritage High School, Colleyville, TX 76034, USA
jzhuang@uta.edu

Abstract. In the field of computational histopathology, both whole slide images (WSIs) and diagnostic captions provide valuable insights for making diagnostic decisions. However, aligning WSIs with diagnostic captions presents a significant challenge. This difficulty arises from two main factors: 1) Gigapixel WSIs are unsuitable for direct input into deep learning models, and the redundancy and correlation among the patches demand more attention; and 2) Authentic WSI diagnostic captions are extremely limited, making it difficult to train an effective model. To overcome these obstacles, we present PathM3, a multimodal, multi-task, multiple instance learning (MIL) framework for WSI classification and captioning. PathM3 adapts a query-based transformer to effectively align WSIs with diagnostic captions. Given that histopathology visual patterns are redundantly distributed across WSIs, we aggregate each patch feature with MIL method that considers the correlations among instances. Furthermore, our PathM3 overcomes data scarcity in WSI-level captions by leveraging limited WSI diagnostic caption data in the manner of multi-task joint learning. Extensive experiments with improved classification accuracy and caption generation demonstrate the effectiveness of our method on both WSI classification and captioning task.

Keywords: Multimodal learning · Multi-instance Learning · Multi-task learning · Histopathology image analysis

1 Introduction

Histopathology remains the gold standard for diagnosing a wide range of cancers[31, 10]. With the rise of deep learning techniques, computational histopathology has made remarkable advances[25, 22], especially in training models on gigapixel whole slide images (WSIs) from Hematoxylin and Eosin (H&E)-stained specimens [27]. Pathologists typically write diagnostic captions informed by their

⋆ Corresponding author.

analysis of WSIs. These captions contribute valuable insights to the diagnostic process. The potential of utilizing such expert knowledge has increasingly attracted interest in developing deep learning models that can process both WSIs and captions to yield more comprehensive and interpretable diagnostic outcomes.

Recent vision-language models have achieved significant success in nature images and texts [16, 1, 11, 12, 28]. Among these, the query-based transformer architecture has been particularly effective due to its capability to capture fine-grained alignment between visual and textual data[1, 11]. Inspired by these developments, the trend towards multimodal learning combining image and text modalities has expanded into the medical domain[20, 29, 13, 15, 19]. Based on these prior studies, we aitm to adapt the query-based transformer to the domain of computational histopathology, with a special focus on the fusion of WSIs and WSI-level diagnostic captions. Nevertheless, integrating WSIs with WSI-level diagnostic captions presents unique challenges. First, WSIs are not suitable for direct input into deep learning models due to their immense size. Moreover, unlike natural images, which are normally independently and identically distributed, the patches extracted from WSIs exhibit redundancy and correlation, which demand specific attention or processing techniques. Second, reliable WSI diagnostic captions require specialized pathologists and are limited by privacy concerns, leading to a scarcity of such captions vital for training effective models.

Multiple Instance Learning (MIL) serves as a popular solution to the first challenge[26]. It processes the instances derived from a "bag" as inputs and predicts the bag-level label. Each WSI is considered as a "bag" and the extracted patches are regarded as instances within this bag. Considering the correlations among instances, we employ an aggregation mechanism to combine instance features into bag-level representation, which can then be used as the input for a query-based transformer for modality fusion. Existing studies mainly address the shortage of WSI-level captions by developing datasets through the use of books, articles, and web sources to compile large-scale histopathology image-caption pairs [6, 7]. However, these efforts generally yield captions limited to the patch level rather than the WSI level. In response, we propose a multi-task joint learning framework that supports both WSI classification and captioning task. This framework aims to maximize the utility of limited captions and enhance learning efficiency and predictive accuracy by leveraging multimodal data sources.

To tackle above challenges, we propose **PathM3**, a **M**ultimodal, **M**ulti-task, **M**ultiple instance learning framework for histopathology image analysis. PathM3 facilitates image-text alignment at the WSI level and is designed to deliver enhanced performance for WSI classification and captioning, even with limited text data during training. Our contributions can be summarized as follows. (1) **WSI-level image and text modality fusion:** PathM3 adapts a query-based transformer to effectively align WSIs with their diagnostic captions, which is pivotal for achieving precise and coherent multimodal understanding in histopathology analysis. (2) **Instance correlation aggregation:** Our aggregation mechanism learns the correlation among instances within WSIs during the multimodal data fusion process, ensuring that spatial redundancy and contextual relationships

are utilized to enhance diagnostic accuracy. (3) **Efficient utilization of limited WSI captions:** Our framework can utilize limited WSI caption data in the training process, significantly improving classification precision and caption generation, which is a capability absent in most current models.

2 Related Work

Current MIL methods can be broadly categorized into two types: bag-level and instance-level. Bag-level MIL transforms instances into low-dimensional embeddings, aggregating these into bag-level representations for analytical tasks, i.e. ABMIL[8] and TransMIL[17]. Some other existing methods perform image-text alignment at the instance level, such as CITE[29] and MI-Zero[13]. Our work is distinct from these works as it seeks alignment at the bag level. We aim to fully utilize the contextual information within WSIs, thereby enabling a more comprehensive interpretation consistent with the diagnostic captions typically drawn by pathologists at the WSI level. Vision-language models are emerging for natural images and texts, i.e., Clip[16] and Flamingo[1]. A notable method, Blip2[11], builds on this foundational work by proposing a query transformer that narrows the gap between visual and textual data and achieves promising results. This multimodal learning paradigm has achieved certain success in applications within the histopathology field, such as CITE[29], MI-Zero[13], FSWC[15]. However, these approaches either focus on captions at the patch level[13], or the text they use is simple category names[29, 13] or generated by Large Language Models(LLM)[15], without considering the integration of WSIs with WSI-level diagnostic captions. Our approach takes a significant step forward by facilitating the fusion of WSIs with diagnostic captions at the WSI level, thereby enriching medical image analysis with a more complete and contextual approach.

3 Method

The overall framework of PathM3 is illustrated in Figure 1, with each component introduced in detail subsequently.

Problem Formulation Consider a dataset comprising a set of $N \times$ WSIs denoted by $\boldsymbol{S} = \{S_1, S_2, \ldots, S_N\}$. Corresponding to each WSI S_i, there exists a WSI-level caption $\boldsymbol{T} = \{T_1, T_2, \ldots, T_N\}$ and a categorical label y_i drawn from $\boldsymbol{Y} = \{1, \ldots, C\}$, where C represents the total number of distinct categories. For the classification task, the sets $\boldsymbol{S}$ and $\boldsymbol{T}$ serve as inputs to our model, which then aims to accurately predict the class labels $\hat{\boldsymbol{Y}}$. The captioning task entails the generation of predicted captions $\hat{\boldsymbol{T}}$ based solely on $\boldsymbol{S}$ as the input.

Correlation of each instance Each WSI S_i consists of M patches, which can be denoted by $\boldsymbol{P} = \{P_1, P_2, \ldots, P_M\}$. To simplify the following learning task, we utilize a frozen image encoder to extract features for each patch. This means that

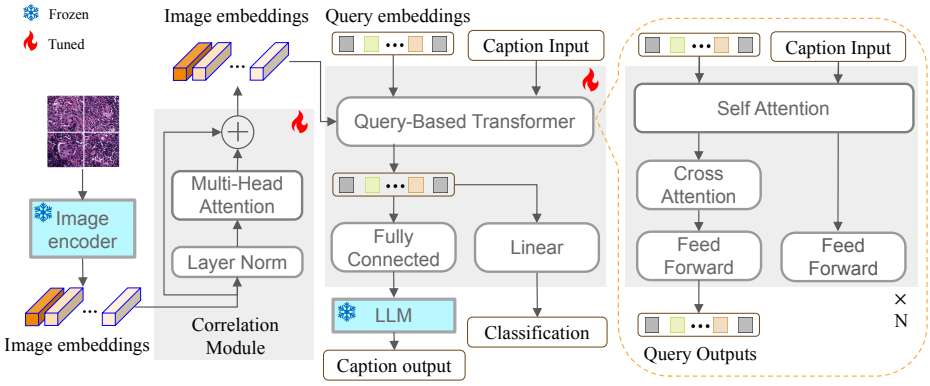

Fig. 1. PathM3 overview. A WSI is fed into a frozen image encoder to generate image embeddings. These embeddings then pass through a correlation module before being fed into a query-based transformer, in which the learnable query embeddings interact with the textual embeddings using self-attention and with image embeddings using cross-attention. The outputs of these queries are then utilized for classification via a linear classifier and for generating captions with a frozen LLM.

each WSI can be represented as a set of embeddings $\boldsymbol{E} = \{E_1, E_2, \ldots, E_M\} \in \mathbb{R}^{M \times d}$ where M denotes the number of patches and d represents the dimensions of each embedding. TransMIL[17] has demonstrated that applying self-attention mechanisms can effectively learn the correlation among multiple input instances. Inspired by this, we incorporate a self-attention mechanism into our correlation module to perform information integration. The computation can be formulated as follows,

$$\boldsymbol{E} = \mathrm{MSA}\left(\mathrm{LN}\left(\boldsymbol{E}\right)\right) + \boldsymbol{E}, \tag{1}$$

where MSA stands for Multi-head Self-Attention and LN denotes Layer Norm. However, in WSIs, each bag may contain a large number of instances ($M > 1000$). Directly computing self-attention among instances results in a time complexity of $\mathcal{O}(M^2)$, which is computationally intensive. Therefore, we employ the Nystrom approximation of attention[23], which can be defined as:

$$\mathrm{softmax}\left(\frac{\tilde{\boldsymbol{Q}}\boldsymbol{K}^T}{\sqrt{d_q}}\right)\left(\mathrm{softmax}\left(\frac{\tilde{\boldsymbol{Q}}\tilde{\boldsymbol{K}}^T}{\sqrt{d_q}}\right)\right)^{+}+\mathrm{softmax}\left(\frac{\tilde{\boldsymbol{Q}}\boldsymbol{K}^T}{\sqrt{d_q}}\right), \tag{2}$$

where $\tilde{\boldsymbol{Q}}$ and $\tilde{\boldsymbol{K}}$ are the m selected landmarks from the original n dimensional sequence of $\boldsymbol{Q}$ and $\boldsymbol{K}$, and $\boldsymbol{A}^{+}$ is a Moore-Penrose pseudoinverse of $\boldsymbol{A}$.

WSI and Caption Fusion We utilize a query-based transformer to establish a connection between WSIs and their corresponding captions. Similar to Flamingo[1] and Blip2[11], this module takes K learnable query embeddings as inputs. Initially, these queries interact with each other through self-attention

layers. When textual data is available, the query further engages with textual information within the same self-attention layers. Following self-attention, the queries proceed to interact with image features via cross-attention. The output of this module is a set of K integrated visual-text vectors, with each vector corresponding to a specific query embedding. After passing through N blocks, these output query vectors are then fed into various specialized modules, tailored for specific tasks. Since our query-based transformer adds cross-attention layers to $\text{BERT}_{\text{base}}$[4], similar to Blip2[11], we initialize it with the pre-trained weights from the first stage of Blip2[11].

Multi-task Joint Learning TieNet[20] leverages multi-task joint learning on X-ray images. We extend this approach to the domain of WSIs. Our PathM3 employs a multi-task joint learning framework by using WSIs as the sole input for inference, while taking advantage of both WSIs and captions during the training phase. This strategy allows us to harness the synergistic effects of multi-modal data, resulting in improved accuracy for both classification and captioning task. Specifically, for the classification task, we introduce both WSIs and captions into our fusion module, which outputs K query embeddings that encapsulate combined visual-textual information. Subsequently, each of these output query embeddings is individually passed through a linear classifier to generate K corresponding logits. Then, the classification prediction p_i is obtained by averaging these K logits across all query embeddings. To optimize the classification performance, we employ the cross-entropy function L_C, which allows us to compute the classification loss:

$$L_C = -\sum_{i=1}^{C} y_i \log(p_i), \tag{3}$$

where p_i denotes the predicted classification outcome and the ground truth label is given by y_i. C represents the total number of distinct categories in the dataset.

In the captioning task, our approach involves inputting WSIs solely into the query-based transformer, where the output query embeddings interact with visual information through cross-attention. It is important to note that text data is not introduced into the query transformer during this process. These query embeddings, now enriched with visual details, are then fed into a frozen LLM (e.g., FlanT5[3]) to leverage its well-established generative language capabilities. Captions in this scenario are employed exclusively as the generational targets for the LLM. Our optimization efforts are geared towards the minimization of the generative loss, denoted as L_G, which is calculated based on the output of the LLM. Since our framework is a multi-task framework, by combining L_C and L_G, the final objective function L_{overall} for the proposed method can be formulated as:

$$L_{\text{overall}} = \alpha L_C + (1 - \alpha) L_G, \tag{4}$$

where α is added to balance the large difference between the two loss types.

4 Experiments and Results

4.1 Dataset

In our study, we employ the PatchGastric dataset [18], which consists of 991 WSIs, sourced from the surgical files of individual patients. This dataset is selected as it is among the few publicly accessible datasets providing high-resolution histology images coupled with WSI-level captions, a requisite for our research. The PatchGastric dataset contains 9 gastric adenocarcinoma subtypes. Drawing on insights from CITE[25], our study on the PatchGastric dataset, which includes 9 gastric adenocarcinoma subtypes, focuses on three main ones: "well differentiated tubular adenocarcinoma", "moderately differentiated tubular adenocarcinoma", and "poorly differentiated adenocarcinoma". The dataset is randomly split into three parts: training (20%), validation (40%), and testing (40%), to mimic the scenario of having limited WSI captions in the real world.

4.2 Comparison with state-of-the-art methods

We compare our PathM3 with other baseline methods on WSI classification tasks under the settings of image only inference and image & text inference. All experiments are run three times with different random seeds. As shown in Table 1, PathM3 achieves an average accuracy of 86.40%, outperforming all other methods by at least 4.08% in accuracy. These results empirically demonstrate the effectiveness of utilizing a query-based transformer to leverage the multimodal relationships within the WSIs and their captions. Considering that captions may not always be available in real-world scenarios, we also evaluate PathM3 under the condition of using only images for inference. Results in Table 1 show that PathM3 obtains an average accuracy of 71.48%, surpassing the comparative methods by up to 4.81%.

The performance of our PathM3 in captioning task is summarized in Table 2, where it is compared against a set of baselines reported in [18]. Our method surpasses all baselines by a significant margin across all three metrics for image captioning, namely BLEU@4, METEOR, and SPICE. With PathM3, a BLEU@4 score of 0.520 is achieved, which reflects a marked improvement in the precision of generated textual descriptions over the best baseline, which scores 0.342. In terms of assessing the alignment with reference captions, the METEOR score of our PathM3 reaches 0.394, exceeding the previously highest score of 0.316. Lastly, for the SPICE metric, which evaluates the semantic propositional content of the generated captions, sees a jump to 0.591 with PathM3, distinctly outperforming the best baseline score of 0.393. These empirical results convincingly demonstrate the efficacy of PathM3 in generating coherent and contextually accurate captions, which have the potential to aid pathologists and enhance the interpretability of deep learning models.

Table 1. Classification results comparison on PatchGastric dataset [18]. The accuracy is reported as mean ± standard deviation. Best results are marked in **bold**.

Data modality (Inference)	Method	Accuracy (%)
image only	ABMIL[8]	66.67 ± 2.63
	CLAM[14]	67.83 ± 1.12
	DSMIL[9]	69.02 ± 0.42
	TransMIL[17]	67.48 ± 1.16
	CITE[29]	69.63 ± 0.91
	ILRA-MIL[21]	70.16 ± 1.11
	PathM3 (Ours)	$\mathbf{71.48 \pm 1.30}$
image & text	MCAT[2]	80.88 ± 0.58
	MOTCat[24]	82.28 ± 0.54
	CMTA[30]	81.23 ± 0.62
	PathOmics[5]	81.32 ± 0.79
	PathM3 (Ours)	$\mathbf{86.40 \pm 0.21}$

Table 2. Comparative performance in captioning of PathM3 against baseline methods[18]. BLEU@4, METEOR, and SPICE scores are presented as mean ± standard deviation. Best results are marked in **bold**.

Method	BLEU@4	METOR	SPICE
DenseNet121 x20 p3x3[18]	0.310 ± 0.020	0.307 ± 0.009	0.382 ± 0.005
EfficientNetB3 x20 p3x3[18]	0.315 ± 0.028	0.293 ± 0.016	0.359 ± 0.017
DenseNet121 x20 pavg[18]	0.324 ± 0.011	0.310 ± 0.001	0.377 ± 0.008
EfficientNetB3 x20 pavg[18]	0.342 ± 0.021	0.316 ± 0.012	0.393 ± 0.008
PathM3 (Ours)	$\mathbf{0.520 \pm 0.011}$	$\mathbf{0.394 \pm 0.006}$	$\mathbf{0.591 \pm 0.011}$

4.3 Ablational studies

Table 3 presents the results of an ablation study. This study is designed to assess the influence of various data modalities for inference and multi-task joint learning. The study also compares performance with and without the correlation module. Analyzing our PathM3 model without correlation, we observe that the

Table 3. Ablation study on the classification accuracy with and without correlation under single-task and multi-task settings with various data modalities.

Data modality (Inference)	Correlation	Multi-task	Accuracy (%)
image only			68.59 ± 0.72
		✓	69.07 ± 1.46
	✓		70.88 ± 1.85
	✓	✓	$\mathbf{71.48 \pm 1.30}$
text only			79.79 ± 1.25
image & text			84.60 ± 0.91
	✓		$\mathbf{86.40 \pm 0.21}$

model achieves an accuracy of 68.59% when only images are used in a single-task setting. Incorporating multiple tasks into the training pushes the accuracy up to 69.07%. We find that using only the textual data modality yields 79.79%, which highlights the value of the textual information provided by pathologists' captions for classification. Notably, the combination of image and text data results in an accuracy of 84.60%, underlining the joint effect of multimodal learning. Upon introducing the correlation mechanism within PathM3, a marked increase is observed across all modalities. The single-task image classification accuracy improves to 70.88%, and further enhancement is noticeable with multi-task learning, yielding an accuracy of 71.48%. Importantly, the fusion of image and text in an integrated, correlated framework achieves an accuracy of 86.40%. This outcome confirms the pivotal role of the correlation mechanism in effectively utilizing and integrating multimodal data to improve classification performance.

Table 4. Ablation study on the effects of correlation and multi-task joint learning on the captioning task. Best results are marked in **bold**.

Correlation	Multi-task	BLEU@4	METEOR	SPICE
		0.503 ± 0.006	0.386 ± 0.002	0.579 ± 0.005
	✓	0.508 ± 0.012	0.388 ± 0.005	0.582 ± 0.011
✓		0.508 ± 0.169	0.388 ± 0.006	$\mathbf{0.591 \pm 0.006}$
✓	✓	$\mathbf{0.520 \pm 0.011}$	$\mathbf{0.394 \pm 0.006}$	$\mathbf{0.591 \pm 0.011}$

Table 4 provides insights from an ablation study that evaluates the impact of two key components on the captioning performance: the correlation module and the multi-task joint learning setup. By empirically adding or omitting these elements, we can observe their individual and combined contributions to the performance metrics. From the results, it is evident that both components offer significant value to the captioning task. The inclusion of the correlation module alone led to a notable performance boost, with BLEU@4 scores rising from 0.508 to 0.520, and an increase in the SPICE metric to 0.591. The use of a multi-task learning framework also shows a positive effect, particularly when used in conjunction with the correlation module, yielding the highest recorded metrics across all categories: BLEU@4 (0.520), METEOR (0.394), and SPICE (0.591). In conclusion, this ablation study confirms the effectiveness of the correlation module and multi-task learning in enhancing captioning performance.

5 Conclusion

We propose PathM3, a multi-modal, multi-task, multi-instance learning framework for histopathology image analysis. The proposed approach addresses key challenges in histopathology image analysis, including the intricate alignment of WSIs with their respective diagnostic captions, and the utilization of limited textual data to enhance model performance. The novel technique delivers a

robust feature extraction and aggregation process by utilizing the query-based transformer and taking into account the correlation among patches within a WSI. Our framework outlines a significant stride in computational histopathology, advancing the integration of deep learning with expert narratives to foster data-efficient, interpretable, and informative diagnostic outcomes.

Acknowledgements. This work was partially supported by US National Science Foundation IIS-2412195, CCF-2400785 and the Cancer Prevention and Research Institute of Texas (CPRIT) award (RP230363).

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

References

1. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al.: Flamingo: a visual language model for few-shot learning. Advances in Neural Information Processing Systems **35**, 23716–23736 (2022)
2. 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)
3. Chung, H.W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, Y., Wang, X., Dehghani, M., Brahma, S., et al.: Scaling instruction-finetuned language models. Journal of Machine Learning Research **25**(70), 1–53 (2024)
4. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 4171–4186 (2019)
5. Ding, K., Zhou, M., Metaxas, D.N., Zhang, S.: Pathology-and-genomics multimodal transformer for survival outcome prediction. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 622–631. Springer (2023)
6. Gamper, J., Rajpoot, N.: Multiple instance captioning: Learning representations from histopathology textbooks and articles. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16549–16559 (2021)
7. Huang, Z., Bianchi, F., Yuksekgonul, M., Montine, T.J., Zou, J.: A visual–language foundation model for pathology image analysis using medical twitter. Nature medicine **29**(9), 2307–2316 (2023)
8. Ilse, M., Tomczak, J., Welling, M.: Attention-based deep multiple instance learning. In: International conference on machine learning. pp. 2127–2136. PMLR (2018)
9. Li, B., Li, Y., Eliceiri, K.W.: 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. pp. 14318–14328 (2021)

10. Li, C., Zhu, X., Yao, J., Huang, J.: Hierarchical transformer for survival prediction using multimodality whole slide images and genomics. In: 2022 26th international conference on pattern recognition (ICPR). pp. 4256–4262. IEEE (2022)
11. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: International conference on machine learning. pp. 19730–19742. PMLR (2023)
12. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems **36** (2024)
13. Lu, M.Y., Chen, B., Zhang, A., Williamson, D.F., Chen, R.J., Ding, T., Le, L.P., Chuang, Y.S., Mahmood, F.: Visual language pretrained multiple instance zero-shot transfer for histopathology images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19764–19775 (2023)
14. 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)
15. Qu, L., Fu, K., Wang, M., Song, Z., et al.: The rise of ai language pathologists: Exploring two-level prompt learning for few-shot weakly-supervised whole slide image classification. Advances in Neural Information Processing Systems **36** (2024)
16. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
17. 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)
18. Tsuneki, M., Kanavati, F.: Inference of captions from histopathological patches. In: International Conference on Medical Imaging with Deep Learning. pp. 1235–1250. PMLR (2022)
19. Wang, P., Wells, W.M., Berkowitz, S., Horng, S., Golland, P.: Using multiple instance learning to build multimodal representations. In: International Conference on Information Processing in Medical Imaging. pp. 457–470. Springer (2023)
20. Wang, X., Peng, Y., Lu, L., Lu, Z., Summers, R.M.: Tienet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 9049–9058 (2018)
21. Xiang, J., Zhang, J.: Exploring low-rank property in multiple instance learning for whole slide image classification. In: The Eleventh International Conference on Learning Representations (2022)
22. Xiao, L., Xu, R., Cang, Y., Chen, Y., Wei, Y.: Advancing surgical imaging with cgan for effective defogging. International Journal of Innovative Research in Computer Science & Technology **12**(3), 135–139 (2024)
23. 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)
24. 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)
25. Yan, Y., He, S., Yu, Z., Yuan, J., Liu, Z., Chen, Y.: Investigation of customized medical decision algorithms utilizing graph neural networks. arXiv preprint arXiv:2405.17460 (2024)

26. Yao, J., Zhu, X., Jonnagaddala, J., Hawkins, N., Huang, J.: Whole slide images based cancer survival prediction using attention guided deep multiple instance learning networks. Medical Image Analysis **65**, 101789 (2020)
27. Yao, J., Zhu, X., Zhu, F., Huang, J.: Deep correlational learning for survival prediction from multi-modality data. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 406–414. Springer (2017)
28. Zhang, Y., Gao, J., Tan, Z., Zhou, L., Ding, K., Zhou, M., Zhang, S., Wang, D.: Data-centric foundation models in computational healthcare: A survey. arXiv preprint arXiv:2401.02458 (2024)
29. Zhang, Y., Gao, J., Zhou, M., Wang, X., Qiao, Y., Zhang, S., Wang, D.: Text-guided foundation model adaptation for pathological image classification. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 272–282. Springer (2023)
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)
31. Zhu, X., Yao, J., Zhu, F., Huang, J.: Wsisa: Making survival prediction from whole slide histopathological images. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7234–7242 (2017)