Exploring Multi-modal Prompt Learning for Few-shot Whole Slide Image Classification with Multiple Instance Learning

Ying Chen[1], **Xiaoqi Chen**[2], **Riyu Qiu**[3], **Yixuan Lin**[4]

[1]23020231154135, [2]23020231154171, [3]23020231154220, [4]23020231154206
All the students are from Information Institute class.

Abstract

The digital transformation of pathological images into Whole Slide Images (WSIs) is pivotal in cancer diagnostics, representing the gold standard for diagnosis. However, classifying WSIs poses substantial challenges, primarily due to their gigapixel resolution and the limited availability of annotated data. To overcome these obstacles, we propose a novel Multimodal Prompt Learning Multiple Instance Learning (MP-MIL) framework, specifically designed for efficient few-shot WSI classification. This innovative approach employs GPT-4 to generate descriptive prompts, thereby enhancing the textual context of WSIs during training. The framework refines a Vision-Language model using solely image and text prompts, leading to significant reductions in computational training costs while maintaining high performance. Additionally, MP-MIL introduces an instance prompt-guided pooling mechanism that effectively captures richer semantic features, aiding in the aggregation of image features. The effectiveness of MP-MIL has been rigorously validated on renowned datasets, including CAMELYON16 and TCGA-LUNG. These evaluations demonstrate the framework's ability to markedly decrease training expenses and enhance few-shot classification accuracy in pathological analysis.

Introduction

Histological Whole-Slide Images (WSIs) are the cornerstone of pathological diagnosis, providing critical insights into oncological assessments and informing treatment strategies. This diagnostic prowess is documented in seminal works such as (Li et al. 2021; Lu et al. 2021), which underscore the preeminence of WSIs in medical diagnostics. Within the analytical framework of WSIs, Multiple Instance Learning (MIL) has gained traction as a compelling methodology. MIL conceptualizes a WSI as a collection, or 'bag', of numerous smaller segments, referred to as 'patches' or 'instances'. The diagnostic paradigm within MIL posits that a bag is classified as negative if all constituent patches are negative, whereas the detection of even a single positive patch suffices to classify the entire bag as positive, a principle detailed in (Qu et al. 2022). Traditional approaches to WSI analysis rely on feature extraction from instances using pre-trained models on natural images, subsequently applying MIL for classification. These methodologies, however, inherently assume the availability of extensive labeled data at the bag level for training. The scarcity of pathological data, however, poses a critical challenge for these traditional MIL approaches. This scarcity stems from a constellation of factors: stringent patient privacy laws, logistical hurdles in procuring pathological specimens, and the rarity of certain pathologies. Such constraints necessitate the exploration of innovative strategies for domain-specific feature extraction and few-shot classification techniques in WSI analysis.

The advent of vision-language models (V-L models) has introduced new avenues for computational diagnostics. Seminal studies (Li et al. 2021; Lu et al. 2021) have illuminated the potential of V-L models, particularly in the context of few-shot learning applications, suggesting their utility in WSI classification tasks. Despite the promise these models hold, the application of V-L models pretrained on general domains to the specialized field of pathology is not straightforward. The divergent characteristics of pathological images and text descriptions from those found in general datasets necessitate the adaptation of these models to the unique domain of pathology. Moreover, the sheer volume of instances in WSIs and the corresponding lack of detailed textual annotations exacerbate the challenge, creating significant barriers to the effective deployment of V-L models in pathological analysis.

To navigate these challenges, research must pivot towards the development of V-L models that are not only robust in few-shot learning scenarios but are also tailored to accommodate the idiosyncrasies of pathological data. Such models must be capable of discerning subtle histological nuances from limited annotations and extrapolating these findings to inform accurate diagnoses. Bridging this gap between general V-L models and domain-specific requirements will be instrumental in realizing the full potential of AI in pathology, potentially revolutionizing the field with enhanced diagnostic precision and efficacy.

To address these challenges, we introduce a novel Multimodal Prompt Learning Multiple Instance Learning (MP-MIL) framework to achieve precise bag-level classification with very few training bags. Leveraging the advanced capabilities of GPT-4, we enrich the sparse textual information of WSI labels with detailed descriptive prompts. We further fine-tune the V-L model only with trainable image and text prompts, obviating retraining the large backbone network.

Unlike traditional MIL, which focuses only on the aggregation of image features, our framework employs an instance prompt-guided pooling mechanism for instance aggregation. This allows for a full interaction of textual and image information, obtaining a more comprehensive and fine-grained representation. Overall, our contributions are summarized as follows:

- We propose a novel Multi-modal Prompt Learning Multiple Instance Learning (MP-MIL) framework to achieve few-shot WSI classification, which address the scarcity of labeled data in the pathological domain.
- We employ GPT-4 to create descriptive prompts that supplement the sparse textual data associated with WSI labels, thereby providing a richer semantic context and enhancing understanding of model.
- The MP-MIL framework innovatively employs multimodal prompts, incorporating both image and text elements, to fine-tune a Vision-Language model specifically for WSI analysis. This approach not only achieves substantial savings in training costs but also markedly enhances the performance of few-shot classification tasks.
- We present a novel instance prompt-guided pooling mechanism that aggregates the features of individual instances within a WSI bag, which allows for more abundant semantic feature representation by incorporating text prompts as a guide for the feature aggregation.

Related Work

Multiple Instance Learning in WSI classification

Existing WSI analysis approaches generally adopt Multiple Instance Learning (MIL) to conduct WSI classification (Campanella et al. 2019; Lerousseau et al. 2020; Xu et al. 2019; Li, Li, and Eliceiri 2021; Zhang et al. 2022; Wang et al. 2018). The high-resolution WSI are partitioned into image patches before further processing to fit the data in modern computation hardware. In this case, MIL fits the classification task of WSI by corresponding the slides to bags, and patches to instances. Many MIL methods have assumption that the instances are independent and identically distributed (i.i.d.), which is not suitable for many application scenarios (e.g., a patient gastroscopy picture set). To deal with this, TransMIL (Shao et al. 2021) proposed a TPT architecture, using two transformer layers and a pyramid encoder to implement correlated-instance learning and mine information between instances. To improve the attention uncertainty, Bayes-MIL (Yufei et al. 2022) is proposed to implement a reglarlization type skill over attention weights. IBMIL (Lin et al. 2023) designs an interventional training to deal with “bag contextual prior”: there exits shared things within bags , which is not directly related to their assigned labels, can still influence the final predictions. MIL could also help vision transformer to utilize patch-level features rather than only cls-token by a MIL head (Yu et al. 2021).

Vision-language Models and Prompt Learning

Pre-trained vision-language (V-L) models such as CLIP (Radford et al. 2021), ALIGN (Jia et al. 2021), and FLIP (Yao et al. 2021) have shown immense promise in the areas of visual representation and transfer learning, having been trained on large volumes of image-text pairs. These V-L models incorporate a dual-tower architecture consisting of visual and text encoders and leverage contrastive learning to synchronize text-to-image and image-to-text correspondences within the feature space. It’s noteworthy that pre-trained V-L models, for example, CLIP (Radford et al. 2021), exhibit impressive transferability in image recognition tasks. By strategically formulating text descriptors, also known as prompts, to match the associated image features within the feature space, these models facilitate classification tasks in a zero-shot or few-shot manner. Building upon the success of CLIP, CoOp (Zhou et al. 2022) substitutes these handcrafted prompts with a learned prompt representation, thus enhancing the applicability of V-L models to downstream few-shot classification (FSC) tasks. Encouraged by the success of V-L models in FSC within the realm of natural imagery, we put forth several techniques to efficiently tailor pre-trained V-L models to tackle the few-shot within-class (FSWC) problem.

Proposed Solution

Problem Formulation

Given a dataset $X = \{X_1, X_2, \dots, X_N\}$ comprising N WSIs, and each WSI X_i is partitioned into non-overlapping small patches $\{x_{i,j}; j = 1, 2, \dots, n_i\}$, where n_i represents the number of patches obtained from X_i. All patches within X_i collectively form a bag, and each patch serves as an instance of that bag. The bag is assigned a label $Y_i \in \{0, 1\}$, where $i = \{1, 2, \dots, N\}$. The labels of each instance $\{y_{i,j}; j = 1, 2, \dots, n_i\}$ are associated with the bag label in the following manner:

$$Y_i = \begin{cases} 0, & \text{if } \sum_j y_{i,j} = 0 \\ 1, & \text{else} \end{cases} \tag{1}$$

This implies that all instances within negative bags are assigned negative labels, whereas positive bags contain at least one positive-labeled instance. In the context of weakly-supervised MIL, only the bag label is provided for the training set, while the labels of individual instances remain unknown. The Few-shot Weakly-supervised WSI Classification (FSWC) task poses an even greater challenge as it allows for only a limited number of labeled bags for training. Typically, only a small number of bags per class, such as 1, 2, 4, 8, or 16, are available. The objective of FSWC is to accurately classify both the bags and individual instances, despite the scarcity of labeled training bags.

Overview

To effectively address the few-shot WSI classification challenge, we introduce a Multi-modal Prompt Learning MIL framework, denoted as MP-MIL, as shown in Figure 1. Firstly, we employ a frozen Image Encoder of CLIP, to process each instance x_i within a bag B_i. This encoder divides each instance into smaller patches, subsequently embedding them into tokens. Alongside this, a Trainable Image Prompt is concatenated and integrated into the image

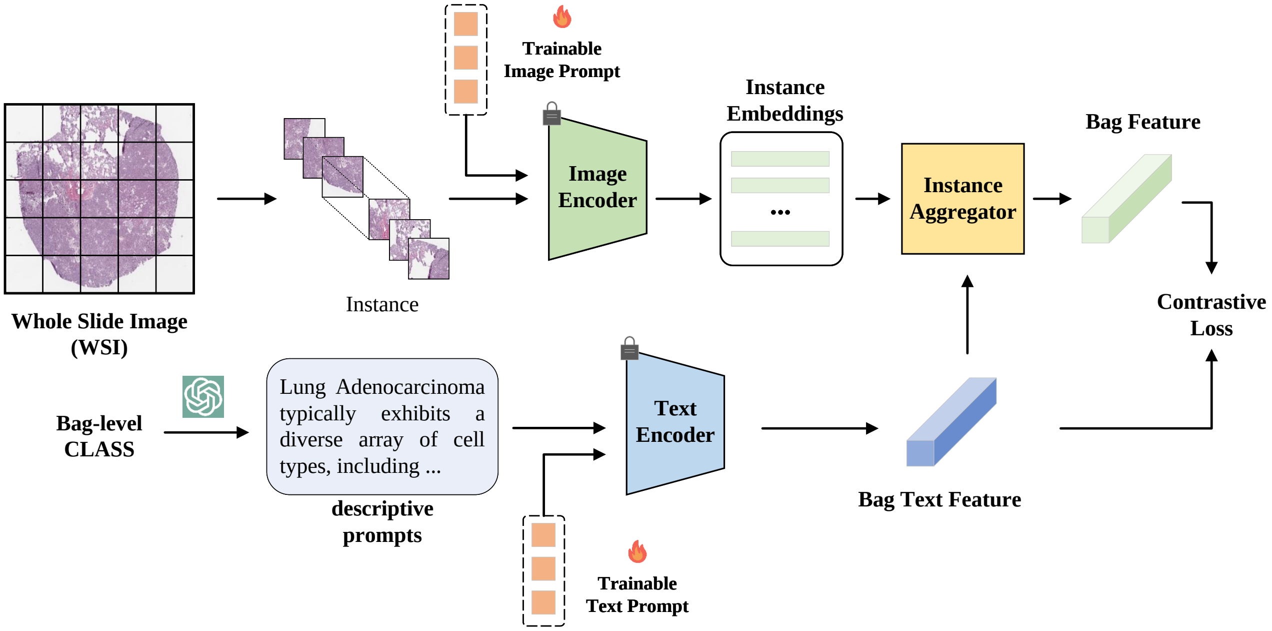

Figure 1: Overview of MP-MIL.

encoder. Following this, we aggregate these individual features into a unified bag feature F_i by instance prompt-guided pooling mechanism. Subsequently, we utilize GPT-4 to intricately describe the bag class label as a Descriptive Prompt, and incorporate an additional Trainable Text Prompt to autonomously capture deeper semantic nuances. Both the Descriptive Prompt and the Trainable Text Prompt are fed into the Text Encoder of CLIP to produce a Bag text token T_i. The primary training goal revolves around leveraging a limited set of bag-level labeled data to finetune the Trainable Image Prompt vectors V_I and the Trainable Text Prompt vectors V_T. In culmination, we use this sparse labeled data to optimize V_I and V_T through a cross-entropy loss function, aiming to maximize similarity between the whole slide image and its descriptive label.

For the inference phase, we evaluate the matching degree between the image features and all corresponding target class bag prompt features to ascertain the classification category in whole slide image bag classification.

Construction of Prompts

We utilized the prompt “Describe the morphological characteristics of the LABEL in a single sentence in English.” to obtain label descriptions through GPT-4. When utilized, the placeholder tag LABEL is substituted with each specific label in the process.

LUAD: LUAD (Lung Adenocarcinoma) typically exhibits a diverse array of cell types, including glandular, papillary, and acinar structures with mucin production, and varying differentiation levels from well-differentiated to poorly differentiated.

LUSC: LUSC (Lung Squamous Cell Carcinoma) is characterized by tumor cells forming sheet-like squamous structures, possibly showing keratinization features like keratin pearl formation, and is typically well-differentiated.

Metastasis: Metastasis slides typically exhibit abnormal cell morphology, such as irregular cell shapes, sizes, and staining characteristics, along with disrupted tissue structures and prominent nuclear changes.

Non-Metastasis: Non-metastasis slides are characterized by normal cell morphology with regular shapes, sizes, and staining, intact tissue structures, and regular nuclear features, indicating the absence of cancerous cell spread.

Instance Aggregator Module

The Instance Aggregator (IA) module is used to aggregate the fine-grained diagnosis prompts and instance features. IA consists of a self-attention module and a cross-attention module.

We employ self-attention to enable feature interaction among instance features $I_i = [e_{i1}, e_{i2}, \cdots, e_{ij}]$, resulting in the feature s_i. Subsequently, utilizing the Bag Text Feature Q to aggregate the instance features and acquire the feature z_i. Then we concatenate s_i and z_i, utilizing the learnable parameter W to fuse these features, ultimately yielding the bag-level feature v_i. The formulas are shown as follows:

$$s_i = SelfAttention(I_i, I_i) + I_i \tag{2}$$

$$z_i = CrossAttention(Q, s_i) \tag{3}$$

$$v_i = concat(mean(s_i), mean(z_i)) \cdot W \tag{4}$$

Ultimately, we acquire the image bag-level features guided by the text prompts, which are then employed to align the Bag Text Features.

Encoder and Loss Function

We divide each WSI into instances x_k and encode these instances into embeddings $e_k \in R^D$ using pre-trained vision encoder E_{img}, composed with ResNet-50 or Transformer structure following (Li, Li, and Eliceiri 2021). Then we send the instance embeddings into the TFS Module to aggregate the instance features and prompts, and obtain the bag-level embeddings $v_i \in R^D$. The formulas are shown as follows:

$$e_k = E_{img}([x_k, V_I]) \tag{5}$$

$$I_i = [e_{i1}, e_{i2}, \cdots, e_{ik}] \tag{6}$$

$$v_i = IA(I_i, E_{txt}(Q)) \tag{7}$$

Besides, we generate pathologically meaningful text embeddings, represented as $t_i \in R^D$, by leveraging the fine-tuned text encoder BioClinicalBERT (Wolf et al. 2019).

$$t_i^c = E_{txt}([x_{txt}, V_T]) \tag{8}$$

where E_{txt} denotes the text encoder, and $x_{txt}^c (c \in [1, C])$ where C denotes the number of categories. Here we use the same embedding dimension D as the vision encoder, suitable for contrastive learning.

Subsequently, the bag-level embeddings v_i are aligned with the text embeddings t_i^c to complete the training process. In this case, prediction $\hat{y}$ is obtained by applying softmax on scaled cosine similarities between the image embeddings and text embeddings:

$$p(\hat{y} = c|I) = \frac{exp(sim(t_i^c, v_i)/\tau)}{\sum_{c'=1}^{C} exp(sim(t_i^{c'}, v_i)/\tau)} \tag{9}$$

where $sim(\cdot, \cdot)$ refers to cosine similarity and τ is the temperature parameter.

The training loss is computed as the cross-entropy between the logits and soft targets as:

$$L^{v \to t} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{i=j}^{N} y_{ij} log(p_{ij}) \tag{10}$$

here N corresponds to the batch size.

Likewise, we can compute $L^{t \to v}$ and serve L as the final training objective.

$$L = \frac{L^{v \to t} + L^{t \to v}}{2} \tag{11}$$

Experiments

Dataset

We evaluated our method on public histopathology WSI datasets: The Cancer Genome Atlas Lung (TCGA Lung) Cancer[1] and Camelyon16 (Bejnordi et al. 2017).

TCGA Lung Cancer. The TCGA Lung Cancer dataset comprises two cancer subtypes: Lung Adenocarcinoma (LUAD) and Lung Squamous Cell Carcinoma (LUSC). It includes diagnostic slides with 541 slides from 478 LUAD cases and 512 slides from 478 LUSC cases. For WSI preprocessing, following the method described by (Li, Li, and Eliceiri 2021), we cropped each WSI into non-overlapping patches of 256 × 256 pixels and removed the background regions. The dataset encompasses approximately 5.2 million patches at 20× magnification, averaging about 5,000 patches per WSI.

Camelyon16. The Camelyon16 dataset (Bejnordi et al. 2017) consists of 399 Hematoxylin and Eosin (H&E) stained slide images, utilized for metastasis detection in breast cancer. We preprocessed each WSI by segmenting it into 256 × 256 non-overlapping patches, excluding background regions. In total, this process yields approximately 2.8 million patches at a 20× magnification level, averaging about 7,200 patches in a Bag.

Results on the TCGA Lung Cancer Dataset

Our model demonstrates adaptability to various tasks even in scenarios with limited data availability. Few-shot experiments were conducted to demonstrate its transferability to downstream tasks. We initialized the networks with pre-trained weights derived from a model trained on TCGA image-report pairs, and subsequently fine-tuned the model on downstream datasets for few-shot image classification. We followed (Qu et al. 2023) and conducted experiments with 1, 2, 4, 8, 16. The results are summarized in Table 1. Mean-pool, Max-pool, and Attn-pool correspond to Linear-Probe implementations with Mean-pooling, Max-pooling, and Attention-pooling, respectively.

Method	16-shot	8-shot	4-shot	2-shot	1-shot
Mean-pool	65.33	53.89	44.85	52.93	45.34
Max-pool	48.48	49.55	44.22	48.39	49.03
Attn-pool	72.50	65.79	62.47	58.36	56.23
CoOp	78.35	67.99	67.60	67.54	67.81
TOP	82.06	80.51	75.41	72.38	71.01
Ours	**84.25**	**82.80**	**80.10**	**75.51**	**73.91**

Table 1: Few-shot classification performance on TCGA Lung Cancer.

The Mean-pool approach yielded moderate accuracy, peaking at 65.33% for the 16-shot case and diminishing to 45.34% in the 1-shot scenario. The Max-pool method demonstrated lower performance, with a maximum accuracy of 49.55% in the 8-shot setting. The Attention-pool (Attn-pool) method showed improved results, particularly in the 4-shot configuration, achieving a 62.47% accuracy rate. The Co-Op strategy further enhanced the outcomes, attaining a peak of 78.35% in the 16-shot case. The TOP method exhibited robust performance with a high of 82.06% accuracy for the 16-shot condition. Our proposed method outperformed all others, achieving the highest accuracy across all few-shot settings, with an impressive 84.25% in the 16-shot framework and maintaining a considerable 73.91% even in the challenging 1-shot scenario. These findings suggest that our method provides a substantial improvement in few-shot learning for WSI classification, potentially setting a new benchmark for future research in AI pathology.

[1] http://www.cancer.gov/tcga

Method	16-shot	8-shot	4-shot	2-shot	1-shot
Mean-pool	67.56	65.11	66.11	66.56	65.11
Max-pool	38.33	62.00	60.11	58.78	58.78
Attn-pool	80.00	74.00	75.33	69.00	52.22
CoOp	69.94	68.00	67.56	69.54	.65.44
TOP	82.33	80.24	78.89	76.22	70.44
Ours	**85.15**	**83.11**	**81.10**	**77.11**	**73.60**

Table 2: Few-shot classification performance on Camelyon16.

TIP	TTP	DP	IA	16-shot	8-shot	4-shot	2-shot	1-shot
✓				64.31	62.23	59.77	52.85	50.58
✓	✓			75.21	73.19	71.21	68.36	61.01
✓	✓	✓		81.22	80.01	78.56	76.01	71.17
✓	✓	✓	✓	**84.25**	**82.80**	**80.10**	**75.51**	**73.91**

Table 3: The ablation experiment on the TCGA Lung Cancer. TIP, TTP, DP and IA correspond to Trainable Image Prompt, Trainable Text Prompt, Descriptive Prompt, and Instance Aggregator, respectively.

Results on the Camelyon16 Dataset

The bag classification and instance classification performance on the Camelyon 16 dataset are shown in Table 2. It can be seen that MP-MIL achieved the best bag classification performance in all few-shot settings, and significantly outperformed all comparison methods by a large margin. It can be observed that Linear-Probe with Mean/Max pooling can hardly work. Although using trainable attention pooling helps learn the importance of each instance and improves the performance of Linear-Probe, it still has limitations in performance. Prompt learning with fully trainable prompts in CoOp outperforms Linear-Probe. In contrast, our method used a multi-modal prompt learning paradigm, which achieved the best performance on both bag.

The Mean-pool strategy exhibits a consistent accuracy, achieving 67.56% in the 16-shot and maintaining a steady performance across all scenarios, dipping slightly to 65.11% in the 1-shot evaluation. The Max-pool method, while starting at a lower accuracy of 38.33% for 16-shot, shows a significant increase, peaking at 62.00% for 8-shot and 4-shot conditions. The Attn-pool method displays a strong performance, particularly in the 16-shot (80.00%) and 4-shot (75.33%) settings, but falls to 52.22% in the 1-shot scenario. The CoOp approach yields a moderate performance with a high of 69.94% for 16-shot, and the TOP method shows commendable results, especially at 78.89% for 4-shot classification. Our proposed method surpasses these figures, registering the highest accuracy of 85.15% in the 16-shot and consistently high performance down to 73.60% in the 1-shot condition. These outcomes underscore the efficacy of our method, indicating its superiority in leveraging few-shot learning for accurate classification in pathology image analysis.

Ablation Study

Table 3 showcases the results from an ablation study on the TCGA Lung Cancer dataset, evaluating the impact of different components on few-shot classification performance. These components include Trainable Image Prompt (TIP), Trainable Text Prompt (TTP), Descriptive Prompt (DP), and Instance Aggregator (IA). The study reveals that the use of TIP alone provides modest accuracy, reaching 64.31% in the 16-shot setting and declining to 50.58% in the 1-shot scenario. The integration of TTP enhances performance, particularly notable at 75.21% for 16-shot and 61.01% for 1-shot. The combination of TIP and DP further improves the results, achieving 81.22% in the 16-shot and 71.17% in the 1-shot context. Incorporating IA with TIP and DP yields the most significant improvements across all few-shot conditions, culminating in a peak accuracy of 84.25% for 16-shot and maintaining a robust 73.91% for 1-shot. These findings illustrate the synergistic effect of the combined components, with the full model configuration providing the most substantial gains in classification performance, underscoring the value of each element in the proposed few-shot learning framework.

Conclusion

In this study, we introduce a Multi-modal Prompt Learning approach for Few-shot Whole Slide Image Classification, termed MP-MIL, designed to address the FSWC challenge effectively. The MP-MIL framework leverages GPT-4 for generating bag-level visual descriptions, aiding in instance feature aggregation and facilitating bag-level prompt learning. Our experiments on the WSI classification task demonstrate the approach's high efficacy in FSWC tasks. The primary objective of this research is to encourage further exploration into the integration of foundational models with large-scale language models for pathology Whole Slide Image classification. We anticipate that such investigative efforts will mark the advent of a transformative phase in AI-driven pathology.

References

Bejnordi, B. E.; Veta, M.; Van Diest, P. J.; Van Ginneken, B.; Karssemeijer, N.; Litjens, G.; Van Der Laak, J. A.; Hermsen, M.; Manson, Q. F.; Balkenhol, M.; et al. 2017. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *Jama*, 318(22): 2199–2210.

Campanella, G.; Hanna, M. G.; Geneslaw, L.; Miraflor, A.; Werneck Krauss Silva, V.; Busam, K. J.; Brogi, E.; Reuter, V. E.; Klimstra, D. S.; and Fuchs, T. J. 2019. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature medicine*, 25(8): 1301–1309.

Jia, C.; Yang, Y.; Xia, Y.; Chen, Y. T.; and Duerig, T. 2021. Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision.

Lerousseau, M.; Vakalopoulou, M.; Classe, M.; Adam, J.; Battistella, E.; Carré, A.; Estienne, T.; Henry, T.; Deutsch, E.; and Paragios, N. 2020. Weakly supervised multiple instance learning histopathological tumor segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part V 23*, 470–479. Springer.

Li, B.; Li, Y.; and Eliceiri, K. W. 2021. 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*, 14318–14328.

Li, H.; Yang, F.; Zhao, Y.; Xing, X.; Zhang, J.; Gao, M.; Huang, J.; Wang, L.; and Yao, J. 2021. DT-MIL: deformable transformer for multi-instance learning on histopathological image. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VIII 24*, 206–216. Springer.

Lin, T.; Yu, Z.; Hu, H.; Xu, Y.; and Chen, C.-W. 2023. Interventional bag multi-instance learning on whole-slide pathological images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19830–19839.

Lu, M. Y.; Chen, T. Y.; Williamson, D. F.; Zhao, M.; Shady, M.; Lipkova, J.; and Mahmood, F. 2021. AI-based pathology predicts origins for cancers of unknown primary. *Nature*, 594(7861): 106–110.

Qu, L.; Luo, X.; Fu, K.; Wang, M.; and Song, Z. 2023. The Rise of AI Language Pathologists: Exploring Two-level Prompt Learning for Few-shot Weakly-supervised Whole Slide Image Classification. *arXiv preprint arXiv:2305.17891.*

Qu, L.; Luo, X.; Liu, S.; Wang, M.; and Song, Z. 2022. Dgmil: Distribution guided multiple instance learning for whole slide image classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 24–34. Springer.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; and Clark, J. 2021. Learning Transferable Visual Models From Natural Language Supervision.

Shao, Z.; Bian, H.; Chen, Y.; Wang, Y.; Zhang, J.; Ji, X.; et al. 2021. Transmil: Transformer based correlated multiple instance learning for whole slide image classification. *Advances in neural information processing systems*, 34: 2136–2147.

Wang, X.; Yan, Y.; Tang, P.; Bai, X.; and Liu, W. 2018. Revisiting multiple instance neural networks. *Pattern Recognition*, 74: 15–24.

Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771.*

Xu, G.; Song, Z.; Sun, Z.; Ku, C.; Yang, Z.; Liu, C.; Wang, S.; Ma, J.; and Xu, W. 2019. Camel: A weakly supervised learning framework for histopathology image segmentation. In *Proceedings of the IEEE/CVF International Conference on computer vision*, 10682–10691.

Yao, L.; Huang, R.; Hou, L.; Lu, G.; Niu, M.; Xu, H.; Liang, X.; Li, Z.; Jiang, X.; and Xu, C. 2021. FILIP: Fine-grained Interactive Language-Image Pre-Training.

Yu, S.; Ma, K.; Bi, Q.; Bian, C.; Ning, M.; He, N.; Li, Y.; Liu, H.; and Zheng, Y. 2021. Mil-vt: Multiple instance learning enhanced vision transformer for fundus image classification. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VIII 24*, 45–54. Springer.

Yufei, C.; Liu, Z.; Liu, X.; Liu, X.; Wang, C.; Kuo, T.-W.; Xue, C. J.; and Chan, A. B. 2022. Bayes-MIL: A New Probabilistic Perspective on Attention-based Multiple Instance Learning for Whole Slide Images. In *The Eleventh International Conference on Learning Representations*.

Zhang, H.; Meng, Y.; Zhao, Y.; Qiao, Y.; Yang, X.; Coupland, S. E.; and Zheng, Y. 2022. DTFD-MIL: Double-tier feature distillation multiple instance learning for histopathology whole slide image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18802–18812.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9): 2337–2348.