Supplementary Materials for PSA-MIL: A Probabilistic Spatial Attention-Based Multiple Instance Learning for Whole Slide Image Classification

Sharon Peled Yosef E. Maruvka* Moti Freiman*
Technion – Israel Institute of Technology, Israel.
sharonpe@campus.technion.ac.il

1. Overview

This supplementary material provides additional experiments and technical details to support our paper on PSA-MIL. We present extensive ablation studies to evaluate the impact of key hyperparameters such as the regularization coefficient α and spatial pruning threshold τ. We also visualize the attention mechanism of PSA-MIL and discuss the dynamics of attention across different heads in the model. Additionally, we provide insights into the initialization and reparameterization strategies used to optimize spatial decay functions.

2. Additional Experiments

2.1. Survival Analysis

We provide additional survival prediction experiments across multiple encoders to further assess the robustness of our approach. Results presented in Tab. 1 show that PSA-MIL$_{[\text{Gau}]}$ consistently outperforms other methods, regardless of the encoder used. Specifically, it maintains top performance on TCGA-CRC and TCGA-STAD, and performs competitively on TCGA-BRCA, further validating its effectiveness.

The results across different encoders highlight PSA-MIL's robustness and generalizability, as it delivers superior or comparable performance even when the underlying encoder changes. This provides stronger evidence for the model's ability to generalize well to diverse settings, reinforcing its potential for real-world applications. The comparative performance of alternative methods like Sm-MIL and BayesMIL underscores the consistent advantage of PSA-MIL in survival prediction tasks.

2.2. Metastatic Detection

For metastatic detection, we use the CAMELYON16 dataset [1], evaluating both slide-level classification and patch-level localization. The ground truth for localization consists of a manually annotated test set of 100 whole-slide images. Each tile is considered cancerous if it contains at least 25% annotated tumor.

For each model, patch-level scores were generated according to the implementation described in its original paper: for PSA-MIL, ABMIL, SM-MIL, and TransMIL, the scores correspond to scaled attention values, which highlight the regions of interest. For DTFD-MIL and BayesMIL, the patch scores are taken from the logits, reflecting the model's classification confidence for each patch. GradCam is used for generating patch-level scores for GTP. Patch scores are crucial for both localization and visualization tasks, providing the necessary information to identify cancerous regions and generate accurate attention maps.

Results are presented in Tab. 2. PSA-MIL demonstrates strong localization performance, achieving the highest AUC-FROC and AUC-ROC scores while also excelling in slide-level classification, with an AUC exceeding 96%.

2.3. Ablation Studies

In this section, we focused on subtyping tasks, utilizing both TCGA-CRC (MSS/MSI) and TCGA-STAD (CIN/GS) datasets, with patient-level AUC as the evaluation metric.

2.3.1. Impact of α Value

Our approach employs diversity loss to encourage the capture of diverse spatial patterns. As shown in Fig. 2 (a-b), TCGA-CRC favored relatively higher α values, indicating the benefits of diverse representations. On the other hand, a moderate α value yields the best results across all decay functions for TCGA-STAD.

2.3.2. Impact of τ Value

Our approach leverages τ to regulate spatial pruning, controlling the range of spatial interactions in PSA-MIL. In our experiments, we typically set it to a constant value (e.g., $1e{-}3$), which we find works well in practice. As shown in Fig. 2 (c-d), Gaussian decay, which models gradual changes, remains relatively stable to variations in τ. In contrast, Cauchy decay, which favors long-range dependencies, is more sensitive, exhibiting steeper performance

*These authors are co-senior authors.

Dataset	Encoder	ABMIL	DTFD-MIL	TransMIL	Sm-MIL	BayesMIL	PSA-MIL$_{[Gau]}$
TCGA-CRC	UNI [4]	63.8 ± 7.1	55.5 ± 6.1	59.9 ± 5.1	63.9 ± 6.3	55.9 ± 5.8	**70.7 ± 3.5**
	Gigapath [16]	65.1 ± 6.8	53.3 ± 9.5	57.5 ± 4.8	55.6 ± 5.8	56.3 ± 9.8	**70.0 ± 5.3**
	Lunit [8]	61.1 ± 3.3	57.3 ± 6.1	57.8 ± 4.0	58.5 ± 6.1	55.5 ± 6.6	**65.8 ± 3.3**
TCGA-STAD	UNI [4]	58.6 ± 5.8	53.3 ± 7.1	52.9 ± 5.0	56.6 ± 5.5	52.4 ± 3.6	**61.1 ± 6.8**
	Gigapath [16]	58.1 ± 4.1	52.9 ± 4.7	50.5 ± 4.7	52.6 ± 6.2	51.2 ± 3.9	**56.6 ± 4.4**
	Lunit [8]	58.3 ± 6.6	49.8 ± 2.6	53.8 ± 6.8	58.5 ± 3.8	51.0 ± 6.2	**59.3 ± 6.1**
TCGA-BRCA	UNI [4]	**61.9 ± 6.7**	53.9 ± 6.2	57.8 ± 4.0	57.2 ± 7.0	49.9 ± 5.3	60.2 ± 6.1
	Gigapath [16]	61.8 ± 6.0	53.1 ± 6.2	58.7 ± 5.8	56.0 ± 5.3	49.3 ± 7.2	**62.3 ± 5.1**
	Lunit [8]	**59.9 ± 5.3**	45.8 ± 3.2	57.7 ± 2.5	50.5 ± 1.8	46.3 ± 3.6	58.0 ± 5.7

Table 1. Concordance index (C-index) comparison of survival prediction models across TCGA-CRC, TCGA-STAD, and TCGA-BRCA datasets using different feature encoders. PSA-MIL$_{[Gau]}$ consistently outperforms all baselines with a clear performance margin in TCGA-CRC and TCGA-STAD across all encoders. In TCGA-BRCA, it achieves either the best or second-best performance, demonstrating strong and robust generalization.

	Slide-level		Patch-level		
Method	AUC (↑)	F1 (↑)	AUC-FROC (↑)	AUC-ROC (↑)	F1 (↑)
ABMIL [7]	95.3 ± 1.8	91.5 ± 1.1	67.8 ± 0.9	88.8 ± 1.7	66.6 ± 1.2
DTFD-MIL [18]	96.1 ± 1.6	92.2 ± 1.8	68.3 ± 2.6	88.1 ± 1.5	65.9 ± 1.4
TransMIL [14]	94.8 ± 1.8	90.5 ± 1.1	67.7 ± 1.5	84.6 ± 2.3	21.1 ± 4.5
GTP [20]	91.7 ± 2.1	90.1 ± 2.1	69.9 ± 1.4	79.6 ± 2.4	46.1 ± 1.0
SM-MIL [2]	**96.8 ± 1.4**	89.6 ± 1.6	69.9 ± 0.9	91.6 ± 1.1	64.6 ± 1.3
BAYES-MIL [5]	95.0 ± 2.2	91.8 ± 2.5	72.5 ± 1.8	93.6 ± 1.2	**76.5 ± 2.0**
PSA-MIL$_{[Gau]}$	96.1 ± 1.6	**92.3** ± 2.1	**75.9 ± 1.6**	**94.7 ± 0.9**	75.4 ± 1.8

Table 2. Slide-level classification and patch-level localization performance on CAMELYON16. PSA-MIL achieved top slide-level performance while maintaining superior localization quality.

drops. Despite the expected performance drops, PSA-MIL maintains strong results even under significant computational constraints.

2.4. Training Dynamics: Attention Visualizations

To demonstrate PSA-MIL's ability to capture diverse spatial patterns, we visualize the attention maps generated during inference. These maps are produced by extracting attention weights from each head during the forward pass, showing which regions of the input the model focuses on.

To assess the accuracy of these heatmaps, we automatically segmented tumor regions using the NCT-CRC-HE-100K dataset [9], a patch-level tissue classification task. We trained a pretrained VisionTransformer and achieved an F1 score greater than 0.99 on an independent test set for the tumor/non-tumor classification task, ensuring reliable tumor localization.

In Fig. 1, Panel (A) shows the tumor segmentation results. Panels (B-D) present the attention heatmaps for individual heads in PSA-MIL$_{[Gau]}$. Head 2 (K=4) shows highly localized attention, while head 3 (K=17) captures global dependencies. Panel (E) aggregates the attention from all heads, revealing the overall focus areas.

These attention patterns demonstrate PSA-MIL's ability to capture diverse spatial patterns, mitigating the risk of learning redundant representations.

Note that this analysis is not about tumor localization accuracy, since the tumor map is generated automatically and not manually annotated (for localization results, see 2.2 on metastatic detection). Instead, this demonstrates PSA-MIL$_{[Gau]}$ ability to capture diverse spatial representations.

3. Comparing to Additional Literature

Our evaluation benchmarks PSA-MIL against strong recent baselines, such as SM-MIL[2] (NeurIPS 2024), BAYES-MIL[5] (ICLR 2023), and IBMIL[11] (CVPR 2023), as well as up-to-date encoders like UNI[4] (Nature Medicine 2025), GigaPath[16] (Nature 2025), and Lunit[8] (CVPR 2023). To further situate PSA-MIL within the broader landscape of computational pathology, we extend this discussion to additional seminal works in the field, including HIPT[3] (CVPR 2022), BEPH[17] (Nat. Commun. 2025),

Figure 1. (A) Tumor segmentation using NCT-CRC-HE-100K [9]. (B-D) Attention heatmaps of individual heads in PSA-MIL$_{[Gau]}$, illustrating diverse spatial focus. (E) Aggregated attention heatmap from the attention pooling operator.

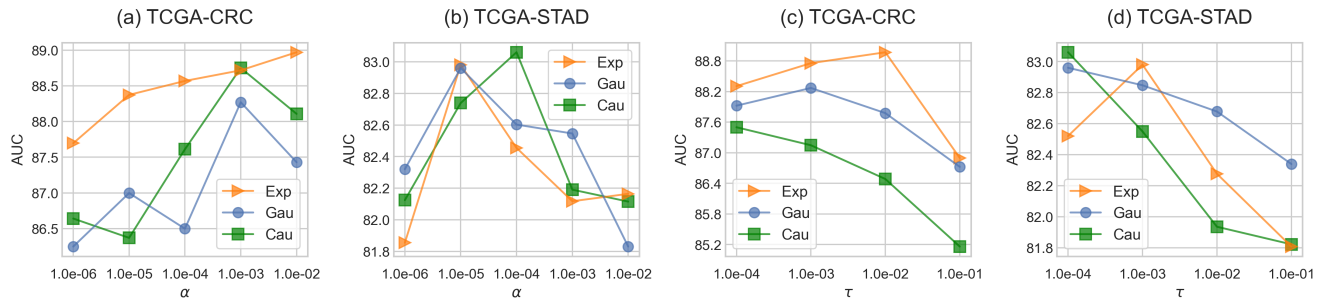

Figure 2. Ablation studies on the impact of α (a-b) and τ (c-d) in PSA-MIL.

and CTransPath[15] (MICCAI 2022).

These works primarily contribute large-scale pretrained encoders, later adapted with MIL heads for downstream classification and survival tasks. In contrast, PSA-MIL focuses on the design of the MIL module itself. Direct benchmarking against these models would require substantial effort to fully align training conditions, which is beyond the scope of this work.

Nevertheless, to still contextualize our contribution within the broader literature, we compare performance on shared benchmarks and provide additional discussion of contributions and limitations. We note that such comparisons are inherently constrained by differences in encoder architectures, model sizes, pretraining data, and computational resources, which we highlight where relevant.

HIPT[3] is one of the first hierarchical self-supervised foundation models, demonstrating strong survival prediction on TCGA-CRC and TCGA-STAD. In our study, PSA-MIL is evaluated on the same benchmarks through three different foundation encoders. Among these, Lunit (d=384) provides the most relevant point of comparison, as HIPT also operates with a relatively small encoder, making this the fairest evaluation setting. On TCGA-CRC survival, HIPT reports a C-index of 0.608, while PSA-MIL with Lunit features achieves 0.658, an 8% improvement. On TCGA-STAD survival, HIPT reaches 0.570, compared to 0.593 for PSA-MIL, a 4% gain. PSA-MIL matches or outperforms HIPT on both tasks, highlighting its contributions to the field. Nevertheless, although we observe substantial improvements, these comparisons are not under identical conditions and should therefore be interpreted with caution.

CTransPath[15] is a foundation model pretrained on millions of histopathology patches from TCGA and PAIP using semantically relevant contrastive learning. Its backbone combines a CNN with a Swin Transformer and produces 768-dimensional features. The main benchmark shared with PSA-MIL is Camelyon16, a widely used dataset for metastasis detection. For this comparison, we report PSA-MIL results using the Lunit encoder (d=384), which is considerably smaller than CTransPath's 768-dimensional backbone. Despite this lower feature dimensionality, PSA-MIL achieves competitive or better results: on Camelyon16, PSA-MIL reaches an AUC of 96.1 compared to 94.2 for CTransPath, and an accuracy of 92.3 versus 92.2. These results highlight the efficiency of our probabilistic MIL design.

BEPH[17] is a recent ViT-Base foundation model (768d, ∼193M parameters) pretrained on 11.7k WSIs from TCGA using masked image modeling. On TCGA-CRC and TCGA-STAD survival prediction, BEPH reports C-index values of ∼0.69–0.70 and 0.60, respectively. PSA-MIL achieves 0.658 (CRC) and 0.593 (STAD) with the compact Lunit encoder (384d), and 0.707 (CRC) and 0.611 (STAD) with the larger UNI encoder (1536d). These results confirm BEPH as an efficient and strong foundation model for survival analysis, while showing that PSA-MIL remains competitive.

4. Clarifications on Mathematical Derivations

We provide additional details to clarify intermediate steps in the derivations presented in the main paper.

4.1. Posterior Reformulation (Sec. 3.1.2)

Equation (3) in the main paper expresses the posterior as

$$p(t_j = 1 \mid q_i) = \frac{\pi_j \, \mathcal{N}(q_i \mid k_j, \sigma^2 I)}{\sum_{j'} \pi_{j'} \, \mathcal{N}(q_i \mid k_{j'}, \sigma^2 I)}.$$

Expanding the Gaussian likelihood and collecting terms gives

$$p(t_j = 1 \mid q_i) = \tag{1}$$

$$\frac{\pi_j \exp\Big(-\frac{1}{2\sigma^2}\|q_i\|^2 - \frac{1}{2\sigma^2}\|k_j\|^2 + \frac{1}{\sigma^2} q_i^\top k_j \Big)}{\sum_{j'} \pi_{j'} \exp\Big(-\frac{1}{2\sigma^2}\|q_i\|^2 - \frac{1}{2\sigma^2}\|k_{j'}\|^2 + \frac{1}{\sigma^2} q_i^\top k_{j'} \Big)}. \tag{2}$$

Under Assumptions 1–3, the norm terms cancel and $\pi_j = 1/N$, yielding

$$p(t_j = 1 \mid q_i) = \frac{\exp\big(q_i^\top k_j / \sqrt{d_k}\big)}{\sum_{j'} \exp\big(q_i^\top k_{j'} / \sqrt{d_k}\big)},$$

which is Eq. (4) in the main text. Relaxing the uniform prior assumption and introducing distance-decayed priors $\pi_j = f(d_{ij}|\theta)$ leads directly to Eq. (6), as the $\log f(d_{ij}|\theta)$ term appears additively inside the exponential.

4.2. Diversity Loss (Sec. 3.3)

Each head is parameterized by a decay θ_h. The entropy term in Eq. (8) is derived as follows. Starting from the entropy

$$H(p) = -\int p(\theta) \log p(\theta) \, d\theta,$$

we approximate $p(\theta)$ using KDE:

$$\hat{p}(\theta) = \tfrac{1}{H\sigma} \sum_{h=1}^{H} K\big(\tfrac{\theta - \theta_h}{\sigma}\big) \, .$$

Substituting $\hat{p}$ into $H(p)$ and approximating the integral by Monte Carlo sampling yields

$$H(p) \approx -\frac{1}{M} \sum_{m=1}^{M} \log \hat{p}(\tilde{\theta}_m),$$

which corresponds to Eq. (8). The diversity loss in Eq. (9) is then defined as $L_{\text{Diversity}} = -H(p)$.

5. Implementation Details

While all tasks used the same overall architecture of a single layer with three heads, there were some task-specific modifications: for subtyping tasks, the model used 32-dimensional heads, whereas for survival prediction and metastatic detection, the model utilized 324-dimensional heads.

Our architecture choice was based on experimentation. In histopathology, models like TransMIL[14] or LongMIL[10] often stack multiple attention layers to capture long-range spatial dependencies. However, our spatial attention is capable of capturing long-range spatial dependencies even when utilizing a single-layer design, through our learned spatial relationships. When stacking layers, in our experiments, we did not observe a significant improvement in performance. Instead, stacking additional layers sometimes led to performance reductions due to overfitting. This behavior can be attributed to the adaptability of our approach and to the nature of histopathology datasets, which are typically small, often consisting of only a few hundred samples. Each input, such as whole-slide images, has high dimensionality, increasing the risk of learning spurious correlations or noise from the data. In such settings, additional layers can compound the problem by further amplifying the model's tendency to overfit.

Our choice of a lightweight model with a single attention layer not only minimizes overfitting but also enhances the model's interpretability, as seen in Fig. 1. Additionally, this design choice leads to significantly faster training times: while other methods typically require around 100 epochs, our model was able to achieve strong performance after only around 15 epochs. As well as inference efficiency presented in Fig. 5 in the main text.

The model was trained using the Adam optimizer [6] for 15-30 epochs (depending on loss plateau), with a learning rate of 1×10^{-4} and a weight decay of 1×10^{-4}. A batch size of 8 was used for all experiments. The learning rate followed a linear scheduling strategy, with 10% of the training iterations allocated for warmup and 10% for cooldown phases. The regularization coefficient α in the diversity loss term was set to 1×10^{-3} for TCGA-CRC and 1×10^{-4} for TCGA-STAD. The primary loss function was the cross-entropy (CE) loss [13]. Unless otherwise specified, experiments were conducted using Lunit foundation backbone [8]. The pruning parameter was set constant to $\tau = 10^{-3}$ during all experiments.

Results were obtained using a 5-fold cross-validation, with data splits stratified based on the target variable (y) at the patient level to prevent data leakage.

For additional details, please refer to our codebase.

5.1. Initialization

The general network weights were initialized randomly (default behavior). However, given the significant influence of θ in $f(d|\theta)$ on the model's spatial constraints, the initialization of these parameters was handled with greater care.

For the spatial decay functions, we observed that the initialization of λ (in the case of exponential decay) and σ (in the case of Gaussian decay) between 0 and 1 had substantial implications for the spatial constraints imposed on the model. For instance, a σ value of 0.5 implies a decay of less than 10^{-85} for instances with a distance greater than 10, while the same value for λ would result in a decay of over 10^{-3}. Consequently, each decay function required its own tailored initialization strategy.

Specifically, θ was initialized by solving the equation:

$$f(d = 10|\theta) = 0.1,$$

ensuring that the decay for instances at a distance of 10 was set to 0.1. This initialization provided a meaningful starting point for spatial constraints. Random sampling was applied around this value to introduce slight variations for each attention head.

5.2. Reparameterization

Rather than learning spatial decay parameters θ directly, which can exhibit varying magnitudes across different parameterizations (as described in the Initialization section), we instead learn a rate parameter that controls the value of θ within a predefined range.

Specifically, we constrain θ within a predefined range by introducing lower and upper bounds, θ_{min} and θ_{max}, and define:

$$\theta = \theta_{\text{min}} + \text{rate} \times (\theta_{\text{max}} - \theta_{\text{min}}),$$

where the learned rate $\in [0, 1]$ ensures that θ remains within a stable and interpretable range.

The bounds θ_{min} and θ_{max} are determined by solving the following equations:

$$f(d = 1|\theta_{\text{min}}) = \tau, \quad f(d = 30|\theta_{\text{max}}) = \tau.$$

This ensures that the dynamic local attention mechanism spans from $K = 1$ up to $K = 30$, allowing the model to adaptively adjust the range of spatial interactions based on the learned decay function. This parameterization stabilizes learning while maintaining flexibility in capturing spatial dependencies.

6. Discussion on Spatial Context Modeling using Graph Neural Networks

Graph Neural Networks (GNNs) are widely used to model spatial dependencies in histopathology, where patches are treated as nodes in a graph [12, 19], and typically only immediate neighbors are being connected. This limits the propagation of information, as the spatial context is confined to a fixed number of graph operations (e.g., GCN, GAT). Increasing the number of operations to capture broader spatial dependencies risks over-smoothing, especially in small datasets.

A different example is Graph-based Transformer (GTP) [20], which tries to identify key nodes before training a transformer. This results in a costly step, which can disrupt spatial continuity, as only a subset of the nodes is used for training, neglecting many spatial relationships in the process.

In contrast, PSA-MIL learns spatial relationships through a flexible, distance-decayed probabilistic self-attention mechanism, capturing spatial context more naturally and efficiently without the need for predefined graph edges or computationally expensive processing.

Overall, GNNs offer a flexible framework for modeling spatial context; however, their application to histopathology is limited by the need to process large graphs with relatively small datasets, which restricts generalization. Recent state-of-the-art approaches instead leverage transformers and other attention-based architectures, which are better equipped to address these challenges.

References

[1] Babak Ehteshami Bejnordi, Mitko Veta, Paul Johannes Van Diest, Bram Van Ginneken, Nico Karssemeijer, Geert Litjens, Jeroen AWM Van Der Laak, Meyke Hermsen, Quirine F Manson, Maschenka Balkenhol, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *Jama*, 318(22):2199–2210, 2017. 1

[2] Francisco M Castro-Macías, Pablo Morales-Álvarez, Yunan Wu, Rafael Molina, and Aggelos K Katsaggelos. Sm: enhanced localization in multiple instance learning for medical imaging classification. *arXiv preprint arXiv:2410.03276*, 2024. 2

[3] Richard J Chen, Chengkuan Chen, Yicong Li, Tiffany Y Chen, Andrew D Trister, Rahul G Krishnan, and Faisal Mahmood. Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16144–16155, 2022. 2, 3

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

[5] Yufei Cui, Ziquan Liu, Xiangyu Liu, Xue Liu, Cong Wang, Tei-Wei Kuo, Chun Jason Xue, and Antoni B Chan. Bayes-mil: A new probabilistic perspective on attention-based multiple instance learning for whole slide images. In *11th In-*

ternational Conference on Learning Representations (ICLR 2023), 2023. 2

[6] P Kingma Diederik. Adam: A method for stochastic optimization. *(No Title)*, 2014. 4

[7] 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

[8] Mingu Kang, Heon Song, Seonwook Park, Donggeun Yoo, and Sérgio Pereira. Benchmarking self-supervised learning on diverse pathology datasets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3344–3354, 2023. 2, 4

[9] Jakob Nikolas Kather, Niels Halama, and Alexander Marx. 100,000 histological images of human colorectal cancer and healthy tissue. *Zenodo10*, 5281:6, 2018. 2, 3

[10] Honglin Li, Yunlong Zhang, Pingyi Chen, Zhongyi Shui, Chenglu Zhu, and Lin Yang. Rethinking transformer for long contextual histopathology whole slide image analysis. *arXiv preprint arXiv:2410.14195*, 2024. 4

[11] Tiancheng Lin, Zhimiao Yu, Hongyu Hu, Yi Xu, and Chang-Wen Chen. Interventional bag multi-instance learning on whole-slide pathological images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19830–19839, 2023. 2

[12] Yangling Ma, Yixin Luo, and Zhouwang Yang. Gcn-based mil: multi-instance learning utilizing structural relationships among instances. *Signal, Image and Video Processing*, 18 (6):5549–5561, 2024. 5

[13] Anqi Mao, Mehryar Mohri, and Yutao Zhong. Cross-entropy loss functions: Theoretical analysis and applications. In *International conference on Machine learning*, pages 23803–23828. PMLR, 2023. 4

[14] 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, 4

[15] Xiyue Wang, Sen Yang, Jun Zhang, Minghui Wang, Jing Zhang, Wei Yang, Junzhou Huang, and Xiao Han. Transformer-based unsupervised contrastive learning for histopathological image classification. *Medical image analysis*, 81:102559, 2022. 3

[16] Hanwen Xu, Naoto Usuyama, Jaspreet Bagga, Sheng Zhang, Rajesh Rao, Tristan Naumann, Cliff Wong, Zelalem Gero, Javier González, Yu Gu, et al. A whole-slide foundation model for digital pathology from real-world data. *Nature*, 630(8015):181–188, 2024. 2

[17] Zhaochang Yang, Ting Wei, Ying Liang, Xin Yuan, Ruitian Gao, Yujia Xia, Jie Zhou, Yue Zhang, and Zhangsheng Yu. A foundation model for generalizable cancer diagnosis and survival prediction from histopathological images. *Nature Communications*, 16(1):2366, 2025. 2, 3

[18] Hongrun Zhang, Yanda Meng, Yitian Zhao, Yihong Qiao, Xiaoyun Yang, Sarah E Coupland, and Yalin Zheng. 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*, pages 18802–18812, 2022. 2

[19] Yu Zhao, Fan Yang, Yuqi Fang, Hailing Liu, Niyun Zhou, Jun Zhang, Jiarui Sun, Sen Yang, Bjoern Menze, Xinjuan Fan, et al. Predicting lymph node metastasis using histopathological images based on multiple instance learning with deep graph convolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4837–4846, 2020. 5

[20] Yi Zheng, Rushin H Gindra, Emily J Green, Eric J Burks, Margrit Betke, Jennifer E Beane, and Vijaya B Kolachalama. A graph-transformer for whole slide image classification. *IEEE transactions on medical imaging*, 41(11):3003–3015, 2022. 2, 5