AEM: Attention Entropy Maximization for Multiple Instance Learning based Whole Slide Image Classification

Yunlong Zhang[1,2], Honglin Li[1,2], Yuxuan Sun[1,2], Zhongyi Shui[1,2], Jingxiong Li[1,2], Chenglu Zhu[2], and Lin Yang[2(✉)]

[1] College of Computer Science and Technology, Zhejiang University
[2] School of Engineering, Westlake University
yanglin@westlake.edu.cn

Abstract. Multiple Instance Learning (MIL) effectively analyzes whole slide images but faces overfitting due to attention over-concentration. While existing solutions rely on complex architectural modifications or additional processing steps, we introduce Attention Entropy Maximization (AEM), a simple yet effective regularization technique. Our investigation reveals the positive correlation between attention entropy and model performance. Building on this insight, we integrate AEM regularization into the MIL framework to penalize excessive attention concentration. To address sensitivity to the AEM weight parameter, we implement Cosine Weight Annealing, reducing parameter dependency. Extensive evaluations demonstrate AEM's superior performance across diverse feature extractors, MIL frameworks, attention mechanisms, and augmentation techniques. Here is our anonymous code: `https://github.com/dazhangyu123/AEM`.

Keywords: Whole slide image · Multiple instance learning · Overfitting.

1 Introduction

Whole slide images (WSIs) are widely recognized as the gold standard for numerous cancer diagnoses, playing a crucial role in ensuring precise diagnosis [2], prognosis [28], and the development of treatment plans [21]. In recent years, attention-based multiple instance learning (ABMIL) [9] has emerged as a promising approach for WSI analysis. However, recent studies have uncovered overfitting issues in MIL due to factors like limited available data [23,31,32,11], class imbalance [32], and staining bias [12,33].

In the attention mechanism, attention values represent the importance or relevance of instances to the bag prediction, influencing both prediction accuracy and result interpretability. Relevant studies [29,32] have revealed that excessive concentration of attention values in ABMIL hinders model interpretability and results in overfitting [32]. There have been several solutions for alleviating attention concentration. Masking-based methods [16,23,32] mask out the instances

Table 1: Comparison of WSI classification methods addressing overfitting.

Method	**Extra Modules/Processing**
DTFD-MIL [31]	Double-tier attention mechanisms
IBMIL [12]	New training stage of interventional training from scratch
C2C [19]	Clustering and sampling process
MHIM-MIL [23]	Teacher model for masking easy instances
ACMIL [32]	Multiple branch attention for extracting pattern embeddings
DGR-MIL [1]	Instance center pushing and DPP-based vector orthogonality
AEM(ours)	**None**

with the highest attention values, allocating their attention values to remaining instances. Clustering-based methods [19,7] group instances into clusters and randomly sample instances from these clusters, ensuring attention values are not overly focused on minority instances. ACMIL [32] generates the heatmap by averaging the attention values generated by multiple attention heads, thereby avoiding the over-concentration of attention values. DGR-MIL [1] addresses this through learnable global vectors capturing diverse patterns via cross-attention, with strategies to push vectors toward positive instance centers and enforce orthogonality using DPP-based diversity loss. However, most solutions add complexity and computational overhead, limiting flexibility (see Table 1).

To address the limitations of existing complex solutions, we propose Attention Entropy Maximization (AEM), a lightweight yet powerful approach for mitigating attention concentration and MIL method overfitting. Our empirical analysis establishes a positive correlation between attention entropy and model performance, which forms the foundation for developing AEM. The approach integrates a negative entropy loss term for attention values into the standard MIL framework (Figure 1), promoting a more uniform distribution of attention across instances. To address sensitivity to the AEM weight parameter, we introduce Cosine Weight Annealing, reducing parameter dependency. Unlike existing overfitting mitigation techniques, AEM requires no additional modules or processing steps, enabling seamless integration with current MIL frameworks while maintaining computational efficiency.

Our experimental evaluations on three datasets (CAMELYON16, CAMELYON17, and our in-house LBC dataset) demonstrate AEM's superior performance over existing methods. Furthermore, extensive experiments showcase AEM's versatility, effectively combining with five feature extractors (Lunit pretrained ViT-S [10], PathGen-CLIP pretrained ViT-L [22], UNI pretrained ViT-L [3], CONCH pretrained ViT-B [14], and GigaPath pretrained ViT-G [27]), Subsampling augmentation technique, two advanced MIL frameworks (DTFD-MIL [31] and ACMIL [32]), and three attention mechanisms (DSMIL [11], LongNet [6] and MHA [24]). These results underscore AEM's potential as a widely applicable enhancement to existing MIL methodologies in medical image analysis.

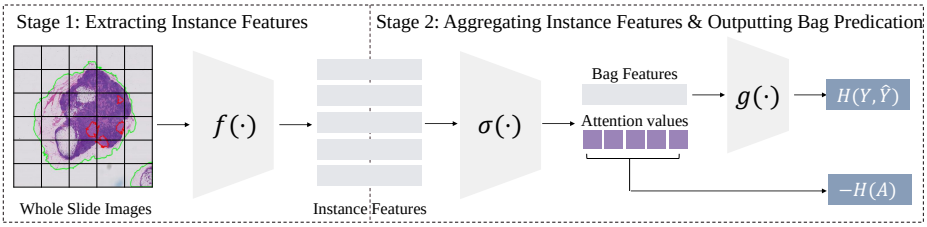


Fig. 1: Overview of plugging AEM into MIL framework. AEM adds only a negative entropy regularization for attention values to the regular MIL framework.

2 Method

2.1 ABMIL for WSI Analysis

MIL formulation. For WSI classification, we have the WSI $\boldsymbol{X}$ with slide-level label $\boldsymbol{Y}$. Due to the extreme resolution of WSIs ($50,000 \times 50,000$ to $100,000 \times 100,000$), direct training is computationally infeasible. ABMIL [9] addresses this by segmenting WSIs into non-overlapping patches $\{\boldsymbol{x}_n\}_{n=1}^{N}$ and employing a two-step process to predict the slide label $\hat{\boldsymbol{Y}}$.

Extracting instance features. Current MIL methods typically use features from a frozen backbone like ImageNet-pretrained ResNet. Recent studies [15,4] show that using encoders pre-trained with self-supervised learning and vision-language pretraining improves performance. To comprehensively verify AEM's effectiveness, we use five feature extractors: DINO pretrained ViT-S/16 [10], PathGen-CLIP pretrained ViT-L/14 [22], UNI pretrained ViT-L [3], CONCH pretrained ViT-B [14], and GigaPath pretrained ViT-G [27]).

Aggregating instance features and outputting bag predication. ABMIL aggregates instance embeddings into the bag embedding using a gated attention operator:

$$\boldsymbol{z} = \sum_{n=1}^{N} a_n \boldsymbol{h}_n, \tag{1}$$

where $a_n = \sigma(\boldsymbol{h}_n)$ represents the attention values for the n-th instance, $\boldsymbol{h}_n$. The bag prediction is then obtained through an MLP layer: $\hat{\boldsymbol{Y}} = g(\boldsymbol{z})$.

While we initially demonstrate our approach with ABMIL, our proposed AEM method (detailed in Section 2.2) can be effectively applied to various attention mechanisms, including DSMIL [11], MHA [32], and LongNet [27].

2.2 Attention Entropy Maximization

Motivation. Studies show that low attention entropy can cause training instability and poor generalization in attention-based models [30,26,5]. To investigate this in WSI classification, we trained ABMIL with 200 different random initializations while keeping training, validation, and test sets fixed. Figure 2 reveals a

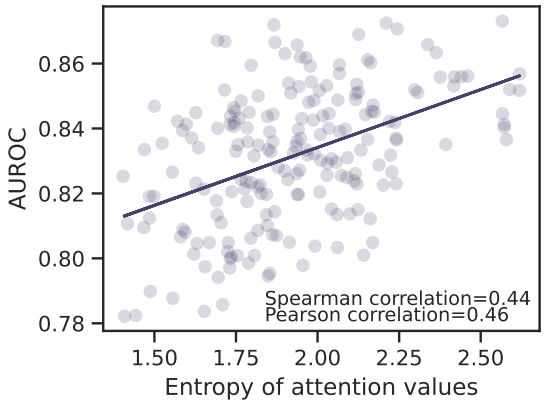


Fig. 2: There exists a positive correlation between AUROC values and entropy of attention values across experimental seeds. One point denotes the outcome of a single seed on the LBC dataset.

positive correlation between AUROC performance and attention entropy values on the test set, with higher entropy consistently associated with better classification results. These findings highlight the importance of attention diversity for effective WSI analysis, demonstrating that maintaining high attention entropy improves model performance and generalization.

Implementation. AEM maximizes the entropy $H(A)$ of attention values, $A = \{a_n\}_{n=1}^{N}$, by formulating it as negative entropy [17]:

$$L_{aem} = -H(A) = \sum_{n} a_n \log a_n. \tag{2}$$

This encourages consideration of more informative regions in WSIs, potentially improving generalization. The final objective is formulated as:

$$L_{total} = L_{ce} + \lambda L_{aem}, \tag{3}$$

where λ is a hyperparameter that controls the relative contribution of the AEM regularization term to the total loss function, balancing it against the task loss L_{ce}. Too small values lead to insufficient instance diversity, while too large values force uniform attention distribution, reducing to mean-pooling behavior—consistently shown inferior to attention-based MIL approaches [9].

We adopt Cosine Weight Annealing [13] as our scheduling strategy, which gradually reduces λ following a cosine curve. This approach naturally supports AEM's progression: initially maintaining high entropy for broad instance exploration when features are less reliable, then transitioning to focused attention as discriminative capabilities improve.

Discussion. AEM serves a similar role to the KL-divergence loss in C2C [19] by promoting attention distribution, but with key differences. AEM operates globally across all instances, while C2C's KL-divergence works only within individual

Table 2: The performance of different MIL approaches across three datasets and two evaluation metrics. The most superior performance is highlighted in **bold**.

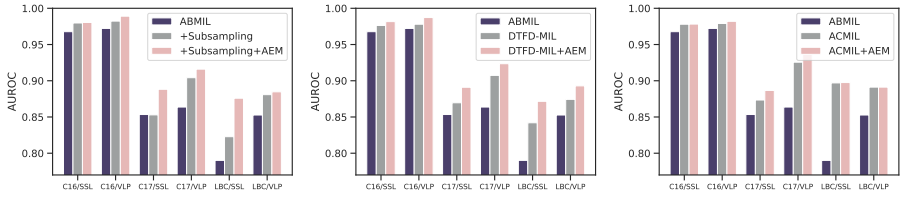
Method	CAMELYON-16		CAMELYON-17		LBC	
	F1-score	AUC	F1-score	AUC	F1-score	AUC
SSL pretrained ViT-S (Lunit [10])						
Clam-SB [15]	0.925±0.035	0.969±0.024	0.523±0.020	0.846±0.020	0.617±0.022	0.865±0.018
LossAttn [20]	0.908±0.031	0.928±0.014	0.575±0.051	0.865±0.016	0.621±0.012	0.843±0.006
TransMIL [18]	0.922±0.019	0.943±0.009	0.554±0.048	0.792±0.029	0.539±0.028	0.805±0.010
DSMIL [11]	0.943±0.007	0.966±0.009	0.532±0.064	0.804±0.032	0.562±0.028	0.820±0.033
IBMIL [12]	0.912±0.034	0.954±0.022	0.557±0.034	0.850±0.024	0.604±0.032	0.834±0.014
MHIM-MIL [23]	0.932±0.024	0.970±0.037	0.541±0.022	0.845±0.026	0.658±0.041	0.872±0.022
ILRA [25]	0.904±0.071	0.940±0.060	0.631±0.051	0.860±0.020	0.618±0.051	0.859±0.017
ABMIL [9]	0.914±0.031	0.945±0.027	0.522±0.050	0.853±0.016	0.595±0.036	0.831±0.022
AEM(ours)	**0.947±0.003**	**0.974±0.007**	**0.647±0.007**	**0.887±0.013**	**0.664±0.021**	**0.879±0.013**
VLM pretrained ViT-L (PathGen-CLIP [22])						
Clam-SB [15]	0.941±0.014	0.960±0.015	0.622±0.031	0.899±0.012	0.641±0.025	0.870±0.013
LossAttn [20]	0.948±0.004	0.981±0.017	0.667±0.023	0.891±0.009	0.657±0.035	0.874±0.006
TransMIL [18]	0.951±0.024	0.968±0.028	0.656±0.021	0.892±0.014	0.573±0.019	0.849±0.010
DSMIL [11]	0.895±0.038	0.949±0.017	0.582±0.062	0.887±0.013	0.586±0.024	0.848±0.010
IBMIL [12]	0.935±0.014	0.953±0.009	0.629±0.027	0.884±0.016	0.640±0.010	0.867±0.007
MHIM-MIL [23]	0.946±0.033	0.984±0.016	0.594±0.090	0.912±0.009	0.660±0.030	**0.890±0.007**
ILRA [25]	0.929±0.018	0.963±0.019	0.662±0.048	**0.914±0.017**	0.626±0.028	0.864±0.014
ABMIL [9]	0.953±0.018	0.972±0.010	0.610±0.025	0.864±0.017	0.621±0.023	0.853±0.013
AEM(ours)	**0.967±0.025**	**0.988±0.013**	**0.688±0.016**	0.905±0.005	**0.691±0.032**	0.884±0.010
SSL pretrained ViT-L (UNI [3])						
ABMIL [9]	0.968±0.011	0.996±0.003	0.605±0.047	**0.885±0.015**	0.580±0.023	0.844±0.024
AEM(ours)	**0.975±0.003**	**0.998±0.003**	**0.633±0.024**	0.863±0.017	**0.645±0.021**	**0.870±0.015**
SSL pretrained ViT-G (GigaPath [27])						
ABMIL [9]	0.978±0.007	**0.984±0.009**	0.555±0.040	0.880±0.023	0.623±0.023	0.866±0.014
AEM(ours)	**0.981±0.009**	0.982±0.011	**0.571±0.029**	**0.886±0.014**	**0.663±0.017**	**0.903±0.014**
VLM pretrained ViT-B (CONCH [14])						
ABMIL [9]	0.932±0.015	0.952±0.017	0.529±0.022	0.862±0.014	0.589±0.036	0.849±0.023
AEM(ours)	**0.942±0.011**	**0.961±0.016**	**0.581±0.013**	**0.893±0.010**	**0.656±0.022**	**0.889±0.011**

clusters. Unlike C2C's strict uniform enforcement, AEM's negative entropy approach provides flexibility by penalizing extreme concentration while allowing meaningful non-uniform distributions when appropriate [8]. Our experiments confirm that replacing AEM with KL-divergence decreases performance.

3 Experiments

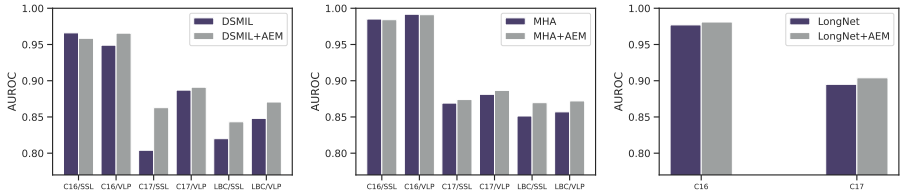
3.1 Experimental setup

Datasets. We evaluate AEM on three WSI datasets: CAMELYON16 (C16) [2], CAMELYON17 (C17) [2], and LBC. C16 contains 270 training WSIs from hospital 1 (split 9:1 for training/validation) and 130 testing WSIs from hospital 2. For C17, we use 500 WSIs in total, with 300 WSIs from three hospitals for training/validation (split 9:1) and 200 WSIs from two other hospitals for testing to evaluate OOD performance. The LBC dataset includes 1,989 WSIs of cervical cancer across four cytological categories: Negative, ASC-US, LSIL, and ASC-H/HSIL, split into 6:2:2 ratios for training, validation, and testing respectively.
Implementation Details. Following [15], we process WSIs by extracting 256 × 256 patches at ×20 magnification. The model architecture consists of a feature dimension reduction layer, gated attention network, and prediction layer, optimized using Adam with cosine learning rate decay. Hyperparameter selection was



(a) Subsampling (b) DTFD-MIL (c) ACMIL

Fig. 3: Performance comparison before and after plugging AEM into the Subsampling augmentation (a) and two advanced MIL frameworks, DTFD-MIL (b) and ACMIL (c). C17/SSL indicates results on C17 using an SSL-pretrained backbone. *AEM improves their performance on 17 out of 18 terms.*



(a) DSMIL (b) MHA (c) LongNet

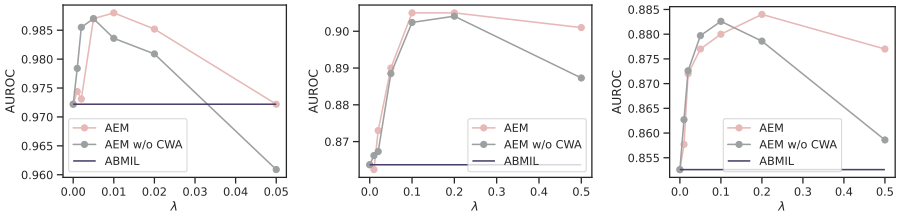
Fig. 4: Performance comparison before and after plugging AEM into DSMIL, MHA, and LongNet attention mechanisms. For LongNet, we used the pretrained Gigapath checkpoint. *AEM improves performance on 11 out of 14 terms.*

based on validation performance optimization, with default λ values of 0.001, 0.1, and 0.2 for C16, C17, and LBC respectively. We report macro-AUC and macro-F1 scores averaged over five runs with different random initializations.

3.2 Main results

AEM's effectiveness across different feature extractors. Table 2 evaluates MIL approaches across three datasets using five backbones. For Lunit-pretrained ViT-S and PathGen-CLIP-pretrained ViT-L, we compare AEM against several advanced MIL methods, with our approach achieving superior performance in 10 out of 12 metrics. With ViT-S, AEM leads across all metrics, while with ViT-L, it dominates in all F1-scores and C16 AUC, with only slight trails in C17 and LBC AUC. For the remaining three backbones (UNI, GigaPath, CONCH), AEM outperforms ABMIL in 16 out of 18 metrics, demonstrating significant improvements across diverse architectures and pretraining strategies. These consistent results confirm AEM's effectiveness as a versatile enhancement applicable to various feature extractors.

AEM enhances Subsampling, DTFD-MIL, and ACMIL. Figure 3a demonstrates AEM's ability to consistently boost performance across multiple MIL



(a) CAMELYON16 (b) CAMELYON17 (c) LBC

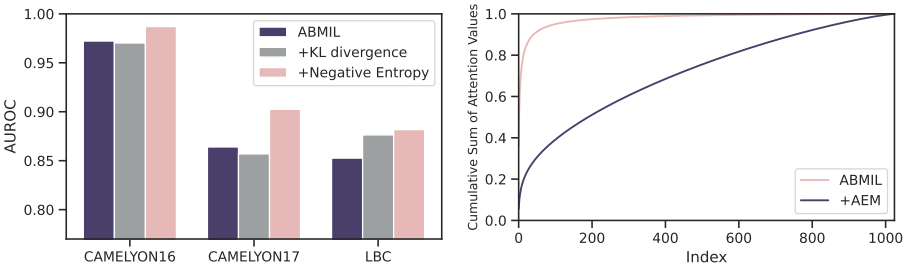
Fig. 5: Sensitivity analysis for hyperparameter λ. *Choosing an appropriate λ is critical for AEM. Moreover, including CWA can improve the stability of AEM.*

frameworks. While Subsampling, DTFD-MIL, and ACMIL all show improvements over standard ABMIL, integrating AEM further elevates their performance. With Subsampling, AEM delivers additional gains, especially in cases where subsampling alone had limited impact. For DTFD-MIL, AEM contributes 2% AUC improvements on C17 and LBC datasets across all backbones. Even when paired with ACMIL, which addresses similar attention concentration issues, AEM still provides notable enhancements on C16 and C17 datasets while maintaining comparable performance on LBC. These consistent improvements across different methods highlight AEM's versatility as a complementary enhancement for diverse MIL approaches.

Performance gains of AEM across different attention mechanisms. To validate AEM's versatility beyond gated attention, we applied it to three additional mechanisms: DSMIL [11], MHA [32], and LongNet [27]. Figure 4 shows AEM's impact across datasets and feature extraction methods. For DSMIL, improvements are most significant with VLM features on C17/SSL and LBC datasets, though C16/SSL performs slightly better without AEM. MHA shows more modest benefits, particularly on C17 and LBC datasets, likely due to its inherent capacity for learning diverse attention values [11,32]. With LongNet using pretrained Gigapath [27] checkpoints, AEM consistently improves finetuning results on both CAMELYON datasets. AEM's effectiveness varies by context, showing particular promise with DSMIL+VLM features, complex datasets, and LongNet architectures.

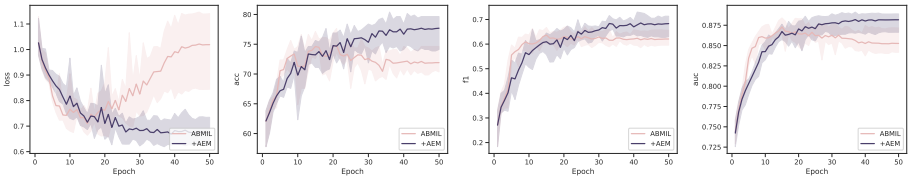
3.3 Further analysis

Ablation Study. We examined the role of λ across three datasets, testing values $\{0, 0.001, 0.002, 0.005, 0.01, 0.02, 0.05\}$ on C16, and $\{0, 0.01, 0.02, 0.05, 0.1, 0.2, 0.5\}$ on C17 and LBC, with $\lambda = 0$ representing baseline ABMIL. Figure 5 reveals that: 1) optimal λ values are approximately 0.01 for C16 and 0.2 for C17/LBC; 2) CWA substantially improves stability, especially at higher λ values where AEM without CWA degrades; 3) both AEM variants outperform the ABMIL baseline; and 4) CWA enables effective operation at larger λ values by adaptively modulating attention weights during training.



(a) Performance comparison of negative entropy vs. KL divergence as AEM loss formulation. Negative entropy shows greater stability and consistent gains.

(b) Mean cumulative sum of top-1000 attention values for LBC test set. AEM effectively mitigates attention over-concentration.



(a) Test Loss (b) Test Accuracy (c) Test F1-score (d) Test AUROC

Fig. 7: Performance comparison between ABMIL [9] and our AEM on LBC test set throughout training. *ABMIL shows clear overfitting with increasing test loss and declining metrics, while AEM effectively prevents this issue.*

Superiority of Negative Entropy over KL Divergence. Comparing loss formulations for alleviating attention concentration, Figure 6a shows AUROC results across three datasets using VLM pretrained embeddings. The negative entropy consistently outperformed both ABMIL baseline and KL divergence approaches. While KL divergence improved results on LBC, it degraded performance on CAMELYON datasets. Negative entropy provides more consistent and stable improvements, making it the preferred formulation for AEM.

AEM effectively mitigates the overfitting. Figure 7 reveals that AEM maintains lower test loss, higher accuracy, and superior F1-score and AUROC compared to ABMIL across training epochs, with ABMIL showing signs of overfitting after epoch 20-30. AEM's consistent outperformance across all metrics demonstrates its superior generalization ability and robustness, establishing it as a more reliable approach less susceptible to overfitting than ABMIL.

AEM effectively mitigates the attention concentration. Figure 6b demonstrates how AEM effectively mitigates the attention concentration problem observed in ABMIL for the LBC test set. The ABMIL curve (purple) rises sharply, indicating that it focuses most of its attention on a small subset of patches. In contrast, the AEM curve (brown) shows a much more gradual increase, suggesting a more balanced distribution of attention across a larger number of patches.

4 Conclusion

This paper introduces AEM, a novel approach addressing attention concentration and overfitting in MIL frameworks through negative entropy regularization of instance attention distributions. AEM effectively mitigates these issues while offering advantages in simplicity—requiring no additional modules or processing. Our experiments demonstrate AEM enhances performance when combined with various MIL frameworks, attention mechanisms, and feature extractors, positioning it as a versatile enhancement for medical image analysis.

Limitation and future work. While currently focused on WSI classification, future work will extend AEM to survival and mutation prediction tasks. Though we introduced cosine weight annealing to stabilize training, the initial weight parameter still requires manual tuning. Future research will develop automatic weight adjustment mechanisms and investigate the theoretical bounds of entropy-based attention regularization.

Acknowledgements. This study was partially supported by the National Natural Science Foundation of China (Grant No. 92270108), Zhejiang Provincial Natural Science Foundation of China (Grant No. XHD23F0201), and the Research Center for Industries of the Future (RCIF) at Westlake University.

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

References

1. Bai, Y., Zhang, B., Zhang, Z., Yan, S., Ma, Z., Liu, W., Zhou, X., Gong, X., Wang, W.: Norma: A noise robust memory-augmented framework for whole slide image classification. In: ECCV. pp. 420–437. Springer (2025)
2. 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.: Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. Jama **318**(22), 2199–2210 (2017)
3. Chen, R.J., Ding, T., Lu, M.Y., Williamson, D.F., Jaume, G., Song, A.H., Chen, B., Zhang, A., Shao, D., Shaban, M., et al.: Towards a general-purpose foundation model for computational pathology. Nature Medicine **30**(3), 850–862 (2024)
4. Dehaene, O., Camara, A., Moindrot, O., de Lavergne, A., Courtiol, P.: Self-supervision closes the gap between weak and strong supervision in histology. arXiv preprint arXiv:2012.03583 (2020)
5. Dehghani, M., Djolonga, J., Mustafa, B., Padlewski, P., Heek, J., Gilmer, J., Steiner, A.P., Caron, M., Geirhos, R., Alabdulmohsin, I., et al.: Scaling vision transformers to 22 billion parameters. In: ICML. pp. 7480–7512. PMLR (2023)
6. Ding, J., Ma, S., Dong, L., Zhang, X., Huang, S., Wang, W., Zheng, N., Wei, F.: Longnet: Scaling transformers to 1,000,000,000 tokens. arXiv preprint arXiv:2307.02486 (2023)
7. Guan, Y., Zhang, J., Tian, K., Yang, S., Dong, P., Xiang, J., Yang, W., Huang, J., Zhang, Y., Han, X.: Node-aligned graph convolutional network for whole-slide image representation and classification. In: CVPR. pp. 18813–18823 (2022)
8. Han, S., Sung, Y.: Diversity actor-critic: Sample-aware entropy regularization for sample-efficient exploration. In: ICML. pp. 4018–4029. PMLR (2021)

9. Ilse, M., Tomczak, J., Welling, M.: Attention-based deep multiple instance learning. In: ICML. pp. 2127–2136. PMLR (2018)
10. Kang, M., Song, H., Park, S., Yoo, D., Pereira, S.: Benchmarking self-supervised learning on diverse pathology datasets. In: CVPR. pp. 3344–3354 (2023)
11. Li, B., Li, Y., Eliceiri, K.W.: Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In: CVPR. pp. 14318–14328 (2021)
12. Lin, T., Yu, Z., Hu, H., Xu, Y., Chen, C.W.: Interventional bag multi-instance learning on whole-slide pathological images. In: CVPR. pp. 19830–19839 (2023)
13. Loshchilov, I.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
14. Lu, M.Y., Chen, B., Williamson, D.F., Chen, R.J., Liang, I., Ding, T., Jaume, G., Odintsov, I., Le, L.P., Gerber, G., et al.: A visual-language foundation model for computational pathology. Nature Medicine **30**(3), 863–874 (2024)
15. 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)
16. Qu, L., Wang, M., Song, Z., et al.: Bi-directional weakly supervised knowledge distillation for whole slide image classification. Neurips **35**, 15368–15381 (2022)
17. Shannon, C.E.: A mathematical theory of communication. The Bell system technical journal **27**(3), 379–423 (1948)
18. 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. Neurips **34**, 2136–2147 (2021)
19. Sharma, Y., Shrivastava, A., Ehsan, L., Moskaluk, C.A., Syed, S., Brown, D.: Cluster-to-conquer: A framework for end-to-end multi-instance learning for whole slide image classification. In: MIDL. pp. 682–698. PMLR (2021)
20. Shi, X., Xing, F., Xie, Y., Zhang, Z., Cui, L., Yang, L.: Loss-based attention for deep multiple instance learning. In: AAAI. vol. 34, pp. 5742–5749 (2020)
21. Song, A.H., Jaume, G., Williamson, D.F., Lu, M.Y., Vaidya, A., Miller, T.R., Mahmood, F.: Artificial intelligence for digital and computational pathology. Nature Reviews Bioengineering **1**(12), 930–949 (2023)
22. Sun, Y., Zhang, Y., Si, Y., Zhu, C., Shui, Z., Zhang, K., Li, J., Lyu, X., Lin, T., Yang, L.: Pathgen-1.6m: 1.6 million pathology image-text pairs generation through multi-agent collaboration (2024), `https://arxiv.org/abs/2407.00203`
23. Tang, W., Huang, S., Zhang, X., Zhou, F., Zhang, Y., Liu, B.: Multiple instance learning framework with masked hard instance mining for whole slide image classification. arXiv preprint arXiv:2307.15254 (2023)
24. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. NeurIPS **30** (2017)
25. Xiang, J., Zhang, J.: Exploring low-rank property in multiple instance learning for whole slide image classification. In: ICLR (2022)
26. Xiong, R., Yang, Y., He, D., Zheng, K., Zheng, S., Xing, C., Zhang, H., Lan, Y., Wang, L., Liu, T.: On layer normalization in the transformer architecture. In: ICML. pp. 10524–10533. PMLR (2020)
27. Xu, H., Usuyama, N., Bagga, J., Zhang, S., Rao, R., Naumann, T., Wong, C., Gero, Z., González, J., Gu, Y., et al.: A whole-slide foundation model for digital pathology from real-world data. Nature pp. 1–8 (2024)
28. 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. MIA **65**, 101789 (2020)

29. Yufei, C., Liu, Z., Liu, X., Liu, X., Wang, C., Kuo, T.W., Xue, C.J., Chan, A.B.: Bayes-mil: A new probabilistic perspective on attention-based multiple instance learning for whole slide images. In: ICLR (2022)
30. Zhai, S., Likhomanenko, T., Littwin, E., Busbridge, D., Ramapuram, J., Zhang, Y., Gu, J., Susskind, J.M.: Stabilizing transformer training by preventing attention entropy collapse. In: ICML. pp. 40770–40803. PMLR (2023)
31. Zhang, H., Meng, Y., Zhao, Y., Qiao, Y., Yang, X., Coupland, S.E., Zheng, Y.: Dtfd-mil: Double-tier feature distillation multiple instance learning for histopathology whole slide image classification. In: CVPR. pp. 18802–18812 (2022)
32. Zhang, Y., Li, H., Sun, Y., Zheng, S., Zhu, C., Yang, L.: Attention-challenging multiple instance learning for whole slide image classification. ECCV (2024)
33. Zhang, Y., Sun, Y., Li, H., Zheng, S., Zhu, C., Yang, L.: Benchmarking the robustness of deep neural networks to common corruptions in digital pathology. In: MICCAI. pp. 242–252. Springer (2022)