GFR-MIL: Glance - Focus - Reflect Based Multiple Instance Learning for Multi-Scale Whole Slide Image Analysis

Mingxin Qi[1,2], Hongyu Li[1,2], Yinuo Lu[1,2], Wei Mu[1,2]✉

[1] Key Laboratory of Precision Diagnosis and Treatment Technology and Intelligent Imaging Equipment, School of Engineering Medicine, Beihang University, 100191, China
[2] Qingdao Research Institute, Beihang University, Qingdao, China

Abstract. Multiple Instance Learning (MIL) has emerged as the dominant paradigm for computational Whole Slide Images (WSIs) analysis. To capture richer diagnostic cues while maintaining computational efficiency, recent MIL methods have adopted coarse-to-fine sampling strategies that use low-magnification overviews to guide high-resolution patch selection. However, the unidirectional nature of such strategies prevents the model from correcting suboptimal region selection, leading to irreversible loss of critical diagnostic information. To address these challenges, we propose GFR-MIL, a novel Glance-Focus-Reflect framework where the Reflect module provides top-down feedback from high-resolution fine-grained features to recalibrate low-resolution global representations, establishing bidirectional cross-scale interaction for MIL-based WSI analysis. This iterative closed-loop mechanism emulates the cognitive workflow of experienced pathologists, who rapidly glance at the global tissue context, focus on suspicious regions, and reflectively revise initial judgments. To further address information loss from unsampled regions, we introduce a Recycling Mechanism and a memory masking strategy, ensuring comprehensive slide-level supervision throughout training. Extensive experiments across 7 public cohorts demonstrate that GFR-MIL consistently outperforms state-of-the-art methods on both cancer subtyping and prognosis prediction tasks, with thorough ablation studies validating the contribution of each proposed component.

Keywords: Multiple Instance Learning, Multi-scale Zooming, Bidirectional Interaction.

1 Introduction

Histopathological assessment is the gold standard for cancer diagnosis and prognosis. The advent of digital pathology has positioned Artificial Intelligence as a transformative tool for analyzing WSIs [1,2]. However, given the gigapixel resolution of WSIs and the scarcity of pixel-level annotations, weakly supervised Multiple Instance Learning has emerged as the dominant paradigm [3-5].

✉ Corresponding author: weimu@buaa.edu.cn

To capture the rich diagnostic information embedded across different resolutions, various multi-scale MIL approaches have been developed. While some methods integrate multi-resolution features through representation fusion [6-10], their computational cost remains prohibitive for gigapixel WSIs. A more efficient alternative, coarse-to-fine sampling methods [11-13] adopt a cascade paradigm. They first employ a low-resolution discriminator to scan the global context and pinpoint key suspicious regions. Guided by these location cues, the model then performs a meticulous diagnosis by selectively extracting and analyzing fine-grained features from the corresponding high-resolution areas. This strategy effectively filters out non-diagnostic background, substantially reducing redundant computation while focusing on clinically relevant details.

However, most existing implementations adopt a unidirectional information flow, in which region selection is determined solely at the coarse stage without any feedback from subsequent fine-grained analysis. Once the coarse-level representation overlooks a diagnostically critical region, that region is permanently excluded from further consideration, as high-resolution fine-grained evidence cannot refine or correct the initial region selection. In addition, many methods employ static Top-K sampling strategies, which further exacerbate this issue. Such mechanisms repeatedly attend to dominant regions while neglecting clinically significant secondary lesions, with information from unsampled regions entirely discarded.

The cognitive process of an experienced pathologist, however, is far from unidirectional. Rather, it constitutes a dynamic iterative loop: rapidly glancing to localize suspicious areas, zooming in to examine micro-environmental details, and reflecting on the global context to identify potentially overlooked lesions. Drawing inspiration from this workflow, we propose GFR-MIL, a novel framework that emulates the Glance-Focus-Reflect cognitive paradigm and establishes bidirectional closed-loop interaction between low-resolution global context and high-resolution local fine-grained details in MIL-based WSI analysis. The main contributions of this paper are as follows:

1. We propose a novel GFRBlock comprising three collaborative modules, namely Glance, Focus, and Reflect, where the Reflect module establishes top-down feedback from high-resolution local features to low-resolution global representations, enabling subsequent blocks to adaptively reselect diagnostically relevant regions.
2. We introduce a Recycling Mechanism and a memory masking strategy, effectively addressing the issues of narrow field-of-view and information loss inherent in traditional hard attention mechanisms, enabling more comprehensive and dynamic region exploration.
3. Extensive experiments across seven large-scale public cohorts demonstrate that GFR-MIL consistently outperforms state-of-the-art methods on both cancer prognosis prediction and cancer subtyping tasks, with thorough ablation studies validating the contribution of each proposed component. The source code is available at https://github.com/longger-z/GFR-MIL.

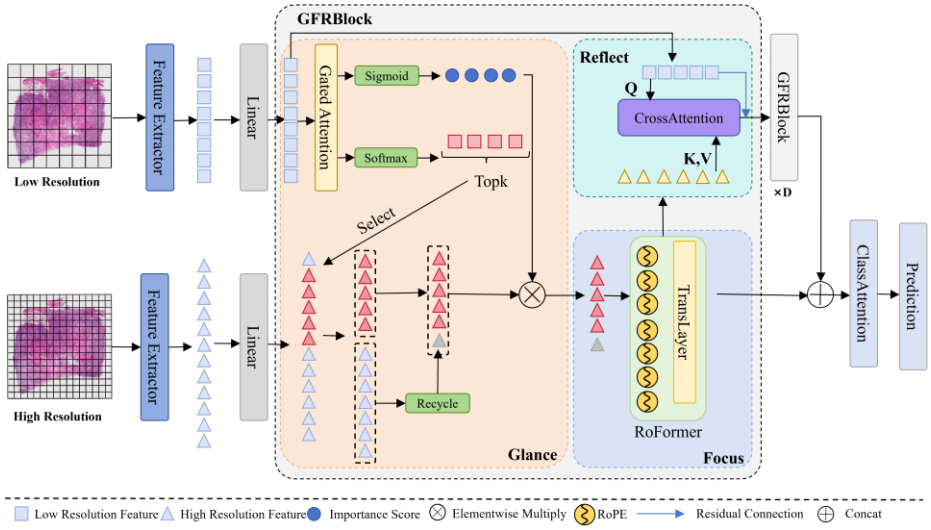

Fig. 1. Overview architecture of our proposed GFR-MIL.

2 Methodology

2.1 Overall Architecture

The overview of the proposed GFR-MIL framework is depicted in Fig. 1. Given a WSI, patches are cropped at two resolution levels and encoded by the pretrained pathology foundation model, yielding low-magnification features and high-magnification features. These dual-scale features are then fed into a cascade of stacked GFRBlocks. Each block employs a Glance-Focus-Reflect workflow comprising three collaborative modules: Glance, Focus, and Reflect, mirroring the diagnostic reasoning process of experienced pathologists. Through iterative information passing across blocks, GFR-MIL achieves progressive bidirectional interaction between global context and local details. The Reflect module is omitted in the final block.

2.2 Glance Module

We extract features from the Whole Slide Image (WSI) at two resolution levels, obtaining a set of instance features at low magnification, $\mathcal{X}^{L} = \{x_j^{L} \in \mathbb{R}^{d}\}_{j=1}^{N_L}$, and a set of instance features at high magnification, $\mathcal{X}^{H} = \{x_i^{H} \in \mathbb{R}^{d}\}_{i=1}^{N_H}$. To enable cross-scale information interaction, we define a mapping function $\mathcal{M}: \{1, \ldots, N_H\} \rightarrow \{1, \ldots, N_L\}$, indicating that the high-resolution patch i spatially belongs to the low-resolution region $\mathcal{M}(i)$.

The core of the Glance Module is a Gated Attention mechanism designed to output two distinct values simultaneously. A gating network maps each low-resolution regional feature x_j^L to a latent space vector s_j. Through Eq. (1), we derive the sampling probability distribution $\mathcal{A} = \{\alpha_j\}_{j=1}^{N_L}$ and an importance score c_j for each instance:

$$\alpha_j = \frac{\exp(s_j)}{\sum_{k=1}^{N_L} \exp(s_k)}, \quad c_j = \sigma(s_j) \tag{1}$$

where σ denotes the Sigmoid function.

Sampling Strategy During training, to enhance the model's exploratory capability, we sample k% of the patches based on the distribution $\mathcal{A}$ using multinomial sampling without replacement to obtain the index set $\mathcal{I}_{sel}$. During testing, to ensure stability, we directly select the top-k% instances. To prevent the attention mechanism from collapsing or degenerating into a uniform distribution, we introduce an entropy regularization term L_{ent} [17] as a constraint, as shown in Eq. (2):

$$L_{ent} = -\sum_i^{N_L} \alpha_i \log \alpha_i \tag{2}$$

The selected high-resolution features x_i^H (where $\mathcal{M}(i) \in \mathcal{I}_{sel}$) are explicitly multiplied by the importance coefficient of their corresponding region before entering the Focus Module:

$$\tilde{\mathbf{x}}_i^H = c_{\mathcal{M}(i)} \cdot \mathbf{x}_i^H \tag{3}$$

This design not only propagates information from low-resolution regions to the subsequent Focus Module but also allows the model to optimize the sampling strategy by refining regional importance scores during training.

Recycling Mechanism Following sampling, unselected regions may still contain critical diagnostic information. To mitigate information loss, we design a recycling mechanism that aggregates high-resolution features from unselected regions (denoted by the index set $\mathcal{I}_{rem}$). We compute a global recycling token z_{rec}^H, which is processed in the Focus Module alongside the selected instances:

$$\mathbf{z}_{rec}^H = \frac{\sum_{j \in \mathcal{I}_{rem}} \sum_{k \in \mathcal{P}_j} c_j \mathbf{x}_k^H}{\sum_{j \in \mathcal{I}_{rem}} |\mathcal{P}_j|} \tag{4}$$

where $\mathcal{P}_j$ represents the set of patches within the j-th region.

Applied solely during the training phase, the recycling mechanism ensures that gradients are backpropagated to the Glance Module even for unselected regions, thereby facilitating the learning of a more effective sampling strategy.

2.3 Focus Module

Subsequently, the selected instance features $\tilde{x}_i^H$ and the global recycling token z_{rec}^H are fed into the Focus Module. The Focus Module is designed to model the relationships between selected regions. Given the significance of spatial relationships within WSIs, we employ a RoFormer layer [15] as the Focus Module. The positional information

coord_i^H for the selected instance features $\tilde{x}_i^H$ is naturally obtained during preprocessing, while the coordinates for the global recycling token z_{rec}^H are calculated using Eq. (5):

$$coord_{rec}^H = \frac{\sum_{j \in \mathcal{I}_{rem}} \sum_{k \in \mathcal{P}_j} c_j \mathbf{coord}_k^H}{\sum_{j \in \mathcal{I}_{rem}} |\mathcal{P}_j|} \tag{5}$$

The RoFormer modeling is defined in Eq. (6):

$$features^H = RoFormer(\widetilde{\mathcal{X}}^H, Coord^H) \tag{6}$$

2.4 Reflect Module

Since a single observation pass may result in omissions, we mimic the pathologist's behavior by introducing a feedback mechanism. The output of the Focus Module from the current round, $features^H$, interacts with $\mathcal{X}^L$ via a Cross-Attention mechanism. This process injects high-resolution, fine-grained features back into the low-resolution feature space to update $\mathcal{X}^L$.

$$\mathcal{X}^{(1)L} = \mathcal{X}^L + \mathrm{LN}(CrossAttention(\mathcal{X}^L, features^H)) \tag{7}$$

This process allows the model to dynamically adjust its attention to global regions based on confirmed pathological micro-environmental details. Subsequently, $\mathcal{X}^{(1)\mathrm{L}}$ and $\mathcal{X}^{\mathrm{H}}$ are forwarded to the next layer to initiate a new iteration.

Simultaneously, to prevent the model from redundantly focusing on the same area, we employ a memory constraint strategy. We maintain a history mask matrix to explicitly suppress regions selected in previous rounds. This forces the model to shift its attention to secondary but clinically significant regions in subsequent iterations, thereby constructing a more comprehensive slide-level representation.

2.5 Aggregation and Loss Function

After D iterations, the model accumulates multiple sets of high-resolution feature sequences. We concatenate the features from all rounds and fuse them using a Class Attention-based aggregation layer. The final slide-level representation z_{slide} is aggregated by ClassAttention [16] and fed into a classifier for prediction.

$$\mathrm{F^H} = \mathrm{concat}[\mathrm{features}^{(1)\mathrm{H}},\ \mathrm{features}^{(2)\mathrm{H}}, \ldots\ \mathrm{features}^{(\mathrm{D})\mathrm{H}}] \tag{8}$$

$$\mathrm{Z_{slide}} = \mathrm{ClassAttention}(\mathrm{F^H}) \tag{9}$$

The total loss function of the model comprises a task-specific loss $\mathrm{L_{task}}$ and the multi-round accumulated entropy regularization term:

$$L_{total} = L_{task} + \frac{\lambda}{D} \sum_{t=1}^{D} L_{ent}^{(t)} \tag{10}$$

This composite loss function optimizes prediction accuracy while ensuring the diversity and robustness of the attention distribution. λ and D are hyperparameters determined based on the specific needs of the task.

Table 1. Survival prediction results on 5 TCGA datasets. The best results are in bold, suboptimal results are underlined.

Model	KIRC	COAD	LUAD	STAD	BLCA
ABMIL[3]	<u>0.7441$_{\pm 0.0317}$</u>	0.6929$_{\pm 0.0599}$	0.6060$_{\pm 0.0674}$	0.6010$_{\pm 0.0627}$	0.6511$_{\pm 0.0370}$
ABMIL-Gated[3]	0.7367$_{\pm 0.0382}$	0.6885$_{\pm 0.0759}$	0.5978$_{\pm 0.0703}$	0.6050$_{\pm 0.0578}$	<u>0.6590$_{\pm 0.0560}$</u>
DSMIL[6]	0.7199$_{\pm 0.0459}$	0.6095$_{\pm 0.0563}$	0.6053$_{\pm 0.0678}$	0.5684$_{\pm 0.0628}$	0.5957$_{\pm 0.0693}$
TransMIL[4]	0.7224$_{\pm 0.0602}$	0.6942$_{\pm 0.0518}$	0.5994$_{\pm 0.0622}$	0.6270$_{\pm 0.0640}$	0.6557$_{\pm 0.0356}$
ILRA[5]	0.7248$_{\pm 0.0568}$	0.7035$_{\pm 0.0499}$	0.6007$_{\pm 0.0478}$	0.6258$_{\pm 0.0567}$	0.6144$_{\pm 0.0435}$
RRT[18]	0.7309$_{\pm 0.0526}$	0.6951$_{\pm 0.0612}$	0.6137$_{\pm 0.0677}$	0.6253$_{\pm 0.0436}$	0.6213$_{\pm 0.0596}$
MambaMIL[19]	0.7303$_{\pm 0.0354}$	<u>0.7081$_{\pm 0.0356}$</u>	<u>0.6444$_{\pm 0.0758}$</u>	0.5820$_{\pm 0.0267}$	0.6494$_{\pm 0.0483}$
PTCMIL[20]	0.7354$_{\pm 0.0596}$	0.6988$_{\pm 0.0437}$	0.6070$_{\pm 0.0838}$	<u>0.6270$_{\pm 0.0530}$</u>	0.6334$_{\pm 0.0407}$
GFRMIL	**0.7453$_{\pm 0.0184}$**	**0.7134$_{\pm 0.0546}$**	**0.6581$_{\pm 0.0639}$**	**0.6360$_{\pm 0.0526}$**	**0.6696$_{\pm 0.0720}$**

3 Experiments

3.1 Tasks and Datasets

To comprehensively validate the effectiveness of GFR-MIL, we conducted extensive experiments across 7 cohorts covering two core downstream clinical tasks.

Cancer Subtyping We tested our model on the TCGA-NSCLC (2 classes) [22] and BRACS (3 classes) [23] datasets. For TCGA-NSCLC, we employed 5-fold cross-validation. For BRACS, we strictly adhered to the official splits.

Prognosis Prediction We conducted experiments on five challenging cohorts: TCGA-KIRC, TCGA-LUAD, TCGA-COAD, TCGA-STAD, and TCGA-BLCA. We used Disease-Specific Survival (DSS) [24] as the ground truth and reported the Concordance Index (C-index) as the performance metric. All prognosis experiments utilized 5-fold cross-validation.

3.2 Implementation Details

We followed the standardized pipeline of CLAM [2], segmenting slides into 224×224 patches at 20× magnification. We used CONCH [21] for feature extraction. For the low-resolution branch, we synchronously performed preprocessing of the same specifications at 2.5× magnification and established cross-resolution index correspondence using the anatomical mapping function $\mathcal{M}$.

Experiments were implemented using the PyTorch framework on an NVIDIA RTX 4090 GPU. We used the Adam optimizer with an initial learning rate of 1e-4 coupled with gradient accumulation with a factor of 4. For classification tasks, Cross-Entropy (CE) loss was used; for survival analysis, we minimized the Negative Log-Likelihood (NLL) loss. During experiments, D was set to 2, λ to 0.1, and the sampling ratios per layer were set to 0.2 and 0.1, respectively.

Table 2. Subtype prediction results on TCGA-NSCLC and BRACS. The best performances are in bold, second best performances are underlined.

Model	TCGA-NSCLC			BRACS		
	AUC	ACC	F1	AUC	ACC	F1
ABMIL[3]	$0.9795_{\pm 0.0028}$	$0.9346_{\pm 0.0107}$	$0.9343_{\pm 0.0109}$	0.8788	0.6744	0.6076
ABMIL-Gated[3]	$\underline{0.9803_{\pm 0.0044}}$	$0.9373_{\pm 0.0184}$	$\underline{0.9371_{\pm 0.0185}}$	<u>0.9019</u>	0.7326	0.7001
DSMIL[6]	$0.9747_{\pm 0.0080}$	$0.9232_{\pm 0.0127}$	$0.9229_{\pm 0.0125}$	0.8532	0.6628	0.6162
TransMIL[4]	$0.9774_{\pm 0.0063}$	$0.9314_{\pm 0.0199}$	$0.9312_{\pm 0.0199}$	0.7867	0.6395	0.5005
ILRA[5]	$0.9738_{\pm 0.0076}$	$0.9237_{\pm 0.0084}$	$0.9235_{\pm 0.0085}$	0.8400	0.6628	0.6274
RRT[18]	$0.9787_{\pm 0.0066}$	$0.9275_{\pm 0.0167}$	$0.9271_{\pm 0.0169}$	0.8934	<u>0.7442</u>	<u>0.7388</u>
MambaMIL[19]	$0.9749_{\pm 0.0090}$	$0.9240_{\pm 0.0201}$	$0.9238_{\pm 0.0201}$	0.8968	0.6744	0.6193
PTCMIL[20]	$0.9789_{\pm 0.0041}$	$\underline{0.9385_{\pm 0.0084}}$	$0.9355_{\pm 0.0113}$	0.9013	0.7209	0.7236
GFRMIL	$\mathbf{0.9805_{\pm 0.0064}}$	$\mathbf{0.9394_{\pm 0.0099}}$	$\mathbf{0.9373_{\pm 0.0096}}$	**0.9069**	**0.7907**	**0.7754**

Table 3. Ablation study of key components in GFR-MIL on TCGA-LUAD and TCGA-COAD datasets using C-index. R&C refers to recycle mechanism and importance score c.

Glance		Reflect	LUAD	COAD
L_{ent}	R&C			
√	√	×	$0.6208_{\pm 0.0878}$	$\underline{0.7022_{\pm 0.0538}}$
×	√	√	$\underline{0.6423_{\pm 0.0669}}$	$0.6849_{\pm 0.0656}$
√	×	√	$0.6090_{\pm 0.0637}$	$0.6512_{\pm 0.0174}$
√	√	√	$\mathbf{0.6581_{\pm 0.0639}}$	$\mathbf{0.7134_{\pm 0.0546}}$

3.3 Comparison with Baselines

Prognosis Prediction As shown in Table 1, we conducted comparative experiments on 5 TCGA survival analysis cohorts. The results demonstrate that our method (GFR-MIL) achieved the best performance across all datasets.

Cancer Subtyping Table 2 presents the performance on binary (NSCLC) and multi-class (BRACS) classification tasks. GFR-MIL achieved the best results on both.

3.4 Ablation Study

As shown in Table 3, we designed three ablation studies to validate the effectiveness of the proposed module.

Value of the Reflect Module Removing the top-down feedback pathway causes the model to degenerate into a unidirectional selection architecture. The ablation results show a marked decline in performance, proving that a feedback loop based on "reflecting on global context via microscopic details" is crucial for correcting initial sampling biases.

Table 4. Theoretical Memory, GFLOPs, and Typical Inference Time of Different Models.

Model	Memory(GB)	FLOPs(G)	Inference-time(ms)
TransMIL[4]	0.93	189.5	14.5
RRT[18]	2.0	571.7	29.8
GFRMIL	0.55	85.7	11.3

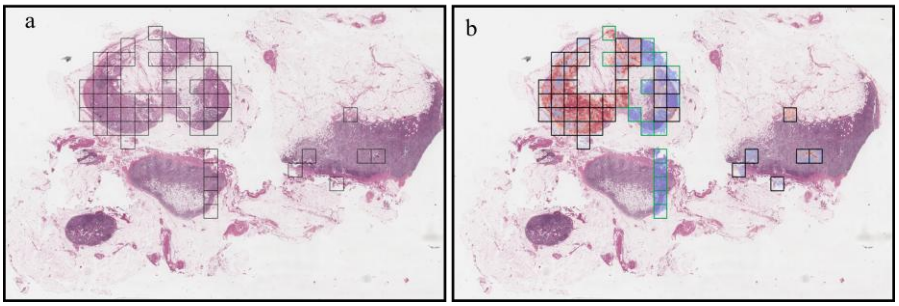

Fig. 2. (a) All regions selected by GFR-MIL. (b) Attention heatmap of GFR-MIL. In (b), black and green bounding boxes denote the regions selected in the first and second rounds, respectively.

Entropy Regularization Term Experiments indicate that without this constraint, the attention mechanism tends to degenerate into a non-informative uniform distribution. The entropy regularization enforces attention sparsity, compelling the model to make sharp, discriminative selections.

Recycling & Importance In the table, R&C represents the Recycling mechanism and Importance Score. To allow for comparison without R&C, we utilized Gumbel-Softmax [14, 25] as the baseline. Experimental results demonstrate that using the product of the recycling mechanism and importance scores to guide gradients not only stabilizes training but also significantly improves performance.

3.5 Computational Complexity Analysis

Table 4 details GFR-MIL's efficiency on TCGA-KIRC (85.7 GFLOPs), outperforming TransMIL (189.5G) and RRT (571.7G). During inference, GFR-MIL extracts features from only ~30% of key regions, reducing WSI processing time from 51.3 s to 30.4 s.

3.6 Visualization and interpretability

As shown in Fig. 2(a), the key regions of interest (ROIs) identified by the GFR-MIL model are indicated by black boxes. The corresponding heatmap in Fig. 2(b) demonstrates that the highlighted red areas accurately encompass the primary lesions (e.g., DCIS). This indicates that the model can effectively localize critical regions to facilitate precise diagnosis.

4 Conclusion

In this paper, we propose GFR-MIL, a multi-scale MIL framework for modeling gigapixel WSIs. By introducing a bidirectional Glance-Focus-Reflect interaction loop, GFR-MIL enables multi-round cross-scale refinement, where the Glance module rapidly localizes suspicious areas, the Focus module scrutinizes micro-environmental details, and the Reflect module identifies overlooked lesions. This design mirrors the diagnostic workflow of experienced pathologists while maintaining algorithmic rigor. Further combination with the Recycling Mechanism and memory masking strategy, the framework ensures comprehensive slide-level supervision while maintaining computational efficiency. Extensive experiments across seven cohorts demonstrate that GFR-MIL achieves state-of-the-art performance while attending to only 30% of the slide regions, significantly balancing efficiency and accuracy. Future work will further explore the potential of the "Glance-Focus-Reflect" paradigm across broader computational pathology tasks.

Acknowledgments. This study was supported by Science, Technology&Innovation Project of Xiongan New Area (2023XAGG0071), National Natural Science Foundation of China (62571019, 92259205), Shandong Provincial Natural Science Foundation (ZR2024QF010), and the fundamental research funds for the central universities.

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

References

1. Saillard, C., Dubois, R., Tchita, O., Loiseau, N., Garcia, T., Adriansen, A., Carpentier, S., Reyre, J., Enea, D., von Loga, K., et al.: Validation of MSIntuit as an AI-based pre-screening tool for MSI detection from colorectal cancer histology slides. Nature Communications **14**, 6695 (2023)
2. Lu, M.Y., Williamson, D.F.K., 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**, 555–570 (2021)
3. Ilse, M., Tomczak, J., Welling, M.: Attention-based Deep Multiple Instance Learning. In: Proceedings of the 35th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 80, pp. 2127–2136. PMLR (2018)
4. Shao, Z., Bian, H., Chen, Y., Wang, Y., Zhang, J., Ji, X., Zhang, Y.: TransMIL: Transformer based Correlated Multiple Instance Learning for Whole Slide Image Classification. In: Advances in Neural Information Processing Systems, vol. 34, pp. 2136–2147. Curran Associates, Inc. (2021)
5. 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 (ICLR) (2023)
6. 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 (CVPR), pp. 14318–14328 (2021)
7. Chen, R.J., Chen, C., Li, Y., Chen, T.Y., Trister, A.D., Krishnan, R.G., Mahmood, F.: Scaling Vision Transformers to Gigapixel Images via Hierarchical Self-Supervised Learning. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 16144–16155 (2022)
8. Deng, R., Cui, C., Remedios, L.W., Bao, S., Womick, R.M., Chiron, S., Li, J., Roland, J.T., Lau, K.S., Liu, Q., et al.: Cross-scale multi-instance learning for pathological image diagnosis. Medical Image Analysis **94**, 103124 (2024)
9. Hou, W., Yu, L., Lin, C., Huang, H., Yu, R., Qin, J., Wang, L.: H^2-MIL: Exploring Hierarchical Representation with Heterogeneous Multiple Instance Learning for Whole Slide Image Analysis. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, pp. 933–941 (2022)
10. Yu, J., Li, S., Tan, L., Zhou, H., Li, Z., Li, J.: HiViT: Hierarchical attention-based Transformer for multi-scale whole slide histopathological image classification. Expert Systems with Applications **277**, 127164 (2025)
11. Xiong, C., Chen, H., Sung, J.J.Y., King, I.: Diagnose Like a Pathologist: Transformer-Enabled Hierarchical Attention-Guided Multiple Instance Learning for Whole Slide Image Classification. In: Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence (IJCAI-23), pp. 1587–1595 (2023)
12. Thandiackal, K., Chen, B., Pati, P., Jaume, G., Williamson, D.F.K., Gabrani, M., Goksel, O.: Differentiable Zooming for Multiple Instance Learning on Whole-Slide Images. In: Computer Vision – ECCV 2022, LNCS, vol. 13681, pp. 699–715. Springer, Cham (2022)
13. Zhao, B., Zhang, J., Ye, D., Cao, J., Han, X., Fu, Q., Yang, W.: RLogist: Fast Observation Strategy on Whole-Slide Images with Deep Reinforcement Learning. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 37, pp. 3570–3578 (2023)
14. Maddison, C.J., Mnih, A., Teh, Y.W.: The Concrete Distribution: A Continuous Relaxation of Discrete Random Variables. In: International Conference on Learning Representations (ICLR) (2017)
15. Su, J., Lu, Y., Pan, S., Murtadha, A., Wen, B., Liu, Y.: RoFormer: Enhanced Transformer with Rotary Position Embedding. Neurocomputing **568**, 127063 (2024)
16. Touvron, H., Cord, M., Sablayrolles, A., Synnaeve, G., Jégou, H.: Going deeper with Image Transformers. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 32–42 (2021)
17. Zhai, S., Likhomanenko, T., Littwin, E., Busbridge, D., Ramapuram, J., Zhang, Y., Gu, J., Susskind, J.: Stabilizing Transformer Training by Preventing Attention Entropy Collapse. In: Proceedings of the 40th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 202, pp. 40770–40803. PMLR (2023)
18. Tang, W., Zhou, F., Huang, S., Zhu, X., Zhang, Y., Liu, B.: Feature Re-Embedding: Towards Foundation Model-Level Performance in Computational Pathology. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 11343–11352 (2024)
19. Yang, S., Wang, Y., Chen, H.: MambaMIL: Enhancing Long Sequence Modeling with Sequence Reordering in Computational Pathology. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2024, LNCS, vol. 15004, pp. 296–306. Springer, Cham (2024)

20. Zhao, B., Kim, S., Chen, H., Zhou, C., Gao, Z.-h., Wang, G., Li, X.: PTCMIL: Multiple Instance Learning via Prompt Token Clustering for Whole Slide Image Analysis. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2025, LNCS, vol. 15974, pp. 502–512. Springer, Cham (2025)
21. Lu, M.Y., Chen, B., Williamson, D.F.K., 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**, 863–874 (2024)
22. Chang, K., Creighton, C.J., Davis, C., Donehower, L., Drummond, J., Wheeler, D., Ally, A., Balasundaram, M., Birol, I., Butterfield, Y.S.N., et al.: The Cancer Genome Atlas Pan-Cancer analysis project. Nature Genetics **45**, 1113–1120 (2013)
23. Brancati, N., Anniciello, A.M., Pati, P., Riccio, D., Scognamiglio, G., Jaume, G., De Pietro, G., Di Bonito, M., Foncubierta, A., Botti, G., et al.: BRACS: A Dataset for BReAst Carcinoma Subtyping in H&E Histology Images. Database **2022**, baac093 (2022)
24. Liu, J., Lichtenberg, T., Hoadley, K.A., Poisson, L.M., Lazar, A.J., Cherniack, A.D., Kovatich, A.J., Benz, C.C., Levine, D.A., Lee, A.V., et al.: An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome Analytics. Cell **173**, 400–416 (2018)
25. Jang, E., Gu, S., Poole, B.: Categorical Reparameterization with Gumbel-Softmax. In: International Conference on Learning Representations (ICLR) (2017)