RetMIL: Retentive Multiple Instance Learning for Histopathological Whole Slide Image Classification

Hongbo Chu[1*], Qiehe Sun[1,3*], Jiawen Li[1,3*], Yuxuan Chen[1], Lizhong Zhang[1,3], Tian Guan[1], Anjia Han[2†], and Yonghong He[1,3†]

[1] Shenzhen International Graduate School, Tsinghua University, China
{zhu-hb23,sunqh21,lijiawen21}@mails.tsinghua.edu.cn
heyh@sz.tsinghua.edu.cn
[2] department of Pathology, The First Affiliated Hospital of Sun Yat-sen University, China
hananjia@mail.sysu.edu.cn
[3] Medical Optical Technology R&D Center, Research Institute of Tsinghua, Pearl River Delta, Guangzhou 510700, China

Abstract. Histopathological whole slide image (WSI) analysis with deep learning has become a research focus in computational pathology. The current paradigm is mainly based on multiple instance learning (MIL), in which approaches with Transformer as the backbone are well discussed. These methods convert WSI tasks into sequence tasks by representing patches as tokens in the WSI sequence. However, the feature complexity brought by high heterogeneity and the ultra-long sequences brought by gigapixel size makes Transformer-based MIL suffer from the challenges of high memory consumption, slow inference speed, and lack of performance. To this end, we propose a retentive MIL method called RetMIL, which processes WSI sequences through hierarchical feature propagation structure. At the local level, the WSI sequence is divided into multiple subsequences. Tokens of each subsequence are updated through a parallel linear retention mechanism and aggregated utilizing an attention layer. At the global level, subsequences are fused into a global sequence, then updated through a serial retention mechanism, and finally the slide-level representation is obtained through a global attention pooling. We conduct experiments on two public CAMELYON and BRACS datasets and an public-internal LUNG dataset, confirming that RetMIL not only achieves state-of-the-art performance but also significantly reduces computational overhead. Our code are available at: https://github.com/Hongbo-Chu/RetMIL

Keywords: Histopathological Whole Slide Image · Multiple Instance Learning · Retention Mechanism.

* Contributed equally.
† Corresponding author.

1 Introduction

Pathological slide scanners store microscopic fields of view as the WSI, laying the foundation for automatic diagnostics based on deep learning [19]. However, the gigapixel-level resolution and the lack of pixel-level annotations pose significant challenges in developing such intelligent tools. In recent years, with the development of weakly-supervised technologies, MIL methods for WSI analysis have been well studied, which treats WSI as bags and cropped patches as instances. By embedding instances into high-dimensional space for aggregation, slide-level representations can be obtained. MIL methods are generally categorized into instance-level [4,7] and embedding-level [12,22] approaches. The former has been gradually replaced due to enormous data requirements and weak generalization.

Embedding-level MIL methods generally focus on proposing effective aggregation strategies to obtain more effective WSI representations. Although mean or max pooling is a direct corollary of the MIL theory, dynamically assigning importance scores to patches has proven more effective [12,17].

As one of the dynamically assigning approach, graph-based WSI analysis methods are currently devoted to modeling through the spatial positional relationships of patches. For example,[10,15], have all demonstrated excellent performance in WSI tasks. Also, in order to model histological and cytological information across different scales, researchers have introduced a multi-scale dynamically assigning approach [8]. This method can explicitly model the histological information at different resolutions, thus improving prediction accuracy.

In addition, due to the wide application of Transformer[25], more research focuses on predicting WSI scores by modeling the correlation between patches through the self-attention mechanism, which helps describe the underlying tumor microenvironment patterns. Transformer-based MIL methods have shown better performance in many WSI analysis tasks[22,6]. However, the square complexity caused by the nonlinear mechanism of self-attention consumes more memory during training and inference, resulting in increased latency and reduced speed, which is not conducive to the actual deployment of algorithms in clinical scenarios.

To alleviate the above challenges, in this paper, we proposed a retentive multiple instance learning neural network called RetMIL, which introduces a retention mechanism to replace nonlinear self-attention, and effectively integrates subsequence information of WSI to obtain global representation with local features by building a hierarchical structure. We conduct experiments on public CAMELYON and BRACS datasets, as well as LUNG dataset for public data training and internal data testing. Results demonstrate that our proposed RetMIL achieves lower memory cost and higher throughput while exhibiting competitive performance.

2 Methodology

In this section, we introduce the methodology of RetMIL. First, WSI is processed into a sequence form. Then local subsequences and the global sequence

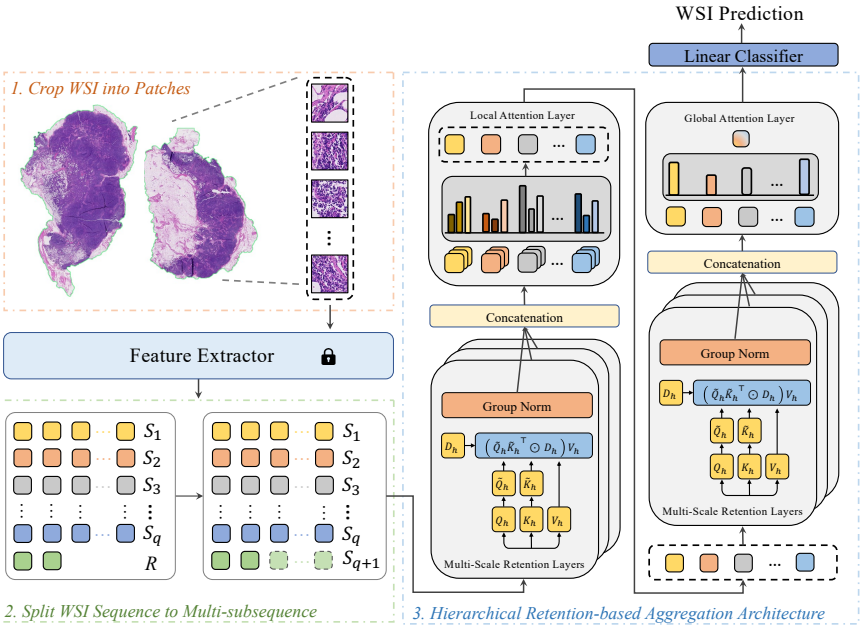


Fig. 1. Overall framework: Tissues in WSI are first cut into patches, each of the patch is encoded into a feature representation. These feature sequences are then composed into multiple subsequences. Next, these combined features are fed into our retention aggregator for feature fusion. Finally, results are output through a linear classifier.

are sequentially updated and aggregated through retention and attention pooling. Finally, the prediction score of WSI is obtained through the classification head. Fig 1 shows the overall framework of RetMIL.

2.1 From WSI to Sequence

We preprocess a WSI in four steps. First, we use the OTSU algorithm [20] to segment the WSI foreground, and then use the sliding window operation to crop patches under a fixed magnification. Secondly, ViT-S/16 [9], which is pretrained based on DINO [5] on large-scale WSIs [13], is used as a feature extractor to encode each patch into a high-dimensional feature embedding $x_i \in \mathbb{R}^{d\times 1}$. Next, we form all x_i into a sequence $X = \{x_1, \ldots, x_N\}$, and split it into multiple subsequences $\{S_1, \ldots, S_q, R\}$. Specifically, let $N = ql + r$, where l is the length of each subsequence $S_j = \{x_{(j-1)l+1}, \ldots, x_{jl}\}$, $j = 1, 2, \ldots, q$ and $r = |R|$. Finally, to ensure that all subsequences have the same length and facilitate parallel calculation, we extend R to $S_{q+1} = Concat(R, X_{l-r})$, where $|X_{l-r}| = l - r$, and there are three situations for X_{l-r} as follows.

– If $r = 0$, then $X_{l-r} = \varnothing$.

– If $0 < r < l/2$, let $l - r = ar + b$ and $A = \{x_{ql+1}, \ldots, x_{ql+r}\}$, then $X_{l-r} = \{\underbrace{A, \ldots, A}_{a \text{ in total}}, x_{ql+1}, \ldots, x_{ql+b}\}$.
– If $r \geq l/2$, then $X_{l-r} = \{x_{ql+1}, \ldots, x_{ql+(l-r)}\}$.

The purpose of processing R in this way is to make each x_i exist in only one subsequence, ensuring that the mapping between feature embeddings and subsequences is satisfied.

2.2 Retention Mechanism

Inspired by applications of the retentive network in large language models [24], RetMIL updates and aggregates sequence tokens through retention mechanisms. Given the matrix form $S \in \mathbb{R}^{|S| \times d}$ of an input sequence, we first use three linear layers to project it into different feature spaces:

$$Q = XW_Q, K = XW_K, V = XW_V, \tag{1}$$

where W_Q, W_K, and W_V are learnable transformation matrices respectively. Next, we split Q, K and V into multiple heads $\{Q_h\}$, $\{K_h\}$ and $\{V_h\}$ and perform rotational position encoding [23] on each Q_h and K_h to obtain $\tilde{Q}_h$ and $\tilde{K}_h$. Then we use the retention layer for processing, which is expressed as follows:

$$Retention(h, X) = (\tilde{Q}_h \tilde{K}_h^\top \odot D_h)V_h, \tag{2}$$

where D_h is a relative distance decay matrix, and each element $D_{h,nm}$ is expressed as:

$$D_{h,nm} = \begin{cases} \gamma^{n-m}, & n \geq m \\ 0, & n < m \end{cases} \tag{3}$$

Finally, we use GroupNorm [26] and swish gate [11,21] to normalize the output, and concatenate all retention head. The above mapping relationship can provide batch-level parallel calculation. When the input batch is B, we denote the entire update operation as $MSR(B; S)$.

2.3 Hierarchical Retentive Aggregation Architecture

For any subsequence matrix S_i,$i \in 1, \ldots, q+1$ in WSI, $MSR(1; S_i)$ represents the result of S_i after passing through the retention mechanism. Our goal is to update all subsequences in parallel, which is expressed as follows:

$$\begin{aligned} (F_1, \ldots, F_{q+1}) =& (MSR(1; S_1), \ldots, MSR(1; S_{q+1})) \\ =& MSR(q+1; (S_1, \ldots, S_{q+1})), \end{aligned} \tag{4}$$

where $F_i \in \mathbb{R}^{l \times d}$ represents the output embedding of subsequence S_i. Next, we use the attention pooling layer to aggregate the element features of each subsequence, which is expressed as:

$$F_{local,i} = \sum_{k=1}^{l} \alpha_{i,k} F_{i,k}, \tag{5}$$

where $F_{i,k}$ represents the kth element of F_i, $F_{local,i} \in \mathbb{R}^{d\times 1}$ represents the feature embedding of subsequence S_i. α_k is calculated through a nonlinear gating mechanism:

$$\alpha_{i,k} = \frac{\exp\{\Gamma_l \tanh(\mathrm{W}_l F_{i,k}) \odot \mathrm{sigm}(\mathrm{U}_l F_{i,k})\}}{\sum_{t=1}^{l} \exp\{\Gamma_l \tanh(\mathrm{W}_l F_{i,t}) \odot \mathrm{sigm}(\mathrm{U}_l F_{i,t})\}}, \tag{6}$$

where $\Gamma_l \in \mathbb{R}^{1\times M}$, $\mathrm{W}_l, \mathrm{U}_l \in \mathbb{R}^{M\times d}$ are learnable parameters, $\tanh(\cdot), \mathrm{sigm}(\cdot)$ are nonlinear activation functions based on tanh and sigmoid respectively.

Next, we convert the feature embeddings of all subsequences into the local WSI feature matrix $F_{local} = (F_{local,1}, \ldots, F_{local,q+1})^\top \in \mathbb{R}^{(q+1)\times d}$, and utilize the retention mechanism to update:

$$G = MSR(1; F_{local}), \tag{7}$$

where $G \in \mathbb{R}^{(q+1)\times d}$. Then attention pooling is used again to aggregate the $(q+1)$ dimension:

$$F_{global} = \sum_{p=1}^{q+1} \beta_p G_p, \tag{8}$$

where G_p represents the pth row element of G, and β_p represents as follows:

$$\beta_p = \frac{\exp\{\Gamma_{global} \tanh(\mathrm{W}_{global} G_p) \odot \mathrm{sigm}(\mathrm{U}_{global} G_p)\}}{\sum_{t=1}^{q+1} \exp\{\Gamma_{global} \tanh(\mathrm{W}_{global} G_t) \odot \mathrm{sigm}(\mathrm{U}_{global} G_t)\}}, \tag{9}$$

where $\Gamma_{global} \in \mathbb{R}^{1\times M}$, $\mathrm{W}_{global}, \mathrm{U}_{global} \in \mathbb{R}^{M\times d}$ are learnable parameters. For the WSI classification task, F_{global} is passed through a linear classifier to obtain the prediction score. The entire RetMIL is trained using the cross-entropy function as the objective loss.

3 Experiment

3.1 Datasets

CAMELYON: The CAMELYON dataset focuses on the binary classification task of lymph node metastases in breast cancer. It includes 399 WSIs from CAMELYON16 [2] and 500 WSIs from CAMELYON17 [1]. We use all data of CAMELYON16 as our training and validation set, to conduct four-fold cross-validation experiments, and choose the CAMELYON17 training set as our testing dataset.

BRACS: The BRACS dataset [3] focuses on multi-classification tasks aimed at subtype analysis of breast cancer. We conduct experiments based on the official

classification. The dataset comprises 395 training samples, 65 validation samples, and 87 test samples. We use four different sets of model initialization parameters for training and testing.

LUNG: The LUNG dataset is a binary classification task focusing on non-small cell lung cancer subtypes. The training set and validation set is collected from TCGA repository [16], containing 541 WSIs of lung adenocarcinoma (LUAD) and 458 lung squamous cell carcinoma (LUSC). The test set is from the cooperative hospital, comprising 105 LUAD and 65 LUSC WSIs. We conduct four-fold cross-validation experiments on the training and validation set and perform inference on the test set.

3.2 Experiment Setup and Evaluation Metrics

During the preprocessing stage, all WSIs are cropped into 224×224 patches at $20\times$ magnification. The length of each subsequence is set to 512. The entire experiment is conducted on one NVIDIA RTX 4090, with 100 epochs, utilizing early stopping with 15 rounds. The batch size is set as 1, and the learning rate is 1e-4, with a weight decay of 1e-5 for the Adam optimizer. All other baseline methods adopt the same experimental settings. We record the Balanced Accuracy(B-Acc) and Weighted F1-score as evaluation metrics to comprehensively evaluate the performance.

3.3 Result

Performance Evaluation against SOTA: Table 1 displays the performance of our proposed RetMIL, and we compare it with the following six state-of-the-art methods: For attention-based MIL: ABMIL [12], DSMIL [14], CLAM-MB [17]. For Transformer-based MIL: TransMIL[22], HIPT [6] and HAG-MIL [27]. In the CAMELYON dataset, our RetMIL surpasses the second-ranked model TransMIL by 3.18% and 3.43% in F1-score and balanced accuracy. In the BRACS dataset, our model leads by 1.52% and 0.86% compared with the second-ranked CLAM-MB, while also achieving the minimum variance among all models. In the LUNG dataset, RetMIL outperforms by 0.13% in balanced accuracy.

We also compared RetMIL with Transformer-based models on the CAMELYON dataset using AUC, which are shown in Fig 2.a. It can be observed that RetMIL achieves a 1.36% improvement in AUC compared to Transformer-based models. Additionally, Fig 2.b demonstrates the results of feature representations. All feature embeddings are reduced to a two-dimensional vector through the t-SNE algorithm [18]. Our observation reveals that RetMIL can better widen the gap between distinct categories while minimizing the separation among patches belonging to the same category compared with the TransMIL algorithm.

Performance at different lengths of sequences: In Transformer-based MIL methods, the length of the WSI sequence represents the number of cropped patches. We analyze model performance under different sequence lengths, and

Table 1. Mean and standard deviation of F1-score and Balanced accuracy (expressed in %) between RetMIL and current powerful MIL method. The best is in **BOLD**, and the second best is indicated with underline.

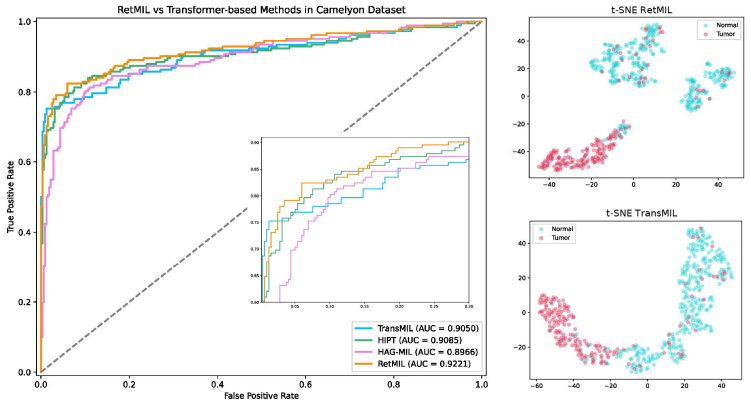
Methods	CAMELYON		BRACS		LUNG	
	F1-score	B-Acc	F1-score	B-Acc	F1-score	B-Acc
ABMIL [12]	$81.27_{3.11}$	$81.60_{2.30}$	$64.11_{5.24}$	$63.17_{4.39}$	$88.68_{3.98}$	$90.71_{3.26}$
CLAM-MB [17]	$83.06_{4.59}$	$83.37_{3.15}$	$\underline{66.99_{4.02}}$	$\underline{66.15_{3.65}}$	$87.67_{2.25}$	$89.73_{1.76}$
DSMIL [14]	$83.98_{1.79}$	$83.77_{1.30}$	$60.12_{4.52}$	$59.22_{3.23}$	$85.86_{9.15}$	$86.57_{8.18}$
TransMIL [22]	$\underline{84.06_{8.19}}$	$\underline{84.10_{5.37}}$	$62.83_{3.97}$	$61.56_{3.53}$	$\mathbf{91.75_{2.73}}$	$\underline{91.43_{3.51}}$
HIPT [6]	$78.92_{8.11}$	$80.17_{5.52}$	$66.19_{8.97}$	$65.73_{6.92}$	$81.55_{6.37}$	$84.85_{4.83}$
HAG-MIL [27]	$79.35_{5.71}$	$80.59_{4.08}$	$66.26_{4.52}$	$64.76_{4.80}$	$85.47_{4.42}$	$87.61_{3.57}$
RetMIL (Ours)	$\mathbf{87.24_{4.22}}$	$\mathbf{87.53_{3.92}}$	$\mathbf{68.51_{0.54}}$	$\mathbf{67.01_{0.71}}$	$\underline{91.51_{2.64}}$	$\mathbf{91.56_{2.77}}$

the result is shown in Table 2. Regardless of the length of the WSI sequence, our proposed method always significantly outperforms Transformer-based methods, especially for ultra-long sequences (i.e., oversized WSI), which demonstrates the effectiveness of RetMIL in long sequence analysis.

Table 2. Performance comparison of RetMIL and Transformer-based models at different sequence lengths.

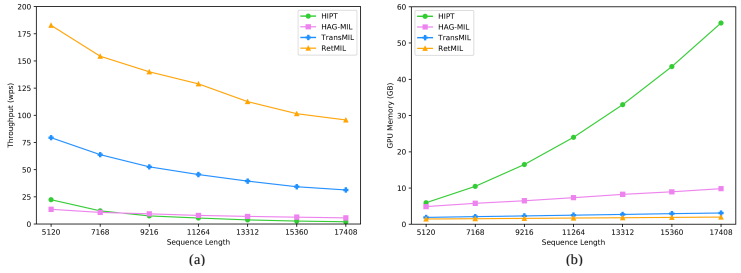
Methods	Patch Number							
	0-5000		5001-10000		10001-15000		15001-	
	F1-score	B-Acc	F1-score	B-Acc	F1-score	B-Acc	F1-score	B-Acc
TransMIL [22]	$81.83_{8.33}$	$81.69_{5.55}$	$86.66_{8.29}$	$86.51_{5.37}$	$84.47_{10.46}$	$84.78_{8.35}$	$79.29_{5.83}$	$79.89_{4.95}$
HIPT [6]	$77.98_{6.89}$	$77.63_{4.74}$	$83.86_{8.59}$	$84.34_{5.95}$	$79.89_{4.93}$	$81.43_{3.08}$	$73.57_{4.45}$	$74.17_{3.81}$
HAG-MIL [27]	$78.10_{5.74}$	$78.18_{4.41}$	$82.32_{5.81}$	$83.39_{4.13}$	$78.74_{6.20}$	$80.72_{4.80}$	$68.73_{7.15}$	$71.14_{5.13}$
RetMIL (Ours)	$\mathbf{86.67_{2.26}}$	$\mathbf{83.99_{1.70}}$	$\mathbf{89.76_{1.85}}$	$\mathbf{88.19_{1.58}}$	$\mathbf{88.59_{0.60}}$	$\mathbf{87.64_{0.42}}$	$\mathbf{82.63_{4.42}}$	$\mathbf{82.50_{4.20}}$

Inference performance: We also analyze the inference throughput and GPU memory usage under different sequence lengths with the Transformer-based models. As shown in Fig 3.b, the GPU memory consumption of HIPT and HAG-MIL almost linearly increases with increasing sequence lengths, except for the lightweight-designed TransMIL model. However, our RetMIL maintains almost constant GPU memory consumption. Fig 3.a shows that our retention model significantly improves model throughput. Even compared to the lightweight-designed TransMIL, our model maintains a nearly 1.5× lead in throughput.



(a) (b)

Fig. 2. (a) ROC curves and corresponding area under the curve(AUC) values for RetMIL and Transformer-based models (b)Visual analysis of feature dimensionality reduction between RetMIL and TransMIL.



(a) (b)

Fig. 3. (a) Comparison of throughput between RetMIL and Transformer-based models at different sequence lengths. (b) Comparison of GPU memory consumption between RetMIL and Transformer-based models at different sequence lengths.

Visulization: Fig 4 presents heatmap visualization results of our RetMIL. We select two macro-metastatic cancer slides and one micro-metastasis cancer slide from the CAMELYON17 to analyze the attention area of our model. For the kth element in subsequence i, the attention score $score_{i,k}$ can be calculated as follow:

$$s_{i,k} = \alpha_{i,k} \cdot \beta_i, \tag{10}$$

For both macro-metastatic and micro-metastatic cancer, our model can accurately and comprehensively pay attention to the cancer area marked by the pathologist, which demonstrates the great interpretability of our model.

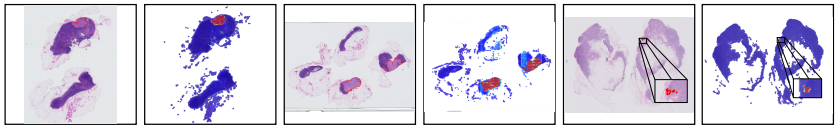


Fig. 4. Heatmap visualization of WSI examples. In each pair of images, the left part displays the standard cancer regions outlined by pathologists (indicated by red contours), while the right side shows the heatmaps generated by our RetMIL.

4 Conclusion

In this paper, we propose a retentive multiple instance learning approach called RetMIL, which uses linear retention mechanisms to reduce the computational overhead while modeling the correlation between patches. In addition, the hierarchical retentive aggregation architecture is designed to update local subsequences and characterize the global WSI sequence comprehensively. We demonstrate the superiority of RetMIL through comparative experiments on three histopathology WSI datasets. At the same time, we also compared the inference performance with the Transformer-based methods, and the results show that our proposed RetMIL has lower computational consumption.

Acknowledgement. This work was supported by The Shenzhen Engineering Research Centre (XMHT- 20230115004), Science and Technology Research Program of Shenzhen City (KCX- FZ20201221173207022), and Jilin Fuyuan Guan Food Group Co., Ltd. In.

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

References

1. Bandi, P., Geessink, O., Manson, Q., Van Dijk, M., Balkenhol, M., Hermsen, M., Bejnordi, B.E., Lee, B., Paeng, K., Zhong, A., et al.: From detection of individual metastases to classification of lymph node status at the patient level: the camelyon17 challenge. IEEE transactions on medical imaging **38**(2), 550–560 (2018)
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. 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)

4. Campanella, G., Hanna, M.G., Geneslaw, L., Miraflor, A., Werneck Krauss Silva, V., Busam, K.J., Brogi, E., Reuter, V.E., Klimstra, D.S., Fuchs, T.J.: Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nature medicine **25**(8), 1301–1309 (2019)
5. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9650–9660 (2021)
6. 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: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16144–16155 (2022)
7. Chikontwe, P., Kim, M., Nam, S.J., Go, H., Park, S.H.: Multiple instance learning with center embeddings for histopathology classification. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part V 23. pp. 519–528. Springer (2020)
8. Ding, S., Wang, J., Li, J., Shi, J.: Multi-scale prototypical transformer for whole slide image classification. In: International conference on medical image computing and computer-assisted intervention. pp. 602–611. Springer (2023)
9. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
10. Guo, Z., Zhao, W., Wang, S., Yu, L.: Higt: Hierarchical interaction graph-transformer for whole slide image analysis. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 755–764. Springer (2023)
11. Hendrycks, D., Gimpel, K.: Gaussian error linear units (gelus). arXiv preprint arXiv:1606.08415 (2016)
12. Ilse, M., Tomczak, J., Welling, M.: Attention-based deep multiple instance learning. In: International conference on machine learning. pp. 2127–2136. PMLR (2018)
13. Kang, M., Song, H., Park, S., Yoo, D., Pereira, S.: Benchmarking self-supervised learning on diverse pathology datasets. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3344–3354 (2023)
14. Li, B., Li, Y., Eliceiri, K.W.: Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14318–14328 (2021)
15. Li, J., Chen, Y., Chu, H., Sun, Q., Guan, T., Han, A., He, Y.: Dynamic graph representation with knowledge-aware attention for histopathology whole slide image analysis. arXiv preprint arXiv:2403.07719 (2024)
16. 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**(2), 400–416 (2018)
17. 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)
18. Van der Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of machine learning research **9**(11) (2008)

19. Madabhushi, A.: Digital pathology image analysis: opportunities and challenges. Imaging in medicine **1**(1), 7 (2009)
20. Otsu, N.: A threshold selection method from gray-level histograms. IEEE transactions on systems, man, and cybernetics **9**(1), 62–66 (1979)
21. Prajit Ramachandran, B.Z., Le, Q.V.: Swish: a self-gated activation function. arXiv: Neural and Evolutionary Computing, 2017 (2017)
22. Shao, Z., Bian, H., Chen, Y., Wang, Y., Zhang, J., Ji, X., et al.: Transmil: Transformer based correlated multiple instance learning for whole slide image classification. Advances in neural information processing systems **34**, 2136–2147 (2021)
23. Su, J., Ahmed, M., Lu, Y., Pan, S., Bo, W., Liu, Y.: Roformer: Enhanced transformer with rotary position embedding. Neurocomputing **568**, 127063 (2024)
24. Sun, Y., Dong, L., Huang, S., Ma, S., Xia, Y., Xue, J., Wang, J., Wei, F.: Retentive network: A successor to transformer for large language models (2023). URL http://arxiv. org/abs/2307.08621 v1
25. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
26. Wu, Y., He, K.: Group normalization. In: Proceedings of the European conference on computer vision (ECCV). pp. 3–19 (2018)
27. Xiong, C., Chen, H., Sung, J.J., King, I.: Diagnose like a pathologist: Transformer-enabled hierarchical attention-guided multiple instance learning for whole slide image classification. arXiv preprint arXiv:2301.08125 (2023)