Position-Aware Masked Autoencoder for Histopathology WSI Representation Learning

Kun Wu[1], Yushan Zheng[2(✉)], Jun Shi[3(✉)], Fengying Xie[1], and Zhiguo Jiang[1]

[1] Image Processing Center, School of Astronautics, Beihang University, Beijing 102206, China
[2] School of Engineering Medicine, Beijing Advanced Innovation Center on Biomedical Engineering, Beihang University, Beijing 100191, China
yszheng@buaa.edu.cn
[3] School of Software, Hefei University of Technology, Hefei 230601, China
juns@hfut.edu.cn

Abstract. Transformer-based multiple instance learning (MIL) framework has been proven advanced for whole slide image (WSI) analysis. However, existing spatial embedding strategies in Transformer can only represent fixed structural information, which are hard to tackle the scale-varying and isotropic characteristics of WSIs. Moreover, the current MIL cannot take advantage of a large number of unlabeled WSIs for training. In this paper, we propose a novel self-supervised whole slide image representation learning framework named position-aware masked autoencoder (PAMA), which can make full use of abundant unlabeled WSIs to improve the discrimination of slide features. Moreover, we propose a position-aware cross-attention (PACA) module with a kernel reorientation (KRO) strategy, which makes PAMA able to maintain spatial integrity and semantic enrichment during the training. We evaluated the proposed method on a public TCGA-Lung dataset with 3,064 WSIs and an in-house Endometrial dataset with 3,654 WSIs, and compared it with 6 state-of-the-art methods. The results of experiments show our PAMA is superior to SOTA MIL methods and SSL methods. The code will be available at https://github.com/WkEEn/PAMA.

Keywords: WSI representation learning · Self-supervised learning

1 Introduction

In the past few years, the development of histopathological whole slide image (WSI) analysis methods has dramatically contributed to the intelligent cancer diagnosis [4,10,15]. However, due to the limitation of hardware resources, it is

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-43987-2_69.

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14225, pp. 714–724, 2023.
https://doi.org/10.1007/978-3-031-43987-2_69

difficult to directly process gigapixel WSIs in an end-to-end framework. Recent studies usually divide the WSI analysis into multiple stages.

Generally, multiple instance learning (MIL) is one of the most popular solutions for WSI analysis [14,17,18]. MIL methods regard WSI recognition as a weakly supervised learning problem and focus on how to effectively and efficiently aggregate histopathological local features into a global representation. Several studies introduced attention mechanisms [9], recurrent neural networks [2] and graph neural network [8] to enhance the capacity of MIL in structural information mining. More recently, Transformer-based structures [13,19] are proposed to aggregate long-term relationships of tissue regions, especially for large-scale WSIs. These Transformer-based models achieved state-of-the-art performance in sub-type classification, survival prediction, gene mutant prediction, etc. However, these methods still rely on at least patient-level annotations. In the network-based consultation and communication platforms, there is a vast quantity of unlabeled WSIs not effectively utilized. These WSIs are usually without any annotations or definite diagnosis descriptions but are available for unsupervised learning. In this case, self-supervised learning (SSL) is gradually introduced into the MIL-based framework and is becoming a new paradigm for WSI analysis [1,11,16]. Typically, Chen et al. [5] explored and posed a new challenge referred to as slide-level self-learning and proposed HIPT, which leveraged the hierarchical structure inherent in WSIs and constructed multiple levels of the self-supervised learning framework to learn high-resolution image representations. This approach enables MIL-based frameworks to take advantage of abundant unlabeled WSIs, further improving the accuracy and robustness of tumor recognition.

However, HIPT is a hierarchical learning framework based on a greedy training strategy. The bias and error generated in each level of the representation model will accumulate in the final decision model. Moreover, the ViT [6] backbone used in HIPT is originally designed for nature sense images in fixed sizes whose positional information is consistent. However, histopathological WSIs are scale-varying and isotropic. The positional embedding strategy of ViT will bring ambiguity into the structural modeling. To relieve this problem, KAT [19] built hierarchical masks based on local anchors to maintain multi-scale relative distance information in the training. But these masks are manually defined which is not trainable and lacked orientation information. The current embedding strategy for WSI structural description is not complete.

In this paper, we propose a novel whole slide image representation learning framework named position-aware masked autoencoder (PAMA), which achieves slide-level representation learning by reconstructing the local representations of the WSI in the patch feature space. PAMA can be trained end-to-end from the local features to the WSI-level representation. Moreover, we designed a position-aware cross-attention mechanism to guarantee the correlation of local-to-global information in the WSIs while saving computational resources. The proposed approach was evaluated on a public TCGA-Lung dataset and an in-

house Endometrial dataset and compared with 6 state-of-the-art methods. The results have demonstrated the effectiveness of the proposed method.

The contribution of this paper can be summarized into three aspects. (1) We propose a novel whole slide image representation learning framework named position-aware masked autoencoder (PAMA). PAMA can make full use of abundant unlabeled WSIs to learn discriminative WSI representations. (2) We propose a position-aware cross-attention (PACA) module with a kernel reorientation (KRO) strategy, which makes the framework able to maintain the spatial integrity and semantic enrichment of slide representation during the self-supervised training. (3) The experiments on two datasets show our PAMA can achieve competitive performance compared with SOTA MIL methods and SSL methods.

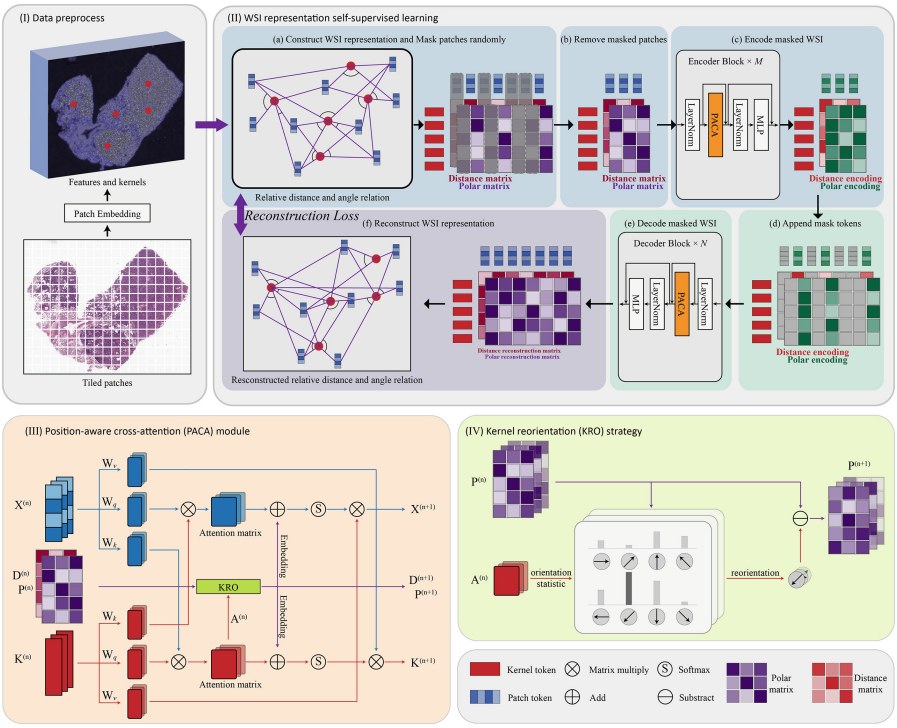

Fig. 1. The overview of the proposed whole slide image representation with position-aware masked autoencoder (PAMA), where (I) shows the data preprocessing including the patch embedding and anchors clustering, (II) describes the workflow of WSI representation self-supervised learning with PAMA, (III) is the structure of the position-aware cross-attention (PACA) module which is the core of the encoder and decoder, and (IV) shows the kernel reorientation (KRO) strategy and the detailed process is described in algorithm 1.

2 Methods

2.1 Problem Formulation and Data Preparation

MAE [7] is a successful SSL framework that learns image presentations by reconstructing the masked image in the original pixel space. We introduced this paradigm to WSI-level representation learning. The flowchart of the proposed work is illustrated in Fig. 1. First, we divided WSIs into non-overlapping image patches and meanwhile removed the background without tissue regions based on a threshold (as shown in Fig. 1(I)). Then, we applied the self-supervised learning framework DINO [3] for patch feature learning and extraction. Afterward, the features for a WSI are represented as $\mathbf{X} \in \mathbb{R}^{n_p \times d_f}$, where d_f is the dimension of the feature and n_p is the number of patches in the WSI. Inspired by KAT [19], we extracted multiple anchors by clustering the location coordinates of patches for the auxiliary description of the WSI structure. We assigned trainable representations for these anchors, which are formulated as $\mathbf{K} \in \mathbb{R}^{n_k \times d_f}$, where n_k is the number of anchors in the WSI. Here, we regard each anchor as an observation point of the tissue and assess the relative distance and orientation from the patch positions to the anchor positions. Specifically, a polar coordinate system is built on each anchor position, and the polar coordinates of all the patches on the system are recorded. Finally, a relative distance matrix $\mathbf{D} \in \mathbb{N}^{n_k \times n_p}$ and relative polar angle matrix $\mathbf{P} \in \mathbb{N}^{n_k \times n_p}$ are obtained, where $D_{ij} \in \mathbf{D}$ and $P_{ij} \in \mathbf{P}$ respectively represent the distance and polar angle of the i-th patch in the polar coordinate system that takes the position of the j-th anchor as the pole. Then, we can formulate a WSI as $S = \{\mathbf{X}, \mathbf{K}, \mathbf{D}, \mathbf{P}\}$.

2.2 Masked WSI Representation Autoencoder

Figure 1(II) illustrates the procedure of WSI representation learning. Referring to MAE [7], we random mask patch tokens with a high masking ratio (*i.e.* 75% in our experiments). The remaining tokens (as shown in Fig. 1(b)) are fed into the encoder. Each encoder block sequentially consists of LayerNorm, PACA module, LayerNorm, and multilayer perceptron (MLP), as shown in Fig. 1(c). Then, masked tokens are appended into encoded tokens to conduct the full set of tokens, which is shown in Fig. 1(d). Next, the decoder reconstructs the slide representation in feature space. Finally, mean squared error (MSE) loss is built between the reconstructed patch features and the original patch features. Referring to MAE [7], a trainable token is appended to the patch tokens to extract the global representation. After training, the pre-trained encoder will be employed as the backbone for various downstream tasks.

2.3 Position-Aware Cross-Attention

To preserve the structure information of the tissue, we propose the position-aware cross-attention (PACA) module, which is the core of the encoder and decoder blocks. The structure of PACA is shown in Fig. 1(III).

The message passing between the anchors and patches is achieved by a bi-directional cross-attention between the patches and anchors. First, the anchors collect the local information from the patches, which is formulated as

$$\mathbf{K}^{(n+1)} = \sigma\left(\frac{\mathbf{K}^{(n)}\mathbf{W}_q^{(n)} \cdot \left(\mathbf{X}^{(n)}\mathbf{W}_k^{(n)}\right)^{\mathrm{T}}}{\sqrt{d_e}} + \varphi_d(\mathbf{D}^{(n)}) + \varphi_p(\mathbf{P}^{(n)})\right) \cdot \left(\mathbf{X}^{(n)}\mathbf{W}_v^{(n)}\right), \quad (1)$$

where $\mathbf{W}_l \in \mathbb{R}^{d_f \times d_e}, l = q, k, v$ are learnable parameters with d_e denoting the dimension of the head output, σ represents the softmax function, and φ_d and φ_p are the embedding functions that respectively take the distance and polar angle as input and output the corresponding trainable embedding values. Symmetrically, each patch token catches the information of all anchors into their own local representations by the equations

$$\mathbf{X}^{(n+1)} = \sigma\left(\frac{\mathbf{X}^{(n)}\mathbf{W}_q^{(n)} \cdot \left(\mathbf{K}^{(n)}\mathbf{W}_k^{(n)}\right)^{\mathrm{T}}}{\sqrt{d_e}} + \varphi_d^{\mathrm{T}}(\mathbf{D}^{(n)}) + \varphi_p^{\mathrm{T}}(\mathbf{P}^{(n)})\right) \cdot \left(\mathbf{K}^{(n)}\mathbf{W}_v^{(n)}\right). \quad (2)$$

The two-way communication makes the patches and anchors timely transmit local information and perceive the dynamic change of global information. The embedding of relative distance and polar angle information helps the model maintain the semantic and structural integrity of the WSI and meanwhile prevents the WSI representation from collapsing to the local area throughout the training process.

In terms of efficiency, the computational complexity of self-attention is $O({n_p}^2)$ where n_p is the number of patch tokens. In contrast, our proposed PACA's complexity is $O(n_k \times n_p)$ where n_k is the number of anchors. Notice that $n_k << n_p$, the complexity is close to $O(n_p)$, i.e. linear correlation with the size of the WSI.

2.4 Kernel Reorientation

As for the polar angle matrix $\mathbf{P} \in \mathbb{N}^{n_k \times n_p}$, we specify the horizontal direction of all the anchors as the initial polar axis. In natural scene images, there is natural directional conspicuousness of semantics. For instance, in the case of a church, it is most likely to find a door below the windows rather than be located above them. But histopathology images have no absolute definition of direction. The semantics of WSI will not change with rotation and flip. Namely, it is isotropic. Embedding the orientation information with a fixed polar axis will lead to ambiguities in various slides.

To address this problem, we design a kernel reorientation (KRO) strategy to dynamically update the polar axis during the training. As shown in Fig. 1(IV), we equally divide the polar coordinate system into N bins and calculate the sum of the attention scores from each bin. Then, the orientation with the highest score is recognized as the new polar axis for the anchor. Based on the updated polar axis, we can then amend $\mathbf{P}^{(n)}$ to $\mathbf{P}^{(n+1)}$. The detailed algorithm is described in Algorithm 1.

3 Experiments and Results

3.1 Datasets

We evaluated the proposed method on two datasets, the public TCGA-Lung and the in-house Endometrial dataset, which are introduced as follows.

Algorithm 1: Kernel Reorientation algorithm.

Input:
$\mathbf{P}^{(n)} \in \mathbb{N}^{H \times n_k \times n_p}$: The relative polar angle matrix of n-th block, where H is the head number of multi-head attention, n_k is the number of anchors in the WSI, n_p is the number of patches in the WSI;
$\mathbf{A}^{(n)} \in \mathbb{R}^{H \times n_k \times n_p}$: The attention matrix from anchors to patches, defined as $\mathbf{A}^{(n)} = \frac{\boldsymbol{\kappa}^{(n)}\mathbf{w}_q^{(n)} \cdot (\mathbf{x}^{(n)}\mathbf{w}_k^{(n)})^{\mathrm{T}}}{\sqrt{d_e}}$;
D^{score}: A dictionary taking the angle as KEY for storing attention scores;
Output: $\mathbf{P}^{(n+1)} \in \mathbb{R}^{H \times n_k \times n_p}$: The updated polar angle matrix.

```
for h in H do
    for i in n_k do
        D^score = 0
        for j in n_p do
            D^score[P^(n)_{h,i,j}] += A^(n)_{h,i,j};
        end
        P^(n)_{h,i,max} = arg max D^score; // Find the orientation that has the highest
           attention score.
        for j in n_p do
            P^(n+1)_{h,i,j} = P^(n)_{h,i,j} − P^(n)_{h,i,max};// Reorientation.
        end
    end
end
```

TCGA-Lung dataset is collected from The Cancer Genome Atlas (TCGA) Data Portal. The dataset includes a total of 3,064 WSIs, which consist of three categories, namely Tumor-free (Normal), Lung Adenocarcinoma (LUAD), and Lung Squamous Cancer (LUSC),

Endometrial dataset includes 3,654 WSIs of endometrial pathology, which includes 8 categories, namely Well/Moderately/Low-differentiated endometrioid adenocarcinoma, Squamous differentiation carcinoma, Plasmacytoid carcinoma, Clear cell carcinoma, Mixed-cell adenocarcinoma, and benign tumor.

Each dataset was randomly divided into training, validation and test sets according to 6:1:3 while keeping each category of data proportionally. We conducted WSI multi-type classification experiments on the two datasets. The validation set was used to perform an early stop. The results of the test set were reported for comparison.

3.2 Implementation Details

The WSI representation pre-training stage uses all training data and does not involve any supervised information. During the downstream classification task,

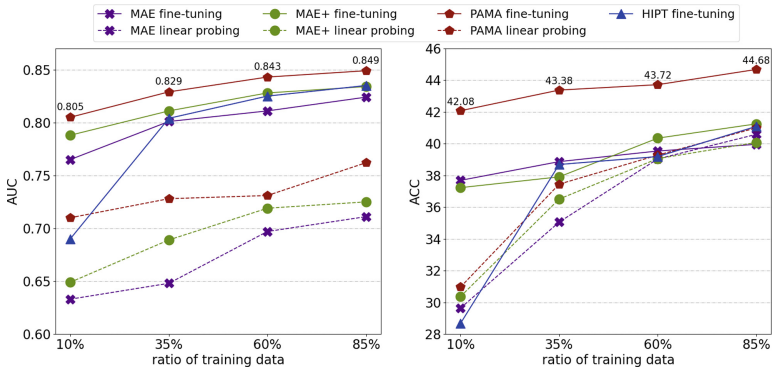

Fig. 2. Semi-supervised experiments with 10%, 35%, 60% and 85% of labelled data on the Endometrial dataset. Solid lines represent fine-tuning results and dotted lines represent liner probing results.

Table 1. Ablation study on 35% of labelled Endometrial dataset.

NO.	**Dis**	**Polar**	**KRO**	AUC	ACC
1	✓	✓	✓	0.829	43.38
2	✓	✓		0.809 (↓0.020)	39.52 (↓3.86)
3	✓			0.808 (↓0.021)	40.82 (↓2.56)
4		✓	✓	0.810 (↓0.019)	39.76 (↓3.62)
5				0.795 (↓0.034)	36.33 (↓7.05)

the pre-trained encoder is utilized as the slide representation extractor, and the [*CLS*] token is fed into the following classifier consisting of a multilayer perceptron (MLP) and a fully connected layer. Following the protocol in self-supervised learning [7], we evaluated the quality of pre-training with the two approaches: 1) **Fine-tuning** is to train the whole network parameters, including WSI encoder and classifier; 2) **Linear probing** is to freeze the encoder and only train the classifier. The usage of [*CLS*] token refers to the MAE [7] framework, which was concatenated with patch tokens. During pre-training, the [*CLS*] token is not involved in loss computation, but it continuously interacts with kernels and receives global information. After pre-training, the pre-trained parameters of the [*CLS*] token will be loaded for fine-tuning and linear probing.

To ensure the uniformity of patch features, we choose DINO [3] to extract patch features on the magnification under 20× lenses. Accuracy (ACC) and area under the ROC curve (AUC) are employed as evaluation metrics. We implemented all the models in Python 3.8 with PyTorch 1.7 and Cuda 10.2 and run the experiments on a computer with 4 GPUs of Nvidia Geforce 2080Ti.

3.3 Effectiveness of the WSI Representation Learning

We first conducted experiments on the Endometrial dataset to verify the effectiveness of self-supervised learning for WSI analysis under label-limited conditions. The results are shown in Fig. 2, where the performance obtained with different ratios of labeled training WSIs are compared. MAE [7] based on the patch features is implemented as the baseline. Furthermore, we applied the proposed distance and polar angle embedding to the self-attention module of MAE [7], which is referred to as MAE+ in Fig. 2.

Table 2. Comparison with weakly-supervised MIL and slide-level self-learning study on the two datasets for sub-type classification.

Methods	**TCGA-Lung**				**Endometrial**			
	35%		100%		35%		100%	
	AUC	ACC	AUC	ACC	AUC	ACC	AUC	ACC
DSMIL [12]	0.911	75.00	0.938	80.11	0.761	38.21	0.786	39.32
TransMIL [13]	0.932	79.62	0.959	84.35	0.783	38.43	0.798	40.01
SETMIL [18]	0.937	80.21	0.962	84.95	0.795	38.71	0.831	40.84
KAT [19]	0.951	83.37	0.965	85.81	0.799	38.89	0.835	41.93
HIPT [5]	0.967	84.23	0.977	87.83	0.804	38.69	0.842	40.63
MAE [7]	0.965	83.90	0.970	87.50	0.801	38.87	0.832	41.95
MAE+	0.969	85.07	0.981	88.25	0.811	37.91	0.845	42.85
PAMA	**0.982**	**90.84**	**0.988**	**92.48**	**0.829**	**43.38**	**0.851**	**43.64**

Overall, PAMA consistently achieves significantly better performance across all the label ratios than MAE [7] and HIPT [5]. These results have demonstrated the effectiveness of PAMA in WSI representation pre-training. Moreover, PAMA achieves the best stability in AUCs and ACCs when the label ratios are reduced from 85% to 10%. This is of practical importance as it reduces the dependence on a large number of labeled WSIs for training robust WSI analysis models. Meanwhile, it means that we can utilize the unlabeled WSIs to improve the capacity of the models with the help of PAMA. HIPT [5] is a two-stage self-learning framework, which first leverages DINO [3] to pre-train patches (256×256) divided from regions (4096×4096) and then utilizes DINO-4k [5] to pre-train regions of WSIs. The multi-stage framework accumulated the training bias and noise, which caused an AUC gap of HIPT [5] to MAE [7] and PAMA, especially trained with only 10% labeled WSIs. We also observed a significant improvement when comparing MAE+ with MAE [7]. It indicates the proposed distance and polar angle embedding strategy is more appreciated than the positional embedding of ViT [6] to describe the structure of histopathological WSIs. Please refer to the supplementary materials for more detailed results.

3.4 Ablation Study

Then, we conducted ablation experiments to verify the necessity of the proposed structural embedding strategy. The detailed results are shown in Table 1, where all the models were fine-tuned with 35% training WSIs. It shows that the AUC decreases by 0.019 and 0.021, respectively, when the distance or polar angle embedding is discarded. And, when removing both the distance and polar angle embedding, the AUC drops by 0.034. These results demonstrate that local and global spatial information is crucial for PAMA to learn WSI representations.

3.5 Comparison with SOTA Methods

Finally, we additionally compared the proposed PAMA with four weakly-supervised methods, DSMIL [12], TransMIL [13], SETMIL [18] and KAT [19]. The results are shown in Table 2. Overall, PAMA consistently achieves the best performance. In comparison with the second-best methods, PAMA achieves an increase of 0.015/0.011 and 0.025/0.009 in AUCs on TCGA and Endometrial datasets, respectively, by using 35%/100% labeled WSIs. Moreover, PAMA reveals the most robust capacity when reducing the training data from 100% to 35%, with AUC decreasing slightly from 0.988 to 0.982 and from 0.851 to 0.829 on the two datasets. TransMIL [13], SETMIL [18] and KAT [19] are state-of-the-art methods for histopathological image classification. They all considered the spatial adjacency of patches but neglected the orientation relationships of the patches. It is the main reason that the three methods cannot surpass our method even with 100% training WSIs.

4 Conclusion

In this paper, we proposed an effective self-supervised representation learning framework for WSI analysis. The experiments on two large-scale datasets have demonstrated the effectiveness of PAMA in the condition of limited-label. The results have shown superiority to the existing weakly-supervised and self-supervised MIL methods. Future work will focus on training the WSI representation model based on datasets across multiple organs, thus promoting the generalization ability of the model for different downstream tasks.

Acknowledgements. This work was partly supported by the National Natural Science Foundation of China (Grant No. 62171007, 61901018, and 61906058), and partly supported by the Fundamental Research Funds for the Central Universities of China (grant No. JZ2022HGTB0285).

References

1. Azizi, S., et al.: Robust and efficient medical imaging with self-supervision. arXiv preprint arXiv:2205.09723 (2022)

2. Campanella, G., et al.: Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nat. Med. **25**(8), 1301–1309 (2019)
3. Caron, M., et al.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 9650–9660 (2021)
4. Chen, C., Lu, M.Y., Williamson, D.F., Chen, T.Y., Schaumberg, A.J., Mahmood, F.: Fast and scalable search of whole-slide images via self-supervised deep learning. Nat. Biomed. Eng. **6**(12), 1420–1434 (2022)
5. 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)
6. Dosovitskiy, A., et al.: An image is worth 16x16 words: transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
7. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16000–16009 (2022)
8. Huang, Z., Chai, H., Wang, R., Wang, H., Yang, Y., Wu, H.: Integration of patch features through self-supervised learning and transformer for survival analysis on whole slide images. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12908, pp. 561–570. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87237-3_54
9. Ilse, M., Tomczak, J., Welling, M.: Attention-based deep multiple instance learning. In: International Conference on Machine Learning, pp. 2127–2136. PMLR (2018)
10. Jaume, G., Song, A.H., Mahmood, F.: Integrating context for superior cancer prognosis. Nat. Biomed. Eng. 1–3 (2022)
11. Koohbanani, N.A., Unnikrishnan, B., Khurram, S.A., Krishnaswamy, P., Rajpoot, N.: Self-path: self-supervision for classification of pathology images with limited annotations. IEEE Trans. Med. Imaging **40**(10), 2845–2856 (2021)
12. 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)
13. Shao, Z., et al.: Transmil: transformer based correlated multiple instance learning for whole slide image classification. Adv. Neural Inf. Process. Syst. **34**, 2136–2147 (2021)
14. Su, Z., Tavolara, T.E., Carreno-Galeano, G., Lee, S.J., Gurcan, M.N., Niazi, M.: Attention2majority: weak multiple instance learning for regenerative kidney grading on whole slide images. Med. Image Anal. **79**, 102462 (2022)
15. Wu, Z., et al.: Graph deep learning for the characterization of tumour microenvironments from spatial protein profiles in tissue specimens. Nat. Biomed. Eng. 1–14 (2022)
16. Yang, P., Hong, Z., Yin, X., Zhu, C., Jiang, R.: Self-supervised visual representation learning for histopathological images. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12902, pp. 47–57. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87196-3_5
17. Yu, J.G., et al.: Prototypical multiple instance learning for predicting lymph node metastasis of breast cancer from whole-slide pathological images. Med. Image Anal. 102748 (2023)

18. Zhao, Y., et al.: Setmil: spatial encoding transformer-based multiple instance learning for pathological image analysis. In: Medical Image Computing and Computer Assisted Intervention-MICCAI 2022: 25th International Conference, Singapore, 18–22 September 2022, Proceedings, Part II, LNCS, pp. 66–76. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16434-7_7
19. Zheng, Y., Li, J., Shi, J., Xie, F., Jiang, Z.: Kernel attention transformer (KAT) for histopathology whole slide image classification. In: Medical Image Computing and Computer Assisted Intervention-MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part II, LNCS. pp. 283–292. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16434-7_28