Patch-Slide Discriminative Joint Learning for Weakly-Supervised Whole Slide Image Representation and Classification*

Jiahui Yu[1], Xuna Wang[1], Tianyu Ma[1], Xiaoxiao Li[2], and Yingke Xu[1(✉)]

[1] Zhejiang University, Hangzhou 310027, China
[2] The University of British Columbia, Vancouver, BC V6T 1Z4, Canada
yingkexu@zju.edu.cn

Abstract. In computational pathology, Multiple Instance Learning (MIL) is widely applied for classifying Giga-pixel whole slide images (WSIs) with only image-level labels. Due to the size and prominence of positive areas varying significantly across different WSIs, it is difficult for existing methods to learn task-specific features accurately. Additionally, subjective label noise usually affects deep learning frameworks, further hindering the mining of discriminative features. To address this problem, we propose an effective theory that optimizes patch and WSI feature extraction jointly, enhancing feature discriminability. Powered by this theory, we develop an angle-guided MIL framework called PSJA-MIL, effectively leveraging features at both levels. We also focus on eliminating noise between instances and emphasizing feature enhancement within WSIs. We evaluate our approach on Camelyon17 and TCGA-Liver datasets, comparing it against state-of-the-art methods. The experimental results show significant improvements in accuracy and generalizability, surpassing the latest methods by more than 2%. Code will be available at: https://github.com/sm8754/PSJAMIL.

Keywords: WSI, Digital pathology, MIL, Classification.

1 Introduction

Modern microscopes digitize traditional slides into WSIs, which often contain gigapixels and cannot be directly analyzed by neural networks. MIL is widely used to solve this problem [1-4]. The pioneering work [5] proposed dividing the WSI foreground region into patches and then classifying based on the feature embeddings of all patches. These early works [6,7] significantly improved the classification performance, proving the effectiveness of MIL. In weakly supervised learning scenarios, the negative instances are unknown, leading to significant uncertainty. Therefore, it is important to carefully adjust each weightsand combine them, which helps extract genuinely

* Jiahui Yu and Xuna Wang—Equal contribution.

distinctive features. Proper calibration of the weighting is pivotal for nuanced comprehension and accurate decision-making [8,9].

In response to this problem, some recent work has made the following efforts. To learn from unlabeled image patches, DTFD-MIL [10] introduced the concept of pseudo-bags and constructed a double-tier MIL framework. Meanwhile, SSL + MIL [11] introduced semi-supervised learning (SSL) to MIL. In addition, exploring the potential hard instances is also a mainstream approach. MHIM-MIL [12] used a Siamese structure with a consistency constraint. CIMIL [13] aggregated instance representations with normalized distances between instances and the critical instance. The above methods are all designed for patch-level feature learning. However, since the proportion of negative instances is not fixed and samples may be incorrectly labeled, the above method still suffers from noisy information interference [14,15]. In addition, previous classification networks used FC layers to convert features into class confidence scores. Based on this method, sample features extracted are poorly discriminative and challenging to optimize intuitively [16-18].

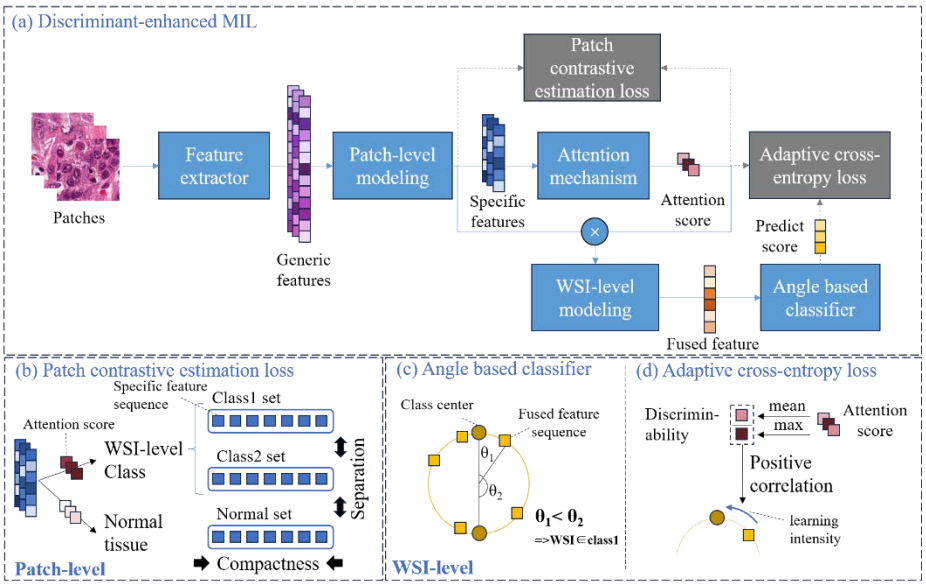

Fig. 1. (a) Implementation process of PSJA-MIL, mainly divided into patch images -> patch features -> WSI feature -> WSI class -> loss. (b) Patch contrastive estimation loss utilizes the attention scores of patches and selects patches for contrastive learning. (c) PCS-classifier improves the interpretability of each WSI-level feature. (d) Adaptive cross-entropy loss uses the discriminability to adjust the learning intensity for samples.

We propose a theory called patch-slide discrimination joint learning. Specifically, we jointly optimize patch and WSI feature extraction across samples to enhance feature discrimination. For patch-level features, we assign contrastive loss and pseudo-labels based on the confidence of each patch. For WSI-level features, the optimization strength of cross-loss is adaptively adjusted by measuring the discriminability of

samples. Based on this theory, we propose an Prototypical Cosine Similarity Guided classification method (PCS-classifier), which enhances the correlation of sample feature vectors to confidence scores. The above design units are shown in Fig. 1(b-d), which will be discussed in the following sections. Finally, we utilize a transformer-based approach to develop a MIL framework called PSJA-MIL, as shown in Fig. 1(a). This framework is designed to learn and integrate dual-level features. We evaluate our method on two challenging datasets, Camelyon17 and TCGA-Liver. PSJA-MIL achieved state-of-the-art results on the pathological image classification task, and each critical component is practical and versatile.

2 Method

2.1 Patch-Slide Discriminative Joint Learning

To improve the discrimination of patch and slide-level features, we design patch contrastive estimation loss and adaptive cross-entropy loss, respectively. These losses are cleverly designed using attention scores, which are generated by the gated attention mechanism [19].

Patch Contrastive Estimation Loss. In the heatmap visualization of previous work [20] and this work, it can be found that the category with the highest attention is usually the target category, while the category with the lowest attention is usually the normal tissue. Based on the above experience, we annotate some of the unlabeled patches based on attention scores. Considering the varying tumor areas in different WSIs, we set a relatively flexible selection method. Specifically, in each WSI, the first 15 patches with an attention score in the $[\max(scores)-0.3, \max(scores)]$ interval are labeled in the same category as the WSI. Meanwhile, patches labeled normal tissue should have attention scores in the range $[\min(scores), \min(scores)+1e-9]$. To avoid focusing on the single characteristic of marginal tissue, we select one patch for every two patches, with a maximum of 5. According to the above selection method, the features of the selected patches in the current batch constitute the set K. Based on the same method, the features of different categories of patches (across batches) are put into different sets, and each set B_* contains a maximum of patches extracted from 20 WSIs. Q is the total set of the above sets.

We introduce the concept of contrastive learning [21], identifying commonalities among patches of the same class by contrasting samples under different labels. Thus, the patch contrastive estimation loss is defined as follows:

$$L_p = -\frac{1}{num(K)} \sum_{f_S^k \in K} \log \frac{\sum_{f_S^b \in B_t} e^{\left(\tau f_S^k \left(f_S^b\right)^T\right)}}{\sum_{f_S^q \in Q} e^{\left(\tau f_S^k \left(f_S^q\right)^T\right)}} \tag{1}$$

where, B_t is the set of features under the same category with f_S^k; $f_S^k \left(f_S^b\right)^T$ and $f_S^k \left(f_S^q\right)^T$ are used to measure the similarity of two vectors; $\tau \in (0,1)$ is a hyperparameter to

prevent overfitting to patches potentially assigned incorrect labels. As shown in Fig 1(b), features within the same class become more similar, while those between different classes become more distinct.

Adaptive Cross-entropy Loss. To avoid the influence of noisy labels and tissues, we adaptively determine the learning intensity for each sample based on its discriminability, as shown in Fig. 1(d). We use the attention score sequence A to measure the confidence of each patch being of the same class as the WSI. If there are more patches of the same category as the WSI, it indicates a higher discriminability of the WSI. Accordingly, we measure the discriminability based on the ratio of the average attention score to the highest attention score, denoted as $D = mean(A)/\max(A)$. The discernibility D is directly proportional to the learning intensity. However, this configuration may lead the model to focus excessively on highly discriminative sample during training. As a result, the model may overfit these specific samples, while neglecting the generalization of other samples. L2 regularization is used to mitigate this issue, with D employed to adjust the strength of the regularization. Conversely, D is inversely correlated with the regularization strength.

Based on the analysis above, we propose the adaptive cross-entropy loss:

$$L_w = -\frac{1}{N}\sum_{n=1}^{N}\log\frac{e^{Y_{n,t}+g(D_n)}}{\sum_{c=1}^{C}e^{Y_{n,c}}} + r_{L2}\left(1-D_n\right)^2\left(\frac{1}{2}\sum_{l=1}^{L}{M_l}^2\right) \tag{2}$$

where N is the batch size, C is the number of categories, t represents the ground-truth label, $Y_{n,t}$ is the predicted score for the right category, and M_l is the layer parameter. $g(D_n)$ is an increasing function whose specific form is provided later; $r_{L2} > 0$ is the regularization coefficient, $(1-D_n)^2$ decreases monotonically within the value interval.

2.2 Prototypical Cosine similarity Guided MIL Framework

Traditional classification models usually use an FC layer to aggregate all features and generate confidence scores for each category. To encourage the model to learn the discriminative features, we designed a new classification mechanism, namely the PCS-classifier. As shown in Fig. 1(c), the sample category is determined by the angle between the generated feature and the class center, which facilitates intuitive optimization of feature learning. Specifically, we establish class center vectors with the maximum margin to enhance the discriminability of WSI-level features. The FC layer is retained for feature aggregation and compression. Its output and the center vectors are normalized to a fixed magnitude ($\|\hat{F}\| = t > 0$, $\|\hat{W}_c\| = 1$). Under this mechanism, the confidence level for each category is calculated as follows:

$$Y_c = \|\hat{W}_c\|\|\hat{F}\|\cos\theta_{\hat{W}_c,\hat{F}} \tag{3}$$

where, $\theta_{\hat{W}_c,\hat{F}}$ is the angle between the feature and the class center.

The adaptive cross-entropy loss optimizes the model based on WSI-level classification results. Under the PCS-classifier, increasing the angle between the feature vector

and the actual class center can enhance the optimization intensity. Thus, the adjusted true class confidence in Eq. (2) can be defined as:

$$\hat{Y}_{n,t} = Y_{n,t} + g\left(D_n\right) = \left\|\hat{W}_{n,t}\right\|\left\|\hat{F}_n\right\|\cos\left(\theta_{\hat{W}_{n,t},\hat{F}_n} + r_a D_n\right) \tag{4}$$

where, $r_a > 0$ is the interval of the angle margin, D_n adjusts the size of the margin.

Before the classification head, the specific implementation details of PSJA-MIL are as follows: Firstly, generic features of patches are extracted offline. These features containing information such as texture and edges are not uniquely specific to the classification task. Hence, an encoder layer is used for feature mapping to generate task-specific patch-level features. Each patch is processed in parallel and optimized by patch contrastive estimation loss. Since the patch is annotated with pseudo-labels, to mitigate the impact of potential erroneous labels on training, the contribution coefficient should be relatively small. Experimental verification has determined that a coefficient of 0.4 is appropriate.

Then, we employ the gated attention mechanism [19] to measure the attention score of each patch as confidence belonging to the same class as the WSI. Based on the attention scores, the features of each patch are weighted, thereby adaptively avoiding the interference of negative patches. Then, a Transformer containing one layer of encoder and decoder is used to learn the correlations between different patches. Finally, the output vectors are averaged, and the result is WSI-level feature. Learning and optimizing WSI-level features and task-specific patch-level features together constitute patch-slide discrimination joint learning.

3 Experiments and Results

3.1 Datasets and Experimental Settings

Datasets. We evaluate the proposed method on two public datasets. 1) Camelyon17 [22]. This dataset identifies lymph node metastases containing normal and tumor categories. Among them, tumor samples can be divided into three types. Isolated tumor cells (ITC) is the minorest type of metastasis, smaller than 0.2 mm or less than 200 cells, which is very challenging. Since only the annotations for the training set are publicly accessible, we used the training set for experiments. These data encompass 500 WSIs from 100 patients in 5 medical centers. 2) TCGA-Liver. This dataset is collected from The Cancer Genome Atlas (TCGA) Data Portal, containing two categories: Liver Cancer (LIHC) and Bile Duct Cancer (CHOL). There is a severe imbalance in the dataset, including 379 LIHC WSIs and 36 CHOL WSIs. This poses a significant challenge to the feature learning. Each dataset is randomly split into a training-validation set and a test set in a 7:3 ratio. The training validation set is performed five-fold cross-validation, and the test set is used to report and compare model performance.

Implementation Details. We refer to the CLAM model [23] for preprocessing WSIs, where all tumor regions in the WSIs are segmented into 256×256 patches. For Camelyon17, the magnification is 40x, and for TCGA-LIBD, it is 20x. Then, 1024-dim

features are extracted from the patches using a Res-Net50 pre-trained on ImageNet. Furthermore, we also extracted features using KimiaNet [24] (a DenseNet121 model pre-trained on TCGA slides). These features were only used in ablation experiments. During training, the Lookahead-RAdam optimizer was used with an initial learning rate of 0.0001 and a batch size 1. Accuracy (ACC) and area under the ROC curve (AUC) were used as evaluation metrics, and we reported the mean and variance of these metrics under five-fold cross-validation. All experiments were completed on a computer with an RTX 4090 GPU, using PyTorch 1.13.0 and Cuda 11.7 in Python 3.9.

Parameter Setting. Experiments were performed on the TCGA-Liver dataset to determine the best configurations for each proposed module. The optimal parameter settings were selected based on the AUC and Acc metrics: 1) In the patch contrastive estimation loss, $\tau = 0.5$; 2) In the PCS-classifier, t=5; 3) In the adaptive cross-entropy loss, $r_a = 0.25$ and $r_{L2} = 0.001$.

3.2 Comparison with SOTA Methods

We compared the proposed PSJA-MIL method with state-of-the-art approaches, including MIL methods and contrastive learning methods (SCL-WC [29]). The results are shown in Table 1. Among the SOTA models, SCL-WC performed exceptionally well on the TCGA-Liver dataset. However, PSJA-MIL improved the AUC by 2.04% and accuracy by 2.5% relative to this model. On the Camelyon17 dataset, our model offered even more significant improvements compared to SOTA models. Specifically, the AUC increased by 2.39%, and accuracy increased by 2.72%. Furthermore, the standard deviation of both metrics indicates that our model had a stable performance in the five-fold cross-validation. Our AUC metric was most stable across two datasets. This demonstrates the outstanding generalization ability of the proposed discriminative enhancement method.

Table 1. Comparison of classification results on two datasets.

Methods	Camelyon17		TCGA-Liver	
	AUC	Acc	AUC	Acc
CLAM [23]	88.01±0.027	82.99±0.021	91.96±0.601	89.15±1.814
TransMIL [20]	92.39±0.028	87.70±0.037	94.19±0.931	93.51±2.648
H2-MIL [25]	90.79±0.029	84.89±0.030	91.85±0.561	86.56±1.438
MHIM-MIL [12]	89.68±0.022	85.13±0.010	93.78±1.642	90.27±3.182
DAS-MIL [26]	**95.78±0.013**	**92.24±0.028**	93.51±0.573	92.92±2.497
HAG-MIL [27]	93.04±0.014	87.23±0.026	95.04±0.724	92.58±2.721
TPMIL [28]	89.22±0.021	85.28±0.023	93.15±0.487	90.66±1.945
SCL-WC [29]	92.20±0.021	87.32±0.025	**95.45±1.437**	**93.92±2.671**
Ours	**98.17±0.011**	**94.96±0.019**	**97.49±0.450**	**96.42±1.471**

3.3 Ablation Study

Method design. To validate the effectiveness and versatility of the proposed method, we improved upon the classic Transformer-based model TransMIL [20] and evaluated different models using two datasets. Specifically, we applied patch contrastive estimation loss to TransMIL's correlation modeling of the sequence, creating Patch-TransMIL. On this basis, changing the prediction head to an PCS-classifier and replacing the cross-entropy loss with adaptive cross-entropy loss resulted in Both-TransMIL. The results are shown in Table 2.

With ResNet as the feature extractor, the improved models showed enhancement over TransMIL. The discriminative enhancement based on patch-level features led to an improvement of over 1% in all metrics for Patch-TransMIL. The discriminative enhancement based on WSI-level features further provided Both-TransMIL with an overall improvement of about 2% compared to the former. PSJA-MIL outperformed Both-TransMIL on two datasets, particularly on Camelyon17, where our model increased accuracy by 3.85%. This demonstrates the rationality of each component and the overall framework.

To further validate the effectiveness of discriminative enhancement based on patch-level features, we compared the results using KimiaNet and ResNet as feature extractors. Unlike ResNet, KimiaNet is pre-trained using pathological images. Patch-TransMIL based on ResNet showed similar results to TransMIL based on KimiaNet, indicating that the patch contrastive estimation loss enabled TransMIL to learn better pathological features. Patch-TransMIL based on KimiaNet also showed an improvement of over 1% compared to TransMIL. This shows that the loss allows the model to further learn task-specific features based on general pathological features.

Table 2. Ablation studies of PSJA-MIL on two datasets.

Extractor	Model	Camelyon17		TCGA-Liver	
		AUC	Acc	AUC	Acc
KimiaNet	TransMIL	93.98±0.012	88.96±0.022	96.11±0.489	95.25±2.074
	Patch-TransMIL	95.16±0.011	90.37±0.019	97.13±0.338	96.49±1.966
ResNet	TransMIL	92.39±0.028	87.70±0.037	94.19±0.931	93.51±2.648
	Patch-TransMIL	94.07±0.019	89.54±0.029	95.35±0.626	94.73±2.145
	Both-TransMIL	96.34±0.013	91.11±0.016	97.02±0.479	96.38±1.924
	Ours	**98.17±0.011**	**94.96±0.019**	**97.49±0.450**	**96.42±1.471**

3.4 Results of Joint Learning

To validate that the proposed method can effectively generate discriminative features at both the patch and WSI levels, we conducted experiments on the Camelyon17 dataset. As shown in Fig. 2(a), we utilized PCA for feature space visualization. For WSI-level features, we further subdivided tumor samples into macro, micro, and ITC categories. All tumor and normal samples tend to gravitate towards the proposed center and

separate from each other, with only a few ITC samples mixing in the normal cluster. Due to the adaptive cross-entropy loss, the distance of tumor samples from the proposed center correlates positively with the size of the tumor tissue. Consequently, the learning avoids the interference of negative tissue features, thereby acquiring discriminative WSI-level features. We assigned values to patches for patch-level features according to the pseudo-labeling approach mentioned in the text. Then, we randomly sampled 1% of the remaining patches from each slide and labeled them agnostic. Patches labeled normal or tumor are divided into different clusters in the feature space and located at the edges of the agnostic feature cluster, respectively. This indicates that the model is capable of learning discriminative patch-level features.

In addition, we compared the actual annotation of the ITC sample with its heatmap. Fig. 2(b) shows an ITC sample magnified ten times, annotated tumor tissue in the left image, and visualized heatmap in the right image. We observe that PSJA-MIL can accurately identify lesion tissue smaller than 200 micrometers. This confirms the effectiveness of the proposed patch-slide discriminative joint learning approach.

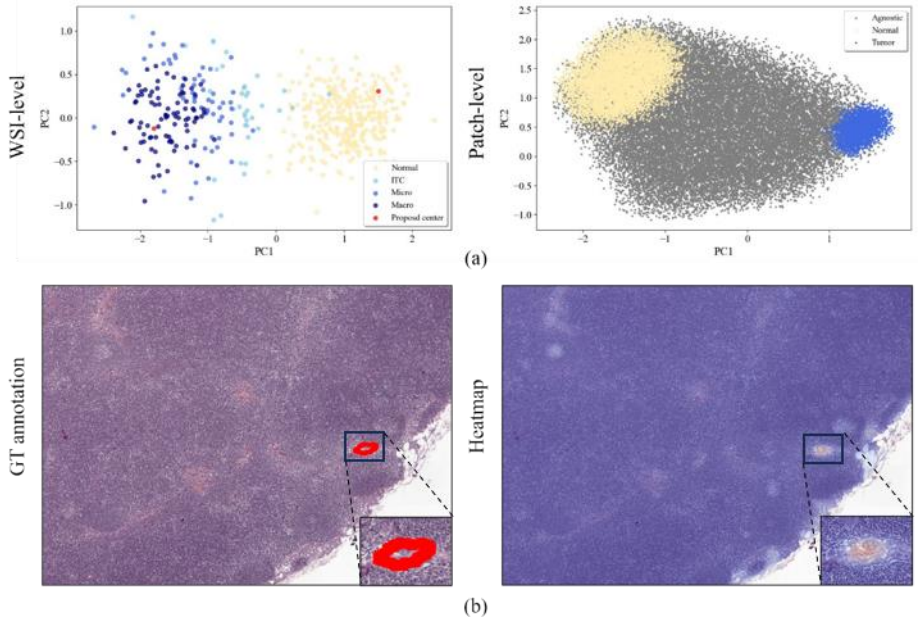

Fig. 2. Results of joint learning on the Camelyon17 dataset. (a) Visualizing the WSI and patch-level feature space. (b) Comparison of lesion annotation with heatmap in an ITC sample.

4 Conclusion

We propose patch-slide discriminative joint learning, which can effectively learn discriminative features at patch and WSI levels. To avoid the impact of negative instances and noise samples, we design two losses for feature optimization at these two levels and design PCS-classifier to further enhance the discriminability of features. To

maximize the effectiveness of these critical components, we developed a Transformer framework called PSJA-MIL. A large number of experiments demonstrate the effectiveness and portability of the proposed method.

Acknowledgments. The authors would like to acknowledge the support from the Zhejiang Provincial Natural Science Foundation of China (LZ23H180002 and LQ23F030001), the Key projects for agriculture and social development in Hangzhou (20231203A13), the Cao Guangbiao High-tech Development Fund (2022RC009), and Autism Research Special Fund of Zhejiang Foundation For Disabled Persons (2023008). The results shown here are in part based upon data generated by the TCGA Re-search Network: https://www.cancer.gov/tcga.

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

References

1. Yu, Z., Lin, T., Xu, Y. : SLPD: Slide-Level Prototypical Distillation for WSIs. In: Greenspan, H., et al. (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2023. LNCS, vol. 14220. Springer, Cham (2023)
2. Pushpak, P., Guillaume, J., Zeineb, A. et al.: Weakly supervised joint whole-slide segmentation and classification in prostate cancer. Medical Image Analysis **89**, 102915 (2023)
3. Wu, K., Zheng, Y., Shi, J., et al.: Position-Aware Masked Autoencoder for Histopathology WSI Representation Learning. In: Greenspan, H., et al. (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2023. LNCS, vol. 14225. Springer, Cham (2023)
4. Jiahui, Y., Tianyu, M., Yu, F. , et al.: Local-to-global spatial learning for whole-slide image representation and classification **107**, 102230 (2023)
5. Pierre, C., Eric, W.T., Marc, S.: Classification and Disease Localization in Histopathology Using Only Global Labels: A Weakly-Supervised Approach. ArXiv **abs/1802.02212**, (2018)
6. Das, K., Conjeti, S., Roy, A.G.: Multiple instance learning of deep convolutional neural networks for breast histopathology whole slide classification. In: 15th International Symposium on Biomedical Imaging, pp. 578-581. Washington, DC, USA (2018)
7. Kanavati, F., Toyokawa, G., Momosaki, S., et al. : Weakly-supervised learning for lung carcinoma classification using deep learning. Sci Rep **10**(9297), 9297 (2020)
8. Raswa, F.H., Lu, C.S., Wang, J.C.: Attention-Guided Prototype Mixing: Diversifying Minority Context on Imbalanced Whole Slide Images Classification Learning. In: 2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pp. 7609-7618. Waikoloa, HI, USA (2024)
9. Qu, L., Luo, X., Liu, S., et al.: DGMIL: Distribution Guided Multiple Instance Learning for Whole Slide Image Classification. In: Wang, L., et al. (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2022. LNCS, vol. 13432. Springer, Cham. (2022)
10. Hongrun, Z., Yanda, M., Yitian, Z. et al.: DTFD-MIL: Double-Tier Feature Distillation Multiple Instance Learning for Histopathology Whole Slide Image Classification. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 18780-18790. New Orleans, LA, USA (2022)
11. Schmidt, A., Silva-Rodríguez, J., Molina, R.: Efficient Cancer Classification by Coupling Semi Supervised and Multiple Instance Learning. IEEE Access **10**, 9763-9773 (2022)

12. Wenhao, T., Sheng, H., Xiaoxian, Z, et al.: Multiple Instance Learning Framework with Masked Hard Instance Mining for Whole Slide Image Classification. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 4055-4064. Paris, France (2023)
13. Zhou, Y., Lu, Y.: Multiple Instance Learning with Critical Instance for Whole Slide Image Classification. In: 2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI), pp. 1-5. Cartagena, Colombia (2023)
14. Weiming, H., Xintong, L. Chen, L. et al.: A state-of-the-art survey of artificial neural networks for Whole-slide Image analysis: From popular Convolutional Neural Networks to potential visual transformers. Computers in Biology and Medicine **161**, 107034 (2023)
15. Girolami, I., Pantanowitz, L., Marletta, S. et al. : Artificial intelligence applications for pre-implantation kidney biopsy pathology practice: a systematic review. J Nephrol **35**, 1801–1808 (2022)
16. Weiyang, L., Yandong, W., Zhiding, Y. et al.: SphereFace: Deep Hypersphere Embedding for Face Recognition. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6738-6746. Honolulu, HI, USA (2017)
17. Barz, B., and Denzler, J.: Deep Learning on Small Datasets without Pre-Training using Cosine Loss. In 2020 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 1360-1369. Snowmass, CO, USA (2020)
18. Jiankang, D., Jia, G., Niannan, X., et al. : ArcFace: Additive Angular Margin Loss for Deep Face Recognition. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4685-4694. Long Beach, CA, USA (2019)
19. Ilse, M., Tomczak, J., Welling, M.: Attention-based deep multiple instance learning. In: International conference on machine learning (PMLR), pp. 2127-2136. (2018)
20. Shao, Z., Bian, H., Chen, Y., et al. : Transmil: Transformer based correlated multiple instance learning for whole slide image classification. Advances in neural information processing systems, 34, 2136-2147 (2021)
21. Ma, Z., Collins, M.: Noise contrastive estimation and negative sampling for conditional models: Consistency and statistical efficiency. ArXiv **abs/1809.01812**, (2018)
22. Péter, B., Oscar, G., Quirine, M.: 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 (2019)
23. Lu, M.Y., Williamson, D.F.K., Chen, T.Y., et al. : Data-efficient and weakly supervised computational pathology on whole-slide images. Nature biomedical engineering **5**(6), 555-570 (2021)
24. Soham, R.C., Sidong, L., Tirtharaj, D., et al.: Domain-Specific Pre-training Improves Confidence in Whole Slide Image Classification: In: 2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), pp. 1-4. Sydney, Australia (2023)
25. Hou, W., Yu, L., Lin, C., et al. : H^ 2-MIL: Exploring Hierarchical Representation with Heterogeneous Multiple Instance Learning for Whole Slide Image Analysis. In: the AAAI conference on artificial intelligence, pp. 933-941. (2022)
26. Bontempo, G., Porrello, A., Bolelli, F., et al. : DAS-MIL: Distilling Across Scales for MIL Classification of Histological WSIs. In: Greenspan, H., et al. (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2023. LNCS, vol. 14220. Springer, Cham (2023)
27. Xiong, C., Chen, H., Sung, J.J.Y., et al. Diagnose like a pathologist: Transformer-enabled hierarchical attention-guided multiple instance learning for whole slide image classification. ArXiv **abs/2301.08125**, (2023)

28. Yang, L., Mehta, D., Liu, S., et al. : TPMIL: Trainable Prototype Enhanced Multiple Instance Learning for Whole Slide Image Classification. ArXiv **abs/2305.00696**, (2023)
29. Wang, X., Xiang, J., Zhang, J., et al. : SCL-WC: Cross-slide contrastive learning for weakly-supervised whole-slide image classification. Advances in neural information processing systems **35**, 18009-18021 (2022)