Hard Negative Sample Mining for Whole Slide Image Classification

Wentao Huang*[1], Xiaoling Hu[2], Shahira Abousamra[1], Prateek Prasanna[1], and Chao Chen[1]

[1] Stony Brook University, Stony Brook, NY, USA
[2] Harvard Medical School, Boston, MA, USA

Abstract. Weakly supervised whole slide image (WSI) classification is challenging due to the lack of patch-level labels and high computational costs. State-of-the-art methods use self-supervised patch-wise feature representations for multiple instance learning (MIL). Recently, methods have been proposed to fine-tune the feature representation on the downstream task using pseudo labeling, but mostly focusing on selecting high-quality positive patches. In this paper, we propose to mine hard negative samples during fine-tuning. This allows us to obtain better feature representations and reduce the training cost. Furthermore, we propose a novel patch-wise ranking loss in MIL to better exploit these hard negative samples. Experiments on two public datasets demonstrate the efficacy of these proposed ideas. Our codes are available at https://github.com/winston52/HNM-WSI.

Keywords: Whole Slide Image · Self-Training · Hard Sample Mining.

1 Introduction

Histopathology image analysis serves as the gold standard for cancer diagnosis and treatment [22,24,2]. Due to the large size of the WSI, the heterogeneity of the tumor microenvironment, and the absence of patch-level labels, Multiple Instance Learning (MIL) [13] schemes are often applied to perform a prediction at the whole slide level. In MIL, each slide is considered a bag. A slide is partitioned into patches to create the instances within the bag. One challenge is that only bag-level (i.e. slide-level) labels are available, but not instance-level labels. Specialized training algorithms have been proposed to learn to make instance-level predictions and aggregate them for bag-level prediction [27,14,19].

Feature representation learning. The performance of MIL heavily relies on feature representation of instances (patches). Due to the huge image size, end-to-end learning is computationally infeasible. Earlier work used convolutional neural networks (CNNs) pretrained on ImageNet to generate patch features [15]. These features are then used for downstream MIL. Recently, more advanced self-supervised learning techniques such as SimCLR [9] and DINO [7] have been

* Email: wenthuang@cs.stonybrook.edu.

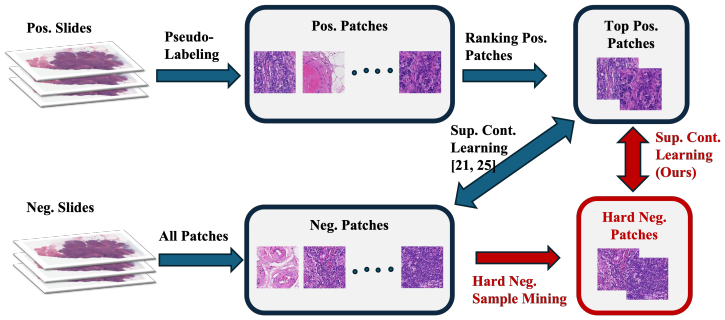

Fig. 1: Feature representation tuning. Previous methods [21,25] perform contrastive learning between top-ranked positive patches and all negative patches. Our method (highlighted in red) selects hard negatives for supervised contrastive learning.

applied to pretrain features using public histopathology image datasets. These features are much more biologically relevant and deliver better performance when combined with MIL [19,8,32,16].

Despite the strong performance of these features, one cannot help but wonder whether they can be further tuned to better fit the downstream prediction task. To this end, new methods [21,25] have been proposed to use supervised contrastive learning to fine-tune the features. Patches are assigned pseudo-labels using a weak patch classifier from the downstream MIL. Supervised contrastive learning is carried out to ensure that patches of the same label are closer to each other in the feature space, and patches of different labels are far away from each other. However, these patch-level pseudo-labels can be noisy, and thus can derail the contrastive learning, leading to deteriorated features. To address this challenge, it was proposed to rank all patches with positive pseudo-labels based on model confidence, and select the top ones for learning. Meanwhile, since all negative slides only contain negative patches, we can just use these patches knowing that they are truly negative patches.

Our contribution: hard negative sampling for patch representation learning. In this paper, we question the design choice of these self-training methods regarding negative sample selection. While it is true that all patches from negative slides are true negative samples, they do not necessarily contribute to learning equally. In particular, we hypothesize that some negative patches are particularly useful in learning. To this end, we propose a novel hard sample mining algorithm to find negative patches that are particularly close to positive patches in feature representation. By focusing on these "hard" negative samples during contrastive learning, we achieve much better patch features for the downstream MIL. Moreover, since we only use a fraction of the negative instances, we are able to reduce the training time considerably. See Figure 1 for illustration.

Indeed, the learning power of these hard negative samples can be further exploited in the downstream MIL. As a second technical contribution, we intro-

duce a novel multiple instance ranking loss that pairwise compares the patch-level classifier's predictions on top positive samples and hard negative samples. By ensuring that the classifier ranks positive and negative patches correctly in terms of their "positiveness", we improve the instance-level classifier and thus the whole-slide-level prediction.

In practice, we perform feature representation tuning and MIL training iteratively to achieve superior performance. Extensive experiments on two public datasets demonstrate the effectiveness of our proposed framework.

2 Related Work

Multiple instance learning in WSI analysis. Multiple Instance Learning [13] (MIL) is a weakly supervised learning framework that can utilize coarse-grained bag labels to train a model when fine-grained instance annotations are not available. The MIL framework for WSI classification is divided into two groups: instance-based and bag-based. The instance-based method first predicts the probability of all instances and then aggregates these to obtain a bag prediction using Mean Pooling or Max Pooling [6,31]. In contrast, the bag-based method involves aggregating the embeddings of all instances into a single bag embedding and then classifying it using a bag classifier. Most current bag-based methods are Attention-based MIL [15,19,26] methods and ViT-based MIL [27,14] methods. Various strategies have been proposed to find positive patches more accurately [23,19,28]. In this paper, we mainly develop an effective and efficient method by focusing on hard negative patches to improve the performance of the WSI classification.

Self-training and pseudo labeling in WSI analysis. Self-training is a widely used technique in semi-supervised learning [18,29,30]. The key idea is to generate pseudo-labels of unlabeled data using a model trained with labeled data and then train the model based on the combination of the labeled data and pseudo labels. In weakly supervised WSI classification, Chen *et al.* [10] proposed a self-training framework and the concept of pseudo labeling to extract the key regions from WSIs. Liu *et al.* [21] proposed a self-paced framework to gradually improve the accuracy of pseudo labels during the training process. However, existing work mainly focuses on pseudo labels from positive slides. Instead of treating all negative patches from negative slides equally as ground truth negative labels, we intend to develop an efficient method to sample part of negative patches based on the pseudo labels from negative slides.

Hard negative sample mining in WSI classification. Hard negative sample mining was first introduced in the object detection task [12], where the main idea is to repeatedly bootstrap negative samples mistakenly classified as false positives. In WSI analysis, Bejnordi *et al.* [3] was the first to enhance model performance on breast cancer by mining difficult negative regions from the training samples. Furthermore, Li *et al.* [20] and Butke *et al.* [5] incorporated hard negative sample mining methods into the MIL framework to improve the performance of the WSI classification task by leveraging attention weights to identify

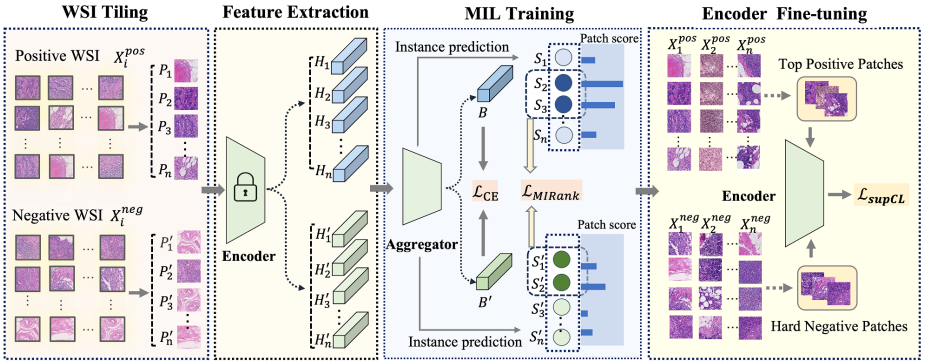

Fig. 2: Overview of our hard negative sample mining framework: WSIs are cut into patches. The encoder generates instance-level features which are aggregated into bag-level features and a pseudo label is assigned to each instance. The multiple instance ranking loss is employed to enhance the accuracy of the pseudo labels. Finally, negative and positive patches are selected based on enhanced pseudo labels to fine-tune the encoder and the process is repeated iteratively.

hard negative instances in false positive bags. Unlike these hard negative mining methods that focus on training a better MIL aggregator, our method utilizes these challenging negative samples to fine-tune the encoder, leading to improved patch-level feature representation.

3 Method

MIL formulation. In the WSI classification task, we are given a dataset D consisting of a set of WSIs $X = \{X_1, X_2, \ldots, X_N\}$ and its corresponding set of slide labels $Y = \{Y_1, Y_2, \ldots, Y_N\}$. Because WSIs are huge size images, each WSI is cut into non-overlapping smaller patches $\{x_{i,1}, x_{i,2} \ldots x_{i,n_i}\}$ where n_i represents the number of patches obtained from X_i. In the setting of MIL, each WSI is considered as a bag, and all patches extracted from the WSI are considered as the instances of the bag. The bag label $Y_i \in \{0, 1\}$ and the instance labels $\{y_{i,1}, y_{i,2} \ldots y_{i,n_i}\}$ are unknown. A bag is labeled as negative only if all its instances are negative, and is labeled positive otherwise.

Method overview. The performance of a WSI classifier is tied to its instance classification performance. The main challenge is the lack of instance-level labels. Additionally, with gigapixel WSIs, the number of instances is huge (in the order of hundreds of thousands for the Camelyon16 [4] dataset), which negatively affects the training time. To improve the accuracy of instance pseudo-label prediction and training efficiency, we propose a negative sampling enhanced self-training MIL framework. Figure 2 provides an overview of our proposed method. Our approach is comprised of two main components. Firstly, we incorporate a

novel multiple instance ranking loss during the training of the aggregator. Subsequently, we design a more efficient strategy for negative patch sampling to fine-tune the encoder. We will next describe these components in more details.

Multiple instance ranking loss. Let $X_i = \{x_{i,1}, x_{i,2} \dots x_{i,n_i}\}$ represents a bag (WSI) and x_{ij} is the j^{th} instance in this bag. After extracting features using encoder f, each patch is projected into instance embedding $h_{ij} = f(x_{ij}) \in \mathbb{R}^{L\times 1}$. An instance classifier converts this embedding into a prediction score $s_{ij} = \phi_{ins}(h_{ij})$, $s_{ij} \in (0, 1)$, where ϕ_{ins} are the weights of the classifier. For the positive bag, the instance-level predicted scores are denoted as $\hat{S}_i^p = \{\hat{s}_{i,1}^p, \hat{s}_{i,2}^p \dots \hat{s}_{i,n_i}^p\}$, and for the negative bag, the instance-level predicted scores are denoted as $\hat{S}_i^n = \{\hat{s}_{i,1}^n, \hat{s}_{i,2}^n \dots \hat{s}_{i,n_i}^n\}$. The instance with the highest prediction score in the positive bag is most likely to be the true positive patch, and the instance with the highest prediction score in the negative bag is the one most similar to a positive patch but is actually negative. This negative instance is considered as a hard instance. To push the scores of positive and negative instances far apart, we propose a novel multiple instance ranking loss that aims to maximize the difference between the sum of scores of the top K instances in the positive and negative bags, respectively. The hinge-based formula of our multiple instance ranking loss is:

$$\mathcal{L}_{MIRank} = \max\left(0, 1 - \frac{1}{K}\sum_{top_K} \hat{s}_{i,k}^p + \frac{1}{K}\sum_{top_K} \hat{s}_{i,k}^n\right) \tag{1}$$

Instance aggregator. To classify the WSI, similar to [19], we first compute the bag embedding as a weighted sum of all instance embeddings. The WSI prediction is then the average of the bag classifier and the instance classifier:

$$\hat{Y_i} = \frac{1}{2}\left(\phi_{ins} h_m + \phi_{bag} \sum_i U(h_i, h_m) h_i\right) \tag{2}$$

where ϕ_{ins} and ϕ_{bag} are the weights of the instance and bag classifiers, respectively. h_m is the embedding of the instance with the highest score and $U(h_i, h_m)$ is the distance between h_m and an instance h_i. Finally, the complete loss function for training the MIL aggregator is given by:

$$\mathcal{L}_{MIL} = w_b * \mathcal{L}_{CE}(\hat{Y_i}, Y_i) + w_r * \mathcal{L}_{MIRank} \tag{3}$$

where $\mathcal{L}_{CE}$ is the cross-entropy loss, w_b and w_r are the weights for the cross-entropy loss and the multiple instance ranking loss, respectively.

Negative sampling enhanced contrastive learning. After each iteration of training the MIL aggregator, we use the trained model to obtain patch-level pseudo labels to fine-tune the encoder. Fine-tuning enables the encoder to learn discriminative representations by pulling together the representations of instances sharing the same pseudo label and pushing apart the representations of instances with different pseudo labels. Let x be the anchor instance, x_s is an instance selected from set $\mathcal{S}_x$ with the same pseudo label as x, and x_d is

an instance selected from set $\mathcal{D}_x$ with a pseudo label different from x. We use supervised contrastive learning as in [17,21] to fine-tune the encoder:

$$\mathcal{L}_{\text{supCL}}(x) = \frac{1}{|\mathcal{S}_x|} \sum_{x_s \in \mathcal{S}_x} -\log \frac{\text{sim}(x, x_s)}{\sum_{x_s \in \mathcal{S}_x} \text{sim}(x, x_s) + \sum_{x_d \in \mathcal{D}_x} \text{sim}(x, x_d)} \tag{4}$$

The similarity score $\text{sim}(x, x')$ is defined as $\exp(f(x) \cdot f(x')/\tau)$, where f is an encoder, and τ is a temperature parameter. The construction of $\mathcal{S}_x$ and $\mathcal{D}_x$ is determined by the pseudo label of x. Let $\mathcal{X}_{pos}$ represent the bank of positive instances, and $\mathcal{X}_{neg}$ represent the bank of negative instances. The construction of $\mathcal{S}_x$ and $\mathcal{D}_x$ is as follows:

$$\text{If } x \in \mathcal{X}_{pos}\text{:} \begin{cases} S_x \leftarrow \mathcal{X}_{pos} \\ D_x \leftarrow \mathcal{X}_{neg} \end{cases}, \qquad \text{If } x \in \mathcal{X}_{neg}\text{:} \begin{cases} S_x \leftarrow \mathcal{X}_{neg} \\ D_x \leftarrow \mathcal{X}_{pos} \end{cases} \tag{5}$$

where $\leftarrow$ represents random sampling from the instance bank. This means that if x is sampled from the positive instance bank $\mathcal{X}_{pos}$, then $\mathcal{S}_x$ and $\mathcal{D}_x$ are constructed by sampling from $\mathcal{X}_{pos}$ and $\mathcal{X}_{neg}$, respectively. Similarly, if x is sampled from $\mathcal{X}_{neg}$, then $\mathcal{S}_x$ and $\mathcal{D}_x$ are constructed by sampling from $\mathcal{X}_{neg}$ and $\mathcal{X}_{pos}$, respectively. Existing methods typically construct $\mathcal{X}_{pos}$ and $\mathcal{X}_{neg}$ as follows: $\mathcal{X}_{pos}$ is the collection of the top $r_p\%$ of positive instances above a preset threshold. $\mathcal{X}_{neg}$ is the collection of all instances in negative slides since, by definition, negative bags only contain negative instances. However, this approach is time-consuming and inefficient for fine-tuning because it includes too many easy negative instances in $\mathcal{X}_{neg}$. Instead, we propose to use hard negative sampling, i.e., construct $\mathcal{X}_{neg}$ from the collection of negative instances with the top $r_n\%$ prediction scores. In this way, training efficiency can be significantly improved by selecting only a fraction of hard negative instances for fine-tuning.

4 Experiment

Datasets. We conduct experiments on two public datasets: Camelyon16 [4] and TCGA-LUAD mutation [1]. Camelyon16 is designed for detecting metastases in lymph node tissue slides. It contains 270 normal slides and 129 tumor slides. The TCGA-LUAD mutation dataset is aimed at detecting gene mutations. We selected four genes related to treatment options: EGFR, KRAS, STK11, and TP53 [21,11]. The dataset comprises 607 WSIs, and the WSI labels indicate whether the corresponding gene is expressed in the slides.

Experiments setup and evaluation metrics. In the WSI preprocessing stage, we cut the slides into non-overlapping patches of size 224×224. For the Camelyon16 dataset, we obtained 0.25 million patches with 5× magnification. For the TCGA-LUAD mutation dataset, we got 0.52 million patches with 10× magnification. We utilize the pre-trained ResNet-18 encoder provided by [19] to extract features for the Camelyon16 and TCGA-LUAD mutation datasets.

Table 1: Main results on Camelyon16 and TCGA-LUAD mutation datasets.

	Camelyon16		TCGA-LUAD mutation			
			EGFR	KRAS	STK11	TP53
Method	ACC	AUC	AUC	AUC	AUC	AUC
Max-pooling	0.8295	0.8641	0.6643	0.5746	0.6702	0.6109
ABMIL[15]	0.8450	0.8653	0.6848	0.5994	0.6784	0.6520
DSMIL[19]	0.8837	0.9095	0.6956	0.6026	0.6885	0.6344
Its2CLR[21]	0.9070	0.9465	0.7103	0.6135	0.7111	0.6703
Ours	**0.9302**	**0.9604**	**0.7235**	**0.6473**	**0.7396**	**0.7071**

We evaluate the performance on WSI classification on Camelyon16 and TCGA-LUAD mutation datasets, in addition to patch-wise classification on Camelyon16 dataset. We report accuracy (ACC) and area under the curve (AUC) evaluation metrics. For Camelyon16 dataset, we reported the results of the official testing set. For TCGA-LUAD mutation dataset, we conducted 5-fold cross-validation on the 607 slides, and the mean and standard deviation of performance metrics are reported. The mean results are presented in Table 1, the detailed results with standard deviation are provided in Tabel 4 in the supplementary material.

Implementation details. When training the MIL aggregator, we follow the settings in [19]. The MIL aggregator was trained for 350 epochs. We employ Adam optimizer with a learning rate of 0.0001. For the multiple instance ranking loss, we set K to 10. The weight of cross-entropy loss w_b and ranking loss w_r are configured to 0.5 and 0.1, respectively. For both the Camelyon16 and TCGA-LUAD mutation datasets, we set the parameters for sampling pseudo labels, r_p and r_n, to 0.2 and 0.05, respectively. The fine-tuning phase was set to 25 epochs. For these hyperparameters, we experiment with different values and select the ones best performing on the validation set. All model training and testing experiments were conducted on Nvidia A5000 GPU.

Quantitative results. Table 1 shows the comparison result on Camelyon16 and TCGA-LUAD mutation datasets. For the Camelyon16 dataset, compared with the classic MIL and self-traning methods, our method achieved the best performance, with an ACC of 0.9302 and an AUC of 0.9604. Furthermore, we also observe improved instance-level prediction accuracy (See Table 2 in the ablation study section). For the TCGA-LUAD mutation dataset, our method achieved the best AUC results over four genes: EGFR reached 0.7235, KRAS reached 0.6473, STK11 reached 0.7396, and TP53 reached 0.7071. The evaluation results show the effectiveness of our proposed framework in bag and instance predictions.

Qualitative results. Figure 3 compares the instance-level prediction in tumor-positive WSIs from the Camelyon16 dataset. Compared to the DSMIL and Its2CLR methods, the instance score maps from our method align best with the ground truth maps. The prediction score for the top negative instances gradually decreases as training progresses. This demonstrates qualitatively that our method enhances the accuracy of instance predictions. The magnified version of Figure 3 is provided in the supplementary materials (Figure 4).

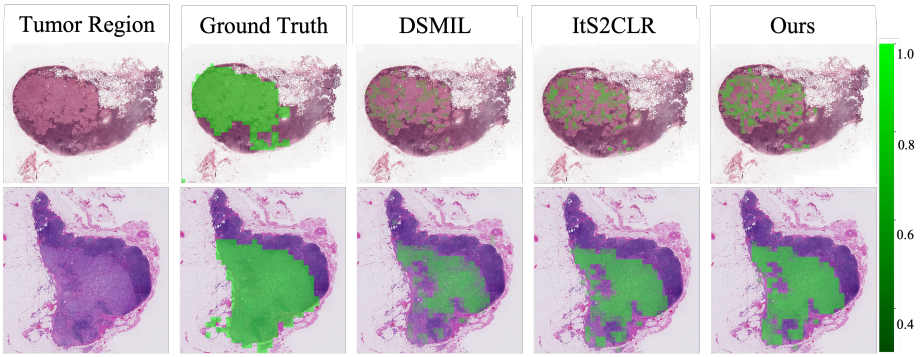

Fig. 3: Visualization of instance prediction probabilities on the Camelyon16 dataset. Patches with probabilities below 0.3 are rendered transparent.

Table 2: Ablation study for our proposed multiple instance ranking loss (MI ranking) with WSI and instance-wise evaluations on Camelyon16 dataset.

	Instances			WSIs	
Method	ACC	AUC	AUPRC	ACC	AUC
DSMIL	0.8941	0.9118	0.8876	0.8837	0.9095
DSMIL + MI ranking	**0.9007**	**0.9176**	**0.8931**	**0.8914**	**0.9151**
ItS2CLR	0.9287	0.9478	0.8974	0.9070	0.9465
ItS2CLR + MI ranking	**0.9291**	**0.9496**	**0.8987**	**0.9147**	**0.9483**
Ours(w/o MI ranking)	0.9341	0.9598	0.9065	0.9225	0.9583
Ours	**0.9374**	**0.9619**	**0.9123**	**0.9302**	**0.9604**

Ablation study: ranking loss. To demonstrate the effectiveness of our multiple instance ranking loss (MI ranking loss), we evaluated the bag-level and instance-level performance on the Camelyon16 dataset before and after integrating our proposed loss into various methods. Table 2 illustrates that incorporating our ranking loss significantly enhances prediction accuracy at both the instance and bag levels for all methods.

Hard negative sample size and training time. We conducted experiments with negative sampling rates of 2%, 5%, 10%, 20%, and 100%. Table 3 illustrates that the optimal performance is achieved at a negative sample ratio of 5% for fine-tuning. Furthermore, we measured the training time per iteration, i.e. the time for training the aggregator, updating the pseudo labels, and fine-tuning the encoder. When the negative sampling rate is 100%, the iteration time is similar to ItS2CLR since fine-tuning dominates the overall training time. However, with our proposed negative sampling strategy, the training time is significantly reduced (around 70% to 80% less time) compared to the ItS2CLR method and with improved performance. These findings demonstrate that our approach can simultaneously enhance performance and training efficiency.

Table 3: Ablation study on proportion of negative sample and training time.

Negative (%)	Camelyon16		Training Time
	ACC	AUC	
2%	90.7	93.12	33 min / iteration
5%	**93.02**	**96.04**	39 min / iteration
10%	92.25	95.76	46 min / iteration
20%	91.47	95.01	72 min / iteration
100%	91.47	94.83	240 min / iteration

5 Conclusion

This work introduces a negative sampling enhanced framework designed to improve performance and training efficiency for self-training frameworks applied to WSI classification tasks. This framework consists of two components: multiple instance ranking loss and negative sampling strategy. The ranking loss enhances instance-level prediction accuracy by differentiating between positive and negative instances, and the negative instance sampling strategy selectively integrates challenging negative samples into the fine-tuning process. Extensive experiments validate the effectiveness and efficiency of our proposed framework.

Acknowledgements: This research was partially supported by the National Science Foundation (NSF) grant CCF-2144901, the National Institute of General Medical Sciences grant R01GM148970, the National Institutes of Health (NIH) grant 5R21CA258493-02, and the Stony Brook Trustees Faculty Award.

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

References

1. The cancer genome atlas program. https://www.cancer.gov/tcga (2019)
2. Barisoni, L., Lafata, K.J., Hewitt, S.M., Madabhushi, A., Balis, U.G.: Digital pathology and computational image analysis in nephropathology. Nature Reviews Nephrology (2020)
3. Bejnordi, B.E., Lin, J., Glass, B., Mullooly, M., Gierach, G.L., Sherman, M.E., Karssemeijer, N., Van Der Laak, J., Beck, A.H.: Deep learning-based assessment of tumor-associated stroma for diagnosing breast cancer in histopathology images. In: ISBI (2017)
4. 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 (2017)
5. Butke, J., Frick, T., Roghmann, F., El-Mashtoly, S.F., Gerwert, K., Mosig, A.: End-to-end multiple instance learning for whole-slide cytopathology of urothelial carcinoma. In: MICCAI Workshop on Computational Pathology (2021)

6. 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 (2019)
7. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: CVPR (2021)
8. 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: CVPR (2022)
9. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: ICML (2020)
10. Chen, Y.C., Lu, C.S.: Rankmix: Data augmentation for weakly supervised learning of classifying whole slide images with diverse sizes and imbalanced categories. In: CVPR (2023)
11. Coudray, N., Ocampo, P.S., Sakellaropoulos, T., Narula, N., Snuderl, M., Fenyö, D., Moreira, A.L., Razavian, N., Tsirigos, A.: Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. Nature medicine (2018)
12. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: CVPR (2005)
13. Dietterich, T.G., Lathrop, R.H., Lozano-Pérez, T.: Solving the multiple instance problem with axis-parallel rectangles. Artificial intelligence (1997)
14. Ding, S., Wang, J., Li, J., Shi, J.: Multi-scale prototypical transformer for whole slide image classification. In: MICCAI (2023)
15. Ilse, M., Tomczak, J., Welling, M.: Attention-based deep multiple instance learning. In: ICML (2018)
16. Kapse, S., Das, S., Zhang, J., Gupta, R.R., Saltz, J., Samaras, D., Prasanna, P.: Attention de-sparsification matters: Inducing diversity in digital pathology representation learning. Medical Image Analysis (2024)
17. Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., Krishnan, D.: Supervised contrastive learning. In: NeurIPS. vol. 33 (2020)
18. Lee, D.H., et al.: Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In: Workshop on challenges in representation learning, ICML (2013)
19. Li, B., Li, Y., Eliceiri, K.W.: Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In: CVPR (2021)
20. Li, M., Wu, L., Wiliem, A., Zhao, K., Zhang, T., Lovell, B.: Deep instance-level hard negative mining model for histopathology images. In: MICCAI (2019)
21. Liu, K., Zhu, W., Shen, Y., Liu, S., Razavian, N., Geras, K.J., Fernandez-Granda, C.: Multiple instance learning via iterative self-paced supervised contrastive learning. In: CVPR (2023)
22. Lu, M.Y., Chen, T.Y., Williamson, D.F., Zhao, M., Shady, M., Lipkova, J., Mahmood, F.: Ai-based pathology predicts origins for cancers of unknown primary. Nature (2021)
23. 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 (2021)
24. Niazi, M.K.K., Parwani, A.V., Gurcan, M.N.: Digital pathology and artificial intelligence. The lancet oncology (2019)

25. Qu, L., Ma, Y., Luo, X., Wang, M., Song, Z.: Rethinking multiple instance learning for whole slide image classification: A good instance classifier is all you need. arXiv preprint arXiv:2307.02249 (2023)
26. Qu, L., Wang, M., Song, Z., et al.: Bi-directional weakly supervised knowledge distillation for whole slide image classification. In: NeurIPS. vol. 35 (2022)
27. 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. In: NeurIPS. vol. 34 (2021)
28. Tang, W., Huang, S., Zhang, X., Zhou, F., Zhang, Y., Liu, B.: Multiple instance learning framework with masked hard instance mining for whole slide image classification. In: CVPR (2023)
29. Wei, C., Sohn, K., Mellina, C., Yuille, A., Yang, F.: Crest: A class-rebalancing self-training framework for imbalanced semi-supervised learning. In: CVPR (2021)
30. Xu, M., Hu, X., Gupta, S., Abousamra, S., Chen, C.: Toposemiseg: Enforcing topological consistency for semi-supervised segmentation of histopathology images. In: ECCV (2024)
31. Zhang, H., Meng, Y., Zhao, Y., Qiao, Y., Yang, X., Coupland, S.E., Zheng, Y.: Dtfd-mil: Double-tier feature distillation multiple instance learning for histopathology whole slide image classification. In: CVPR (2022)
32. Zhang, J., Kapse, S., Ma, K., Prasanna, P., Vakalopoulou, M., Saltz, J., Samaras, D.: Precise location matching improves dense contrastive learning in digital pathology. In: MICCAI (2023)