Clustering-Based Multi-instance Learning Network for Whole Slide Image Classification

Wei Wu, Zhonghang Zhu, Baptiste Magnier, Liansheng Wang

Clustering-based Multi-instance Learning Network for Whole Slide Image Classification

Wei Wu[1], Zhonghang Zhu[1], Baptiste Magnier[2], and Liansheng Wang[1](✉)

[1] Department of Computer Science at School of Informatics, Xiamen University,
Xiamen, China, {weiwu,zzhonghang}@stu.xmu.edu.cn, lswang@xmu.edu.cn
[2] Euromov Digital Health in Motion, Univ Montpellier, IMT Mines Ales, Ales,
France, baptiste.magnier@mines-ales.fr

Abstract. Automated and accurate classification of Whole Slide Image (WSI) is of great significance for the early diagnosis and treatment of cancer, which can be realized by Multi-Instance Learning (MIL). However, the current MIL method easily suffers from over-fitting due to the weak supervision of slide-level labels. In addition, it is difficult to distinguish discriminative instances in a WSI bag in the absence of pixel-level annotations. To address these problems, we propose a novel Clustering-Based Multi-Instance Learning method (CBMIL) for WSI classification. The CBMIL constructs feature set from phenotypic clusters to augment data for training the aggregation network. Meanwhile, a contrastive learning task is incorporated into the CBMIL for multi-task learning, which helps to regularize the feature aggregation process. In addition, the centroid of each phenotypic cluster is updated by the model, and the weights of the WSI patches are calculated by their similarity to the phenotypic centroids to highlight the significant patches. Our method is evaluated on two public WSI datasets (CAMELYON16 and TCGA-Lung) for binary tumor and cancer sub-types classification and achieves better performance and great interpretability compared with the state-of-the-art methods. The code is available at: https://github.com/wwu98934/CBMIL.

Keywords: Whole slide image · Multiple instance learning · Multi-task.

1 Introduction

Whole Slide Images (WSIs) which are digital visualization of tissue section are widely used in disease diagnosis [5,22]. Recently, deep learning approaches have been used in WSI analysis, which is a long-standing challenge due to the gigapixel resolution and the lack of pixel-level annotations [24]. Therefore, the analysis of WSI which is a weakly supervised learning problem usually follows a MIL problem formulation [7,20], where each WSI is regarded as a bag containing many instances that are patches of the WSI.

In previous MIL approaches for WSI analysis, a WSI has been tiled into a large number of small patches and further extracted into features by a pre-trained Convolutional Neural Network (CNN) *e.g.*, ResNet-18 [11]. Then, patch-level features are aggregated, and examined by a classifier that predicts slide-level labels. For aggregation operator, a straightforward method is named pooling, such as mean-pooling and max-pooling [8,27,13]. However, the pooling operation is a handcrafted method that guides limited performance. To address this problem, Ilse *et al.* [12] proposed an attention-based aggregation operator parameterized by deep neural networks, assigning the contribution to each instance for aggregating all instance-level features to a bag-level embedding. Recently, Li *et al.* [16] proposed a non-local attention aggregator that gives the contribution to each instance by the similarity between the highest-score instance and others. Shao *et al.* [23] introduced the self-attention mechanism into the MIL framework which considers the contextual and spatial information between different instances. Notably, WSI contains rich phenotypic information that reflects underlying molecular processes and disease progression. Several studies have shown phenotypic information could provide a convenient visual representation of disease aggressiveness [31,21,29]. Yao *et al.* [29] proposed a MIL framework for survival prediction that considers phenotype clusters as instances instead of patches.

Nevertheless, there are several challenges that exist in developing robust deep MIL models to learn rich representation. First, a positive WSI might contain few disease-positive patches as well as a lot of redundant instances [12,19,16,23,29], leading to the prediction failure of the models due to the weak supervision of the bag-level labels. Second, the model can easily suffer from over-fitting with limited number of training data (WSIs) [16,18] and labels.

To address these challenges, we propose a novel Clustering-Based Multi-Instance Learning (CBMIL) model, which constructs discriminative set from phenotypic clusters that highlight the significant patches of WSI. Meanwhile, a random set is constructed to augment training data for the contrastive task in our multi-task learning module. Hence, the main contributions of our work are summarized as follows:

- A novel clustering-based multi-instance learning model is proposed: it constructs discriminative set by adaptively sampling from phenotypic clusters based on the similarity between instance and phenotypic centroid.
- A mechanism for updating centroid of the phenotypic cluster is designed, which is to calculate the aggregation feature of each phenotypic cluster as the new cluster centroid in each epoch to improve the reliability of prediction.
- The contrastive learning is set as an auxiliary task of the classification task to regularize the feature aggregation process.
- CBMIL is evaluated for WSI classification on two public WSI datasets, namely: CAMELYON16 and TCGA-Lung. Great performances over these datasets and interpretability demonstrate the superiority of the proposed model compared with other state-of-the-art methods.

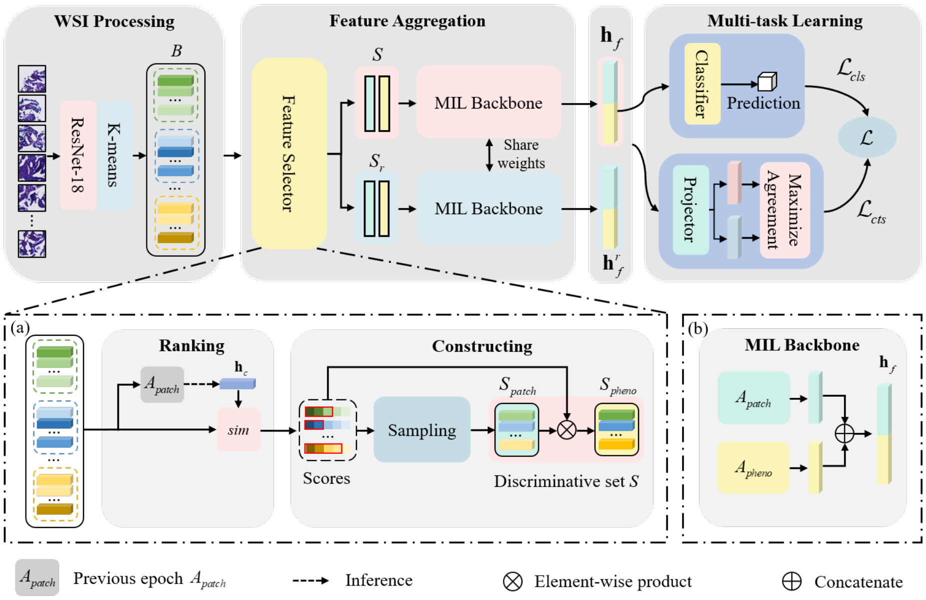


Fig. 1. The pipeline of our method. With the input feature bag, first, a feature selector constructs discriminative set S and random set S_r. Then a MIL backbone encodes the two sets to obtain high-level representations. Finally, the whole model is jointly trained by classification loss $\mathcal{L}_{cls}$ and contrastive loss $\mathcal{L}_{cts}$. (a) represents the construction of discriminative set using the selector, and (b) depicts the framework of the MIL backbone which is consisted of patch-level aggregator A_{patch} and phenotype-level aggregator A_{pheno}.

2 Method

Fig. 1 depicts the overall architecture of our proposed MIL-based framework. Given an input feature bag of a WSI after clustering, two separate sets (*i.e.*, discriminative and random sets) are constructed by a feature selector, then, the selector and a MIL backbone are trained to maximize the agreement of the sets using a contrastive loss. Meanwhile, the discriminative set is involved in classification training, establishing a multi-task learning framework with the contrastive task. Specifically, in Fig. 1(a), where the construction of the discriminative set is illustrated for each training epoch. With the input feature bag, the patch-level aggregator of the previous epoch produces a sequence of centroids for each phenotypic cluster. These centroids are used to select discriminative features based on distance measurement. These discriminative features are aggregated to generate the phenotype-level features to form the discriminative set.

2.1 Clustering-based MIL Framework

As shown in Fig. 1, a clustering-based multi-instance learning framework with multi-task learning is built for WSI classification, in which a feature selector is used to construct discriminative set fed into the MIL backbone to obtain the high-level representation (see respectively Figs. 1(a) and (b)). Then, the representation $\mathbf{h}_f$ is used to generate bag-level prediction which will be used to calculate the cross-entropy loss with the slide-level ground truth labels. Also, a small neural network projector that maps $\mathbf{h}_f$ and $\mathbf{h}_f^r$ to the latent space where contrastive loss is applied.

Let $B = \{\mathbf{B}_i\}_{i=1}^C$ denotes a bag of the clustered features of a WSI, where C is the number of clusters, $\mathbf{B}_i = \{\mathbf{x}_{i,j}\}_{j=1}^{N_i}$ is the i^{th} phenotypic cluster that consists of patch features $\mathbf{x}_{i,j} \in \mathbb{R}^{L\times 1}$ extracted by pre-trained ResNet-18 [11] from image patches, where N_i is the number of patches of i^{th} cluster could vary for different clusters and L is the dimension of the patch feature.

As detailed in Fig. 1(a), a discriminative set is generated by the two following processes: ranking and constructing. In the ranking phase, different non-local attention scores are assigned to patches within each cluster respectively. In a phenotypic cluster, the score of a patch is obtained based on the similarity of the patch feature to the centroid $\mathbf{h}_c$ of the cluster. The centroid is inferred from the patch-level aggregator A_{patch} of the previous epoch during training. Given a phenotypic cluster $\mathbf{B}_i = \{\mathbf{x}_{i,j}\}_{j=1}^{N_i}$, the score $r_{i,j}$ of the j^{th} patch can be formulated as:

$$r_{i,j} = \frac{\exp(\langle \mathbf{W}_q \mathbf{h}_{c,i}, \mathbf{W}_q \mathbf{x}_{i,j}\rangle)}{\sum_{k=1}^{N_i} \exp(\langle \mathbf{W}_q \mathbf{h}_{c,i}, \mathbf{W}_q \mathbf{x}_{i,k}\rangle)}, \tag{1}$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product of two vectors, and $\mathbf{W}_q$ is a weight matrix of fully-connected layer. In constructing phase, N patches are sampled with top scores from all phenotypic clusters, given by:

$$\mathbf{B'}_i = T(\mathbf{B}_i; K_i, \mathbf{r}_i), \quad K_i = \left[N_i \times \frac{N}{\sum_{k=1}^{C} N_k} \right], \tag{2}$$

where T is the top-k operation of choosing patches from $\mathbf{B}_i$ according to the scores $\mathbf{r}_i = \{r_{i,j}\}_{j=1}^{N_i}$. Also, K_i denotes the number of chosen patches in the i^{th} cluster. Then, compose all the $\mathbf{B'}_i$ to get the subset of WSI $S_{patch} = \{\mathbf{x}_n\}_{n=1}^N$. Furthermore, the phenotype-level feature is aggregated by the scores within each sampled phenotypic cluster, is given by $\mathbf{x}_i^p = \sum_{j=1}^{K_i} r_{i,j}\mathbf{W}_v\mathbf{x}_{i,j}$, where $\mathbf{W}_v$ is a weight matrix used to transform $\mathbf{x}_{i,j} \in \mathbf{B'}_i$ into an information vector. The phenotype-level features are represented as $S_{pheno} = \{\mathbf{x}_i^p\}_{i=1}^C$. Finally, S_{patch} and S_{pheno} together form the discriminative set $S = \{S_{patch}, S_{pheno}\}$. Notably, the only difference between the construction of random set and discriminative set is that the features in the random set are sampled randomly and do not depend on the attention scores. As S_{patch} is sampled from the phenotypic cluster whose patches are uniformly distributed in WSI and which has the same proportion of phenotypic features as the WSI. As A_{patch} is updated during training, the selector can sample the more informative patches from each phenotypic cluster, and the model benefits from the selector as well.

Meanwhile, the network of our MIL backbone includes two feature aggregators: A_{patch} and A_{pheno}, as shown in Fig. 1(b). These two aggregators encode the constructed WSI set to patch-level and phenotype-level features which are concatenated to obtain the high-level representation of WSI. Given a WSI set $S = \{S_{patch}, S_{pheno}\}$, the fused representation $\mathbf{h}_f$ is given by:

$$\mathbf{h}_f = Cat(A_{patch}(S_{patch}), A_{pheno}(S_{pheno})), \tag{3}$$

where Cat is a concatenation operator. With the two aggregators and concatenation operator, the MIL backbone generates a high-level representation of WSI, providing rich information for following the multi-task learning module.

2.2 Multi-task Learning

In this sub-section, a multi-task learning module is detailed, it is designed to improve the representational power of our model and mitigate over-fitting as shown in Fig. 1. Inspired by recent contrastive algorithms [2,3,10,9,4], we propose an auxiliary contrastive task based on our adaptive selector and MIL backbone to update our model together with the classification task.

The contrastive algorithm learns representations by maximizing agreement between differently augmented views of the same data example via a contrastive loss in the latent space [2]. The two different views of a sample are generated by a stochastic data augmentation module in previous works. Different from this, in CBMIL, a discriminative set S and a random set S_r are generated by the proposed feature selector from the same bag B, as shown in Fig. 1.

Then, the two sets from the same WSI bag as a positive pair will be transformed into two representations $\mathbf{h}_f$ and $\mathbf{h}_f^r$ by our MIL backbone. Then, the representations are mapped to vectors in latent space and *NT-Xent* contrastive loss is applied to maximize their agreement. In addition, $\mathbf{h}_f$ is also used for the classification task, which is trained using a standard cross-entropy loss. The total loss $\mathcal{L}$ for a given mini-batch WSIs is the weighted sum of both the contrastive loss $\mathcal{L}_{cts}$ and classification loss $\mathcal{L}_{cls}$, given by $\mathcal{L} = \beta \cdot \mathcal{L}_{cts} + (1 - \beta) \cdot \mathcal{L}_{cls}$, where $\beta \in [0, 1]$ is a scalar for scaling.

2.3 Model Structure and Training Procedure

In the proposed MIL backbone, the aggregator could be an arbitrary MIL-based model that satisfies the permutation-invariant MIL formulation, such as in [12,19,16]. We use CLAM-SB [19], a solid MIL aggregator, as our A_{patch} to aggregate the sampled features of WSI, and, A_{pheno}, a simple gated attention [12] is used to aggregate the phenotype-level features. As denoted in Eq. (2), N is a constant number that denotes the number of selected patch features. For a few WSIs with patches less than N, we will pad the bag with 0 vectors.

Stochasticity is important in contrastive learning, previous works [2,10,9,4] usually use stronger data augmentation on images. But the WSI bag is a feature-level data sample, consequently, the natural data augmentation methods are not available. To address this problem, we apply Mixup [30] based data interpolation for S_{patch} inspired by [26]. The Mixup operation is only used during the training phase. Given a mini-batch of M bags $\mathcal{B} = \{\mathbf{B}_m\}_{m=1}^{M}$ with the same constant shape, the augmented sample for a $\mathbf{B}$ is created by taking its random interpolation with another randomly chosen sample $\tilde{\mathbf{B}}$ from $\mathcal{B}$, formulated as:

$$\mathbf{B}^{+} = \lambda \cdot \mathbf{B} + (\mathbf{1} - \lambda) \cdot \tilde{\mathbf{B}}, \tag{4}$$

where λ is a coefficient sampled from a uniform distribution $\lambda \sim U(\alpha, 1.0)$. The value of α is usually high such as 0.9. It means that $\mathbf{B}^{+}$ is closer to $\mathbf{B}$ than $\tilde{\mathbf{B}}$, and the $\tilde{\mathbf{B}}$ could be thought of as a data noise being added.

In the inference step, we throw away the contrastive branch and the generated random set, and keep only the discriminative set for predicting the WSI label.

3 Experiments and Results

In this section, the implementation of the proposed method is detailed; also, experiments and results are reported. Our experiments are conducted on two public datasets: CAMELYONG16 [1] and the lung cancer dataset of The Cancer Genome Atlas (TCGA-Lung) [25].

3.1 Dataset and Evaluation Metrics

CAMELYON16 is a widely used public dataset for metastasis detection in breast cancer, including 270 training WSIs and 129 test WSIs. TCGA-Lung consists of two subtype projects, *i.e.*, Lung Squamous Cell Carcinoma (TGCA-LUSC) and Lung Adenocarcinoma (TCGA-LUAD), which contains 529 LUAD WSIs and 512 LUSC WSIs.

For all WSIs in both datasets, tissue segmentation of the WSI was performed by applying a combination of filters [28]. Each WSI is tiled into a series of 256×256 patches without overlap at 20× magnification, where the background patches (tissue region $< 35\%$) are discarded. After pre-processing, CAMELYON16 yields about 6881 patches per WSI, and TCGA-Lung yields about 11540 patches per WSI. As in [16], the feature of each patch is embedded in a 512-dimensional ($L = 512$, L is defined at the beginning of the Sec. 2.1) vector by a ResNet-18 [11] model pre-trained by [16]. Then, we adopt K-means algorithm to cluster patch features into $C = 10$ phenotypic cluster to form bag B, following [29].

Regarding CAMELYON16 dataset, the training set is done after splitting the 270 WSIs into approximately 80% training and 20% validation and tested on the official test set. For TCGA-Lung, we randomly split the data in the ratio of training:validation:test = 60:15:25. For evaluation metrics, the accuracy, Area Under Che curve (AUC) scores and F1-score are reported in Sec. 3.3 on both datasets. The average results are obtained by 4-fold cross-validation on TCGA-Lung dataset.

Table 1. Results on CAMELYON16 and TCGA-Lung, respectively.

	CAMELYON16			TCGA-Lung		
	Accuracy	AUC	F1-score	Accuracy	AUC	F1-score
MinMax [6]	0.8504	0.8757	0.7800	0.8373	0.9088	0.8396
ABMIL [12]	0.8640	0.8939	0.7988	0.8457	0.9073	0.8419
ABMIL-Gated [12]	0.8550	0.8766	0.7833	0.8468	0.9078	0.8426
SetTransformer [15]	0.7775	0.8493	0.7415	0.6758	0.7800	0.7176
DeepAttnMISL [29]	0.8791	0.9213	0.8236	0.7992	0.8744	0.79506
CLAM-SB [19]	0.8713	0.8926	0.8107	0.8687	0.9412	0.8697
CLAM-MB [19]	0.8508	0.8938	0.7866	0.8661	0.9420	0.8660
DSMIL [16]	0.8682	0.8832	0.7952	0.8597	0.9300	0.8590
CBMIL (ours)	**0.9380**	**0.9541**	**0.9184**	**0.8849**	**0.9429**	**0.8853**

3.2 Implementation Details

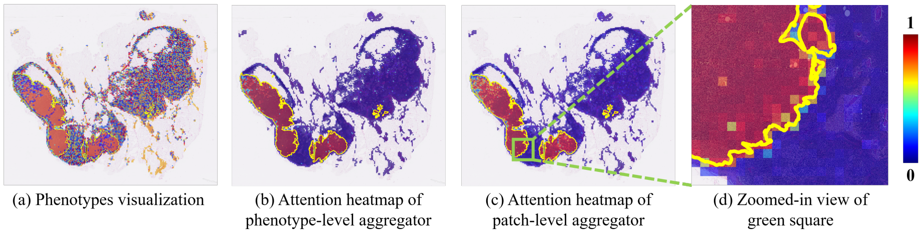
The number of sampled patches N is experimentally set to 1024. In the training step, we use Adam [14] optimizer with an initial learning rate of 0.0001, a cosine annealing (without warm restarts) scheme for learning rate scheduling [17], and a mini-batch size of 16. The parameter α of Mixup is set to 0.8, the temperature parameter τ defined in *NT-Xent* loss [2] is set to 1.0, and the loss scaling parameter β is set to 0.1. The classifier and projector are two Multilayer Perceptron (MLP) with one hidden layer, where the classifier calculates the prediction scores and the projector maps the representations to a 128-dimensional latent space.

3.3 Experimental Results

To demonstrate the performance of our model, we first compare our proposed model with the current state-of-the-art deep MIL models [6,12,15,29,19,16]. All the results are provided in Table 1. In CAMELYON16, only a small portion of regions in a positive slide contains tumor (roughly $<$ 10% of the total tissue area per slide) which leads to the positive bags being highly unbalanced. CBMIL outperforms its A_{patch} CLAM-SB [19] (*i.e.*, 5% and 6% higher in accuracy and AUC) and other deep MIL-based models. In TCGA-Lung, a positive slide contains a relatively larger area of tumor region (roughly $>$ 80% of the total tissue area per slide). CBMIL also outperforms all the other methods. Overall, the results demonstrate the superiority of our CBMIL model.

Table 2. Effects of the adaptive sampling and multi-task module.

Method	A	M	CAMELYON16			TCGA-Lung		
			Accuracy	AUC	F1-score	Accuracy	AUC	F1-score
Random MIL			0.8915	0.9173	0.8409	0.8626	0.9344	0.8603
Adaptive MIL	✓		0.9302	0.9438	0.9032	0.8769	0.9301	0.8755
CBMIL	✓	✓	**0.9380**	**0.9541**	**0.9184**	**0.8849**	**0.9429**	**0.8853**



(a) Phenotypes visualization (b) Attention heatmap of phenotype-level aggregator (c) Attention heatmap of patch-level aggregator (d) Zoomed-in view of green square

Fig. 2. The visualization of phenotypes and attention heatmaps: (a) is the visualization of phenotypes of a WSI from CAMELYON16 testing set, (b) and (c) are heatmaps of attention weights in aggregators. Note: for (b) and (c), attention weights are re-scaled from min-max to [0, 1] and used for patch intensities. (The details and colors are better seen by zooming on a computer screen.)

In addition, to further determine the effect of the adaptive sampling mechanism and multi-task module combined with contrastive learning, we report ablation study results as shown in Table 2. This table shows the experimental results of whether our proposed model has adaptive sampling and multi-task module. Here, the A indicates whether to sample patch features based on the attention scores in the feature selector, and the M indicates whether to add contrastive learning branch in the training phase to establish multi-task learning. It could be noted that the performance of classification can be substantially improved by the adaptive sampling mechanism, and the performance can be further improved by adding the multi-task learning module.

In closing, we also show the interpretability of CBMIL as displayed in Fig. 2. The yellow curve depicts the official pixel-level annotation of the tumor region in CAMELYON16. Fig. 2(a) allows the visualization of phenotypic clusters after clustering, where each color represents a cluster, and it can be noticed that the phenotypic cluster of the tumor region are very obvious, while other normal tissues are uniformly distributed throughout the WSI. It is remarkable in Fig. 2(b) that the phenotypic clusters belonging to the tumor region are given high weights. Finally, Fig. 2(c) shows a more fine-grained attention heatmap: the boundaries of which can highly overlap with the labeled regions. These visualization results demonstrate the reliable interpretability of our proposed model.

4 Conclusion

In this paper, a novel Clustering-Based Multi-Instance Learning framework (CBMIL) is proposed for weakly supervised classification of Whole Slide Image (WSI). Firstly, we design a feature selector that constructs discriminative set of WSI from phenotypic clusters by sampling patches based on centroids. The centroids are updated during training and are used to sample patches that are highly correlated with the prediction results. In addition, with the representational power of contrastive learning, we integrate contrastive learning task directly into MIL, establishing a multi-task learning framework to improve the performance of our method. Meanwhile, a Mixup operator is introduced for feature-level data augmentation. Most importantly, the proposed method outperforms the state-of-the-art MIL algorithms in terms of accuracy, AUC and F1-score over two public datasets, namely: CAMELYON16 and TCGA-Lung. Eventually, CBMIL can provide great interpretability by visualizing the attention weights in the MIL backbone.

References

1. 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)
2. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International conference on machine learning. pp. 1597–1607. PMLR (2020)
3. Chen, X., Fan, H., Girshick, R., He, K.: Improved baselines with momentum contrastive learning. arXiv preprint arXiv:2003.04297 (2020)
4. Chen, X., He, K.: Exploring simple siamese representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15750–15758 (2021)
5. Cornish, T.C., Swapp, R.E., Kaplan, K.J.: Whole-slide imaging: routine pathologic diagnosis. Advances in anatomic pathology **19**(3), 152–159 (2012)
6. Courtiol, P., Tramel, E.W., Sanselme, M., Wainrib, G.: Classification and disease localization in histopathology using only global labels: A weakly-supervised approach. arXiv preprint arXiv:1802.02212 (2018)
7. Dietterich, T.G., Lathrop, R.H., Lozano-Pérez, T.: Solving the multiple instance problem with axis-parallel rectangles. Artificial intelligence **89**(1-2), 31–71 (1997)
8. Feng, J., Zhou, Z.H.: Deep miml network. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 31 (2017)
9. Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al.: Bootstrap your own latent-a new approach to self-supervised learning. Advances in Neural Information Processing Systems **33**, 21271–21284 (2020)
10. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9729–9738 (2020)
11. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (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. Kanavati, F., Toyokawa, G., Momosaki, S., Rambeau, M., Kozuma, Y., Shoji, F., Yamazaki, K., Takeo, S., Iizuka, O., Tsuneki, M.: Weakly-supervised learning for lung carcinoma classification using deep learning. Scientific reports **10**(1), 1–11 (2020)
14. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)

15. Lee, J., Lee, Y., Kim, J., Kosiorek, A., Choi, S., Teh, Y.W.: Set transformer: A framework for attention-based permutation-invariant neural networks. In: International Conference on Machine Learning. pp. 3744–3753. PMLR (2019)
16. 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)
17. Loshchilov, I., Hutter, F.: Sgdr: Stochastic gradient descent with warm restarts. arXiv preprint arXiv:1608.03983 (2016)
18. Lu, M.Y., Chen, R.J., Wang, J., Dillon, D., Mahmood, F.: Semi-supervised histology classification using deep multiple instance learning and contrastive predictive coding. arXiv preprint arXiv:1910.10825 (2019)
19. 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)
20. Maron, O., Lozano-Pérez, T.: A framework for multiple-instance learning. Advances in neural information processing systems **10** (1997)
21. Mobadersany, P., Yousefi, S., Amgad, M., Gutman, D.A., Barnholtz-Sloan, J.S., Vega, J.E.V., Brat, D.J., Cooper, L.A.: Predicting cancer outcomes from histology and genomics using convolutional networks. Proceedings of the National Academy of Sciences **115**(13), E2970–E2979 (2018)
22. Pantanowitz, L., Valenstein, P.N., Evans, A.J., Kaplan, K.J., Pfeifer, J.D., Wilbur, D.C., Collins, L.C., Colgan, T.J.: Review of the current state of whole slide imaging in pathology. Journal of pathology informatics **2**(1), 36 (2011)
23. 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** (2021)
24. Srinidhi, C.L., Ciga, O., Martel, A.L.: Deep neural network models for computational histopathology: A survey. Medical Image Analysis **67**, 101813 (2021)
25. Tomczak, K., Czerwińska, P., Wiznerowicz, M.: The cancer genome atlas (tcga): An immeasurable source of knowledge. wspolczesna onkol. 2015; 1a: A68–a77 (2014)
26. Verma, V., Luong, T., Kawaguchi, K., Pham, H., Le, Q.: Towards domain-agnostic contrastive learning. In: International Conference on Machine Learning. pp. 10530–10541. PMLR (2021)
27. Wang, X., Yan, Y., Tang, P., Bai, X., Liu, W.: Revisiting multiple instance neural networks. Pattern Recognition **74**, 15–24 (2018)
28. Yamashita, R., Long, J., Saleem, A., Rubin, D.L., Shen, J.: Deep learning predicts postsurgical recurrence of hepatocellular carcinoma from digital histopathologic images. Scientific reports **11**(1), 1–14 (2021)
29. Yao, J., Zhu, X., Jonnagaddala, J., Hawkins, N., Huang, J.: Whole slide images based cancer survival prediction using attention guided deep multiple instance learning networks. Medical Image Analysis **65**, 101789 (2020)
30. Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D.: mixup: Beyond empirical risk minimization. arXiv preprint arXiv:1710.09412 (2017)
31. Zhu, X., Yao, J., Zhu, F., Huang, J.: Wsisa: Making survival prediction from whole slide histopathological images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7234–7242 (2017)