Weakly supervised histopathological image representation learning based on contrastive dynamic clustering

Jun Li[a], Zhiguo Jiang[a,b], Yushan Zheng[b,*], Haopeng Zhang[a,b], Jun Shi[d], Dingyi Hu[a,b], Wei Luo[a,b], Zhongmin Jiang[c], and Chenghai Xue[e]

[a]Image Processing Center, School of Astronautics, Beihang University, Beijing 102206, China
[b]Beijing Advanced Innovation Center for Biomedical Engineering, Beihang University, Beijing 100191, China
[c]Department of Pathology, Tianjin Fifth Central Hospital, Tianjin 300450, China
[d]School of Software, Hefei University of Technology, Hefei 230601, China
[e]Wangkangyuan Tianjin Gene Technology, Inc, Tianjin, 300220, China
[*]Corresponding author, E-mail: yszheng@buaa.edu.cn

ABSTRACT

Feature representations of histopathology whole slide images (WSIs) are crucial to the downstream applications for computer-aided cancer diagnosis, including whole slide image classification, region of interest detection, hash retrieval, prognosis analysis, and other high-level inference tasks. State-of-the-art methods for whole slide image feature extraction generally rely on supervised learning algorithms based on fine-grained manual annotations, unsupervised learning algorithms without annotation, or directly use pre-trained features. At present, there is a lack of research on weakly supervised feature learning methods that only utilize WSI-level labeling. In this paper, we propose a weakly supervised framework that learns the feature representations of various lesion areas from histopathology whole slide images. The proposed framework consists of a contrastive learning network as the backbone and a designed contrastive dynamic clustering (CDC) module to embedding the lesion information into the feature representations. The proposed method was evaluated on a large scale endometrial whole slide image dataset. The experimental results have demonstrated that our method can learn discriminative feature representations for histopathology image classification and the quantitative performance of our method is close to the fully-supervision learning methods. The code is available at https://github.com/junl21/cdc.

Keywords: Weakly supervised learning, Representation learning, Histopathological image analysis, Clustering

1. INTRODUCTION

Local representation extraction based on convolutional neural networks (CNNs) has become essential in the recent studies for WSI analysis.[1–3] However, Learning good representations of different lesion tissues is a challenging task, as the structure and the morphology of WSIs are complex.

Most of the previous methods for WSI analysis are based on fully supervised paradigm,[4–8] which depends on the manual annotations of pathologists. Although the supervised methods perform well in various downstream tasks, the cost of pathological image annotation has become its development bottleneck. Recently, self-supervised representation learning methods are widely developed, which can alleviate the above problems.[9–11] In the field of pathological image analysis, there have been some works using self-supervised learning to pre-train the feature extractor.[12–14] However, the representations obtained by these self-supervised methods are difficult to distinguish the subtle inter-class morphology.

In this situation, the weakly supervised methods are introduced to the domain of histopathology image representation learning.[15–17] Most current weakly supervised methods[17–19] are based on multi-instance learning (MIL)[20] and depend on the quality of the pre-trained representations. Moreover, the methods based on MIL generally focus on binary classification tasks. There is a lack of research on weakly supervised representation learning based on multi-class WSI labels. Typically, Lu et al.[3] uses representations pre-trained on ImageNet[21] and MIL method to classify WSIs. However, the features pre-trained on the ImageNet are not discriminative enough to represent the morphological and structural differences of tissues. In addition, this weakly supervised method only provides a *positive vs. negative* prediction for each patch but cannot classify the subtypes of the positive patches.

In response to the above problems, we propose a weakly supervised histopathological image representation learning based on the framework proposed in ”Bootstrap Your Own Latent” (BYOL)[11] and a designed dynamic clustering module (CDC). BYOL is used to initialize the image representations. The CDC module is proposed to enable the weak-supervision task to be trained as the full-supervision paradigm. Experiments on a endometrial dataset have proven the effectiveness of the proposed and shown that the proposed method is competitive to the full-supervision method.

2. MATERIALS AND METHODS

2.1 Dataset Setup

Table 1. The number of different types of slides in our dataset.

	Normal	WDEA	MDEA	LDEA	SEIC
Train	18	186	166	74	42
Test	9	81	71	32	19

The experimental dataset contains 698 histopathological WSIs of endometrial cases, which includes 5 categories, namely Normal, well-differentiated endometrioid adenocarcinoma (WDEA), moderately-differentiated endometrioid adenocarcinoma (MDEA), lowly-differentiated endometrioid adenocarcinoma (LDEA) and Serous endometrial intraepithelial carcinoma (SEIC). The dataset was randomly split into training and testing parts following the ratio of 7:3 at the WSI-level, and the details are shown in Table 1. All the WSIs are stained by H&E and scanned at X40 by PRECICE 500 . Each WSI is manually annotated by the pathologists. Then, the images from the annotated regions are regarded as positive and the others are negative in this paper. We crop these WSIs into non-overlapping image patches in size of 256×256 pixels. For weakly supervised method, we randomly crop 700 patches from each WSI. For fully supervised method, we crop 500 positive patches which contain at least 70% positive pixels and 200 negative patches which contain 100% negative pixels from each WSI.

In this paper, we address the weakly supervised problem for patch representation learning with only the WSI-level labels. Specifically, we assign a pseudo-label for the patches from a training WSI as the same of the WSI. In this case, the pseudo-labels for the patches from the annotated regions are consistent to the true labels of these patches and pseudo-labels for the negative patches are opposite to the true labels, which are then corrected in our method. And for the test set, the patches are labeled based on the pathologists’ annotations for quantitative evaluation.

2.2 Method

The overview of the proposed framework is illustrated in Figure 1. The pipeline consists of two steps: pre-training and fine-tuning. In the pre-training step, we apply BYOL[11] to initializing the feature representations of patches. In the fine-tuning step, we use the proposed contrastive dynamic clustering (CDC) module to embedding the semantic information into the feature representations. The details of the two steps are described as following.

Pre-training step. We introduce BYOL to pre-train a network to initialize representations of the patches. It’s a siamese network consisting of two branches, namely an online network and a target network. The online network is composed of an encoder f_θ, a projector g_θ and a predictor q_θ, which are defined by a set of trainable weights θ. Correspondingly, the target network contains an encoder f_ξ and a projector g_ξ that share the same structures of f_θ and q_θ but are determined by a set of weights ξ.

In the training stage, two different augmented views for a patch x are fed into the online network and the target network, respectively. Then, an L2 loss is built between the outputs of the two branches to train the

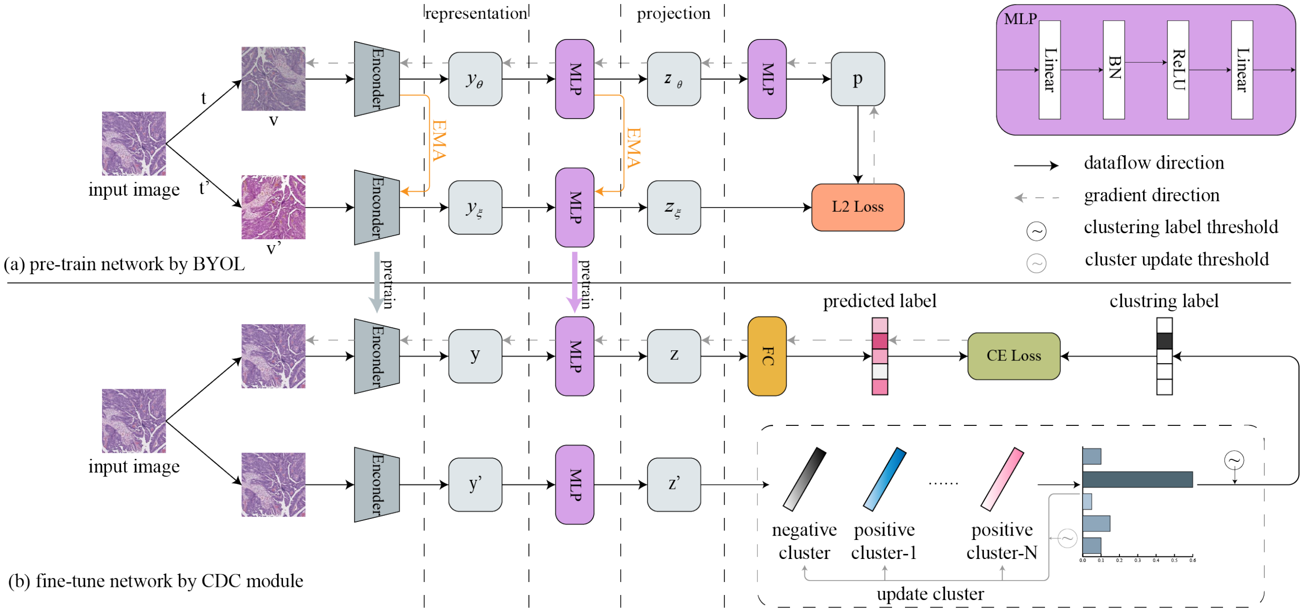


Figure 1. The overview of our proposed framework, where (a) is the pre-train step, which aims to learning image representations by BYOL, and (b) is the fine-tune step, which aims to embedding the lesion information to image representations by CDC module.

weights θ of the online network. The target network is updated by the exponential moving average (EMA) mechanism, as illustrated in Figure 1(a).

Fine-tuning step. Afterwards, we append a classification layer on the end of g_θ and use the proposed CDC module to fine-tune the f_θ and g_θ. The motivation of CDC is that the average distance from a positive patch to the patches with the same pseudo-label should be closer than those with the other pseudo-labels. Based on this hypothesis, we design a clustering mechanism to correct the pseudo-labels, i.e., to identify the negative patches with false positive pseudo-labels.

Specifically, we build a cluster based on the output of g_ξ for each category from the same category of WSIs. As shown in Figure 1(b), letting $\mathbf{z}'_i$ denote the output of g_ξ for the i-th patch and $\bar{\mathbf{z}}'_k$ denote the center for the k-th cluster, $k = 0, 1, ..., C-1$, the correction of the pseudo-label $\tilde{y}_i$ is formulated as

$$y_i = \begin{cases} 0 & \tilde{y}_i = 0 \ or \ s_{i0} > T_{neg} \\ \tilde{y}_i & otherwise, \end{cases}, \quad s_{ik} = e^{d_{ik}} / \sum\nolimits_j e^{d_{ij}}, \quad d_{ik} = \langle \mathbf{z}'_i, \bar{\mathbf{z}}'_k \rangle, \tag{1}$$

where d_{ik} is the similarity between the i-th patch and k-th cluster, s_{ik} is the probability the patch belongs to the k-th class, and s_{i0} is the probability the patch belongs to the negative class (normal tissue). A high probability of s_{i0} indicates the patch is close to the negative cluster. Therefore, a threshold T_{neg} is used to correct the pseudo-label to be negative, i.e., $y_i = 0$.

In our method, the k-th cluster is defined by a set $\mathbb{C}_k = \{\mathbf{z}'_i\}$ and implemented by a queue in a consistent length. Correspondingly, the cluster center $\bar{\mathbf{z}}'_k$ is dynamically calculated in each step of training by equation $\bar{\mathbf{z}}'_k = 1/|\mathbb{C}_k| \sum_{\mathbf{z}'_i \in \mathbb{C}_k} \mathbf{z}'_i,$. To ensure $\bar{\mathbf{z}}'_k$ to be consistently representative to the corresponding class during the optimization of the network, we proposed to update $\mathbb{C}_k$ in each step of training. Specifically, for the k-th cluster, a set of samples $\mathbb{C}_k^+$ are recognized from the mini-batch. The representations in $\mathbb{C}_k^+$ are pushed to the cluster queue $\mathbb{C}_k$ and simultaneously pop $|\mathbb{C}_k^+|$ oldest samples from $\mathbb{C}_k$. Correspondingly, $|\mathbb{C}_k^+|$ is recognized by the following strategy

$$\mathbb{C}_k^+ = \begin{cases} \{\mathbf{z}'_i | s_{i0} > T_{neg}\} & k = 0 \\ \{\mathbf{z}'_i | s_{i0} < (1 - T_{pos})\} & otherwise \end{cases} \tag{2}$$

where T_{pos} is threshold to filter the representative positive patches. Particularly, $\mathbb{C}_k$ is filled by the representations extracted by the pre-trained BYOL's target network during the early steps of the training.

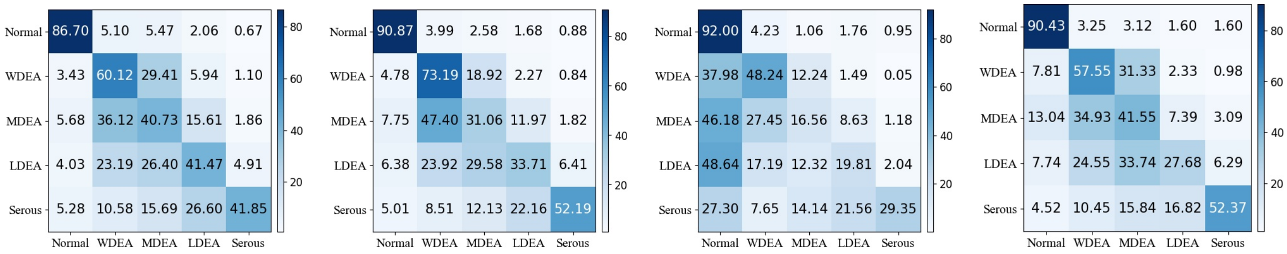
After correction, y_i is used as the label of i-th patch to fine-tune the model based on cross-entropy loss function. Finally, the encoder f_θ and the predictor g_θ as our feature extractor to obtain the image representations.

3. RESULTS

The ResNet50[22] is used as the baseline to evaluate the effectiveness of the proposed representation learning method in the downstream task of histopathological image classification. We compared our model with three methods: 1) train the ResNet50[22] by full-supervised learning, 2) pre-train the ResNet50 by BYOL[11] and fine-tune it by full-supervised learning, 3) The method proposed by Lerousseau et al.[15] It should be noted that we modified the method proposed by Lerousseau et al.[15] from a binary task to a multi-class task by performing the threshold on the negative probability.

Table 2. Comparison of different representation learning methods for the task of histopathology image classification, where the best result in each column is printed in bold.

Methods	Label level	Multi-class task		Binary task	
		Accuracy	AUC	Sensitivity	Specificity
Full-supervision[22]	Patch	0.595	0.819	**0.955**	0.867
BYOL+fine-tuning[11]	Patch	**0.618**	**0.861**	0.939	0.909
Lerousseau et al.[15]	WSI	0.482	0.721	0.584	**0.920**
BYOL+CDC (Ours)	WSI	0.594	0.832	0.907	0.904



(a) Full-supervision (b) BYOL+fine-tuning (c) Lerousseau et al. (d) BYOL+CDC

Figure 2. The confusion matrix for different methods.

The results are presented in Table 2. Overall, the proposed method, which only depends on the WSI-level labels, achieved a competitive performance to the ResNet directly trained by patch-level labels. The result demonstrates the proposed CDC module has effectively corrected the false pseudo-labels, which enables the weak-supervision task to be trained as the full-supervision paradigm. It also indicates that the representations extracted by our method are equally discriminative to those by the full-supervision model. The network pre-trained by BYOL and fine-tuned with patch labels achieves the best performance, especially in distinguishing the negative and positive samples. However, it is difficult to distinguish the subtypes (e.g., WDEA, MDEA and LDEA), as shown in Figure 2, for the reason that the model for representation extraction does not utilize subtype information of lesions. Lerousseau et al.[15] proposed to filter the pseudo-labels and learn the representations by a single ResNet, which is sensitive to noisy labels and suffers from the collapse of feature space. In contrast, our CDC module is based on the siamese network structure, which is more robust than the single branch structure. It brings an improvement of 11.2 % in the classification accuracy and 0.111 in AUC compared to Lerousseau et al.[15]

4. CONCLUSION

In this paper, we proposed a weakly supervised representation learning method based on WSI-level labels. BYOL was used to initialize the image representations. Then, a contrastive dynamic clustering (CDC) module was proposed to enable the weak-supervision task to be trained as the full-supervision paradigm. Experiments in an endometrial dataset consisting of 698 WSIs show that the proposed method is competitive to the full-supervision method, and achieves an 11.2% improvement in the classification accuracy than SOTA weak-supervised method.

ACKNOWLEDGMENTS

This work was partly supported by the National Natural Science Foundation of China [grant number 62171007, 61901018, 61771031, and61906058], partly supported by the China Postdoctoral Science Foundation [grant number 2019M650446], partly supported by Tianjin Science and Technology Major Project [grant number 18ZXZNSY00260], partly supported by the Anhui Provincial Natural Science Foundation [grant number 1908085MF210], partly supported by the Fundamental Research Funds for the Central Universities of China [grant number JZ2020YYPY0093], and partly supported by the 111 Project [grant number B13003].

REFERENCES

[1] Madabhushi, A. and Lee, G., "Image analysis and machine learning in digital pathology: Challenges and opportunities," *Medical Image Analysis* **33**, 170–175 (2016).

[2] Zheng, Y., Jiang, Z., Xie, F., Shi, J., Zhang, H., Huai, J., Cao, M., and Yang, X., "Diagnostic regions attention network (dra-net) for histopathology wsi recommendation and retrieval," *IEEE Transactions on Medical Imaging* **40**(3), 1090–1103 (2021).

[3] Lu, M. Y., Williamson, D. F. K., Chen, T. Y., Chen, R. J., Barbieri, M., and Mahmood, F., "Data-efficient and weakly supervised computational pathology on whole-slide images.," *Nature Biomedical Engineering* **5**(6), 555–570 (2021).

[4] Chan, L., Hosseini, M., Rowsell, C., Plataniotis, K., and Damaskinos, S., "Histosegnet: Semantic segmentation of histological tissue type in whole slide images," in [*2019 IEEE/CVF International Conference on Computer Vision (ICCV)*], 10662–10671 (2019).

[5] Lin, H., Chen, H., Wang, X., Wang, Q., Wang, L., and Heng, P.-A., "Dual-path network with synergistic grouping loss and evidence driven risk stratification for whole slide cervical image analysis.," *Medical Image Analysis* **69**, 101955–101955 (2021).

[6] Coudray, N., Ocampo, P. S., Sakellaropoulos, T., Narula, N., Snuderl, M., Fenyö, D., Moreira, A. L., Razavian, N., and Tsirigos, A., "Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning," *Nature medicine* **24**(10), 1559–1567 (2018).

[7] Arvaniti, E., Fricker, K. S., Moret, M., Rupp, N., Hermanns, T., Fankhauser, C., Wey, N., Wild, P. J., Rueschoff, J. H., and Claassen, M., "Automated gleason grading of prostate cancer tissue microarrays via deep learning," *Scientific reports* **8**(1), 1–11 (2018).

[8] Gertych, A., Swiderska-Chadaj, Z., Ma, Z., Ing, N., Markiewicz, T., Cierniak, S., Salemi, H., Guzman, S., Walts, A. E., and Knudsen, B. S., "Convolutional neural networks can accurately distinguish four histologic growth patterns of lung adenocarcinoma in digital slides," *Scientific reports* **9**(1), 1–12 (2019).

[9] He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R., "Momentum contrast for unsupervised visual representation learning," in [*2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*], 9729–9738 (2020).

[10] Chen, T., Kornblith, S., Norouzi, M., and Hinton, G., "A simple framework for contrastive learning of visual representations," in [*International conference on machine learning*], 1597–1607, PMLR (2020).

[11] Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P. H., Buchatskaya, E., Doersch, C., Pires, B. A., Guo, Z. D., Azar, M. G., Piot, B., Kavukcuoglu, K., Munos, R., and Valko, M., "Bootstrap your own latent: A new approach to self-supervised learning," in [*Advances in Neural Information Processing Systems*], **33**, 21271–21284 (2020).

[12] Wang, X., Yang, S., Zhang, J., Wang, M., Zhang, J., Huang, J., Yang, W., and Han, X., "Transpath: Transformer-based self-supervised learning for histopathological image classification," in [*International Conference on Medical Image Computing and Computer-Assisted Intervention*], 186–195, Springer (2021).

[13] Huang, Z., Chai, H., Wang, R., Wang, H., Yang, Y., and Wu, H., "Integration of patch features through self-supervised learning and transformer for survival analysis on whole slide images," in [*International Conference on Medical Image Computing and Computer-Assisted Intervention*], 561–570, Springer (2021).

[14] Yang, P., Hong, Z., Yin, X., Zhu, C., and Jiang, R., "Self-supervised visual representation learning for histopathological images," in [*International Conference on Medical Image Computing and Computer-Assisted Intervention*], 47–57, Springer (2021).

[15] Lerousseau, M., Vakalopoulou, M., Classe, M., Adam, J., Battistella, E., Carré, A., Estienne, T., Henry, T., Deutsch, E., and Paragios, N., "Weakly supervised multiple instance learning histopathological tumor segmentation," in [*International Conference on Medical Image Computing and Computer-Assisted Intervention*], 470–479 (2020).

[16] Wang, S., Zhu, Y., Yu, L., Chen, H., Lin, H., Wan, X., Fan, X., and Heng, P.-A., "Rmdl: Recalibrated multi-instance deep learning for whole slide gastric image classification.," *Medical Image Analysis* **58**, 101549 (2019).

[17] Zhu, W., Lou, Q., Vang, Y. S., and Xie, X., "Deep multi-instance networks with sparse label assignment for whole mammogram classification," in [*International Conference on Medical Image Computing and Computer-Assisted Intervention*], 603–611, Springer (2017).

[18] Li, W., Zhang, J., and McKenna, S. J., "Multiple instance cancer detection by boosting regularised trees," in [*International Conference on Medical Image Computing and Computer-Assisted Intervention*], 645–652, Springer (2015).

[19] Courtiol, P., Tramel, E. W., Sanselme, M., and Wainrib, G., "Classification and disease localization in histopathology using only global labels: A weakly-supervised approach," *arXiv preprint arXiv:1802.02212* (2018).

[20] Dietterich, T. G., Lathrop, R. H., and Lozano-Pérez, T., "Solving the multiple instance problem with axis-parallel rectangles," *Artificial intelligence* **89**(1-2), 31–71 (1997).

[21] Krizhevsky, A., Sutskever, I., and Hinton, G. E., "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems* **25**, 1097–1105 (2012).

[22] He, K., Zhang, X., Ren, S., and Sun, J., "Deep residual learning for image recognition," in [*2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*], 770–778 (2016).