Detection-Free Pipeline for Cervical Cancer Screening of Whole Slide Images

Maosong Cao[1], Manman Fei[2], Jiangdong Cai[1], Luyan Liu[1], Lichi Zhang[2], and Qian Wang[1(✉)]

[1] School of Biomedical Engineering, ShanghaiTech University, Shanghai, China
qianwang@shanghaitech.edu.cn

[2] School of Biomedical Engineering, Shanghai Jiao Tong University, Shanghai, China

Abstract. Cervical cancer is a significant health burden worldwide, and computer-aided diagnosis (CAD) pipelines have the potential to improve diagnosis efficiency and treatment outcomes. However, traditional CAD pipelines have limitations due to the requirement of a detection model trained on a large annotated dataset, which can be expensive and time-consuming. They also have a clear performance limit and low data utilization efficiency. To address these issues, we introduce a two-stage detection-free pipeline, incorporating pooling transformer and MoCo pre-training strategies, that optimizes data utilization for whole slide images (WSIs) while relying solely on sample-level diagnosis labels for training. The experimental results demonstrate the effectiveness of our approach, with performance scaling up as the amount of data increases. Overall, our novel pipeline has the potential to fully utilize massive data in WSI classification and can significantly improve cancer diagnosis and treatment. By reducing the reliance on expensive data labeling and detection models, our approach could enable more widespread and cost-effective implementation of CAD pipelines in clinical settings. Our code and model is available at https://github.com/thebestannie/Detection-free-MICCAI2023.

Keywords: Detection-free · Contrastive Learning · Pathology Image Classification · Cervical Cancer

1 Introduction

Cervical cancer is a common and severe disease that affects millions of women globally, particularly in developing countries [9]. Early diagnosis is vital for successful treatment, which can significantly increase the cure rate [17]. In recent years, computer-aided diagnosis (CAD) methods have become an important tool in the fight against cervical cancer, as they aim to improve the accuracy and efficiency of diagnosis.

Several computer-aided cervical cancer screening methods have been proposed for whole slide images (WSIs) in the literature. Most of them are detection-based methods, which typically contain a detection model as well as some post-processing modules in their frameworks. For instance, Zhou et al. [29] proposed a

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14225, pp. 243–252, 2023.
https://doi.org/10.1007/978-3-031-43987-2_24

three-step framework for cervical thin-prep cytologic test (TCT) [12]. The first step involves training a RetinaNet [13] as a cell detection network to localize suspiciously abnormal cervical cells from WSIs. In the second step, the patches centered on these detected cells are processed through a classification model, to refine the judgment of whether they are positive or negative. Finally, the positive patches refined by the patch-level classification are further combined to produce an overall positive/negative diagnosis for the WSI at the sample level.

Some methods improve the final classification performance by improving the detection model to identify positive cells more reliably. Cao et al. [1] improved the detection performance by incorporating clinical knowledge and attention mechanism into their cell detection model of AttFPN. Wei et al. [24] adopted the Yolo [20] architecture with a variety of convolution kernels of different sizes to accommodate diverse cell clusters. Other methods improve the classification performance by changing the post-processing modules behind the detection model. Cheng et al. [5] proposed a progressive identification method that leveraged multi-scale visual cues to identify abnormal cells and then an RNN [27] for sample-level classification. Zhang et al. [28] used GAT [23] to model the relation of the suspicious positive cells provided by detection, thus obtaining a global description of the WSI and performing sample-level classification.

These methods have achieved good results through continuous improvement on the detection-based pipeline, but there are some common drawbacks. First, they are not able to get rid of their reliance on detection models, which means they have a high need for expensive detection data labeling to train the detection model. Cervical cancer cell detection datasets involve labeling individual and small bounding boxes in a large number of cells. It often requires multiple experienced pathologists to annotate [15], which is very time-consuming and labor-intensive. Second, the widely used detection-based pipeline has not fully utilized the massive information in WSIs. A WSI is typically large (sized of about 20000×20000 pixels). A lot of data would be wasted if only a small part of annotated images (e.g., corresponding to positive cells and bounding boxes) was used as training data. Finally, many existing methods focus on detecting and classifying individual cells. The tendency to neglect effective integration of the overall information across the entire WSI results in poor performance in sample-level classification.

To address the aforementioned issues, we propose a detection-free pipeline in this paper, which does not rely on any detection model. Instead, our pipeline requires only sample-level diagnosis labels, which are naturally available in clinical scenarios and thus get rid of additional image labeling. To attain this goal, we have designed a two-stage pipeline as in Fig. 1. In the coarse-grained stage, we crop and downsample a WSI into multiple images, and conduct sample-level classification roughly based on all resized images. The coarse-grained classification yields attention scores, from which we perform attention guided selection to localize these key patches from the original WSI. Then, in the fine-grained stage, we use these key patches for fine prediction of the sample. The two stages in our pipeline adopt the same network design (i.e., encoder + pooling trans-

former), which makes our solution friendly to develop and to use. We also adopt contrastive learning to effectively utilize the massive information in WSIs when training the encoder for classification. As a summary, our pipeline surpasses previous detection-based methods and achieves state-of-the-art performance with large-scale training. Our experiments show that our method becomes more effective when increasing the data size for training. Moreover, while many pathological images are also based on WSIs, our pipeline has a high potential to extend to other pathological tasks.

2 Methodology

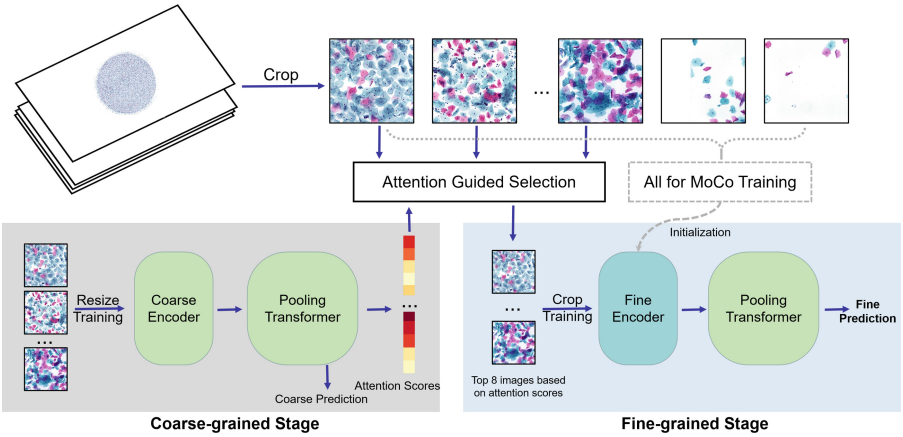

Fig. 1. The overview of our proposed method. For feasibility of computation, we crop a WSI into mutiple images. The cropped images are passed through the coarse-grained and fine-grained stages, where only sample-level diagnosis labels of WSIs, instead of any additional manual labeling, are required for training.

Two-Stage Pipeline with Attention Guided Selection. The overview of our two-stage pipeline is shown in Fig. 1. The input WSI is typically too big to be directly processed by a common deep learning pipeline [18], so we crop each WSI into local images sized 1024×1024 pixels. The images are then processed through the coarse-grained and fine-grained stages in order to obtain the WSI-level classification results, respectively. In general, the purpose of the coarse-grained stage is to replace the detection model and identify local images that may contain abnormal positive cells. The fine-grained stage then integrates these key regions, producing refined classification for the sample.

To complete sample-level classification, both stages share basically the same network architecture. The input images are first processed by a CNN encoder to

extract features. Then, we propose the pooling transformer, which is modified from the basic transformer module in Sect. 2, to integrate these features for WSI classification. Additionally, the input images for both stages are 256×256. In the coarse-grained stage, in order to allow the model to examine as many local images as possible, we resize the cropped local images from 1024×1024 to 256×256. In the fine-grained stage, we enlarge suspicious local abnormality and thus crop input images to 256×256 from 1024×1024.

For the coarse-grained stage, after passing the resized local images through encoder and pooling transformer, we obtain a rough prediction result at the sample level. We then use the Cross-Entropy (CE) loss to minimize the difference between the predicted WSI label and the ground truth. In addition, we calculate the attention score to identify the local image inputs that are most likely to yield positive reading. We describe the attention score as

$$AS(x_0, f) = Softmax(\frac{x_0 \cdot f^T}{\sqrt{d_{x_0}}})f, \tag{1}$$

where x_0 represents classification token (which is a commonly used setting in transformer [7,22]), and d_{x_0} is 512 in our implementation, f represents the feature vector of a certain input local image. After calculating attention scores, we preserve top-8 (resized) local images with the highest scores from the entire WSI for subsequent fine-grained classification.

Next, in the fine-grained stage, each local image that has passed attention guided selection is cropped into 16 patches of the size 256×256. We expect that those patches contain positive cells and are thus critical to diagnosis at the sample level. The network of the fine-grained stage is the same as that of the coarse-grained stage, but the weights of the encoder is pre-trained in an unsupervised manner (Sect. 2). The same CE loss supervised by sample-level ground truth is used for the fine-grained stage here. For inference, the output of the fine-grained stage will be treated as the final result of the test WSI.

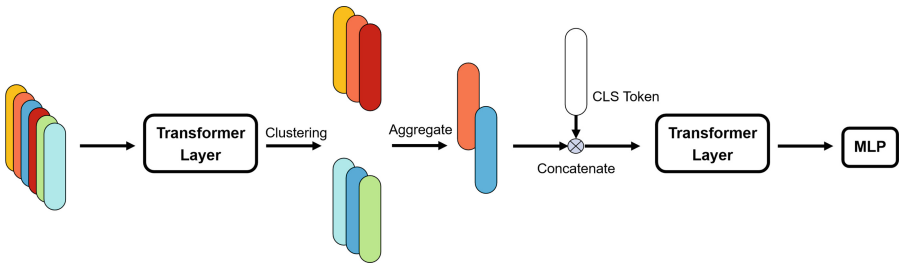

Fig. 2. Details of pooling transformer. It contains a token pooling layer in the middle of two transformer layers to cluster and aggregate redundant tokens.

Pooling Transformer. We use a transformer network to aggregate features of multiple inputs and to derive the sample-level outcome in both coarse-grained and fine-grained stages. We have observed that different local images of the same sample often have patterns of grouped similarity (such as the first two images in the upper-right of Fig. 1). For negative samples, most of the local images are similar with each other. For positive samples, the images of abnormal cells are inclined to be grouped into several clusters.

Therefore, inspired by [2,14], we propose pooling transformer that is effective to reduce the redundancy and distortion from the input images. The pooling transformer in Fig. 2 is designed to integrate all inputs toward the sample-level diagnosis. To remove redundant features, between two transformer layers, we use the affinity propagation algorithm [8] to cluster the inputs into several classes. Within each clustered class, we average the features and aggregate a single token. Finally, the classification (CLS) token is concatenated with all tokens after clustering-based pooling, and passed through the rest of the network to obtain the classification result. In this way, we find that the similar yet redundant input features can be fused, making the network more concise and efficient to calculate the attention between pooled features.

Contrastive Pre-training of Encoder. To make full use of WSI data and provide a better feature encoder, inspired by MoCo [4,11] and other contrastive learning methods [25,26] pre-training on ImageNet [6], we also perform pre-training for fine-grained encoder on a large scale of pathology images. Generally, large-scale pre-training usually requires a massive dataset and a suitable loss function. For data, WSI naturally has the advantage of having a large amount of training data. A WSI (20000×20000) can be cropped into about 5000–6000 patches (256×256). Therefore, we only need 2,000–3,000 WSI samples to obtain a dataset that can even be compared to ImageNet in quantity. For the loss function, there are typically two ways: one is like MAE [10] to model the loss function using masks, and the other is to use contrastive learning as in MoCo and CLIP [19]. In our task, since the structural features of cells are relatively weak compared to natural images, it is not suitable to model the loss function using masks. Therefore, we adopt a contrastive learning approach.

Specifically, in the same training batch, a patch (256×256, the same to the input size of the fine-grained stage) and its augmented patch are treated as a positive pair (note that here "positive/negative" is defined in the context of contrastive learning), and their features are required to be as similar as possible. Meanwhile, their features are required to be as dissimilar as possible from those of other patches. So the loss function can be described as

$$L = -\sum_{i=0}^{n}\left(\frac{f_i \cdot f_{i_a}}{|f_i||f_{i_a}|} - \sum_{j=1}^{n}\frac{f_i \cdot f_j}{|f_i||f_j|}\right) \tag{2}$$

f_i and f_{i_a} represent the positive pair, and f_j represents another patch negatively paired with f_i. Using this method, we can pre-train a feature encoder in an unsupervised manner and initialize it into our encoder for the fine-grained stage.

3 Experiment and Results

Dataset and Experimental Setup. In this study, we have collected 5384 cervical cytopathological WSI by 20x lens, each with 20000×20000 pixels, from our collaborating hospitals. Among them, there 2853 negative samples, and 2531 positive samples (962 ASCUS, and 1569 high-level positive samples). All WSIs only have diagnosis labels at the sample level, without annotation boxes at the cell level. And all sample labels are strictly diagnosed according to the TBS [16] criterion by a pathologist with 35 years of clinical experience. We conduct the experiment with initial learning rate of 1.0×10^{-4}, batch size of 4, and SGD optimizer [21] for 30 epochs each stage. For contrastive pre-training, we follow the settings of MoCov2 [3] and trained for 300 epochs.

Comparison to SOTA Methods. In this section, we experiment to compare our method with popular state-of-the-art (SOTA) methods, which are all fully supervised and detection-based. To the best of our knowledge, there are few good methods to train cervical cancer classification models in weakly supervised or unsupervised learning ways. No methods can achieve the detection-free goal either.

All the detection-based methods are evaluated in the following way. First, we label a dataset with cell-level bounding boxes to train a detection model. The detection dataset has 3761 images and 7623 cell-level annotations. After obtaining the suspicious cell patches provided by the detection model, we use the subsequent classification models used in these SOTA works to classify them and obtain the final classification results.

Table 1. Comparison with SOTA methods (%).

Method	Accuracy↑	Precision↑	Recall↑	F1-Score↑
AttFPN+Average [1]	78.33 ± 1.51	71.23 ± 2.10	79.39 ± 1.56	74.10 ± 2.33
RetinaNet+MLP [29]	74.31 ± 2.31	65.34 ± 3.57	78.12 ± 1.59	73.44 ± 2.78
RetinaNet+SVM [29]	72.37 ± 1.40	73.96 ± 0.79	77.38 ± 0.88	75.86 ± 0.95
LRModel+RNN [5]	80.55 ± 1.53	74.59 ± 1.66	82.31 ± 1.74	78.51 ± 1.32
RetinaNet+GAT [28]	83.74 ± 1.35	$\mathbf{80.38 \pm 1.43}$	84.58 ± 1.48	81.93 ± 1.02
Ours (Detection-Free)	$\mathbf{83.84 \pm 1.56}$	78.36 ± 1.23	$\mathbf{85.22 \pm 0.98}$	$\mathbf{82.12 \pm 0.93}$

As shown in Table 1 for fair five-fold cross-validation, our method outperforms all compared detection-based methods. While our method has a large margin with most methods in the table, the improvement against [28] (top-ranked in current detection-based methods) is relatively limited. On one hand, in [28], GAT aggregates local patches that are detected by Retinanet. And the attention mechanism of GAT is similar with the transformer used in our pipeline to certain extent. On the other hand, the result implies that our coarse-grained task has replaced the role of cell detection in early works. Thus, we conclude that a

detection model trained with an expensive annotated dataset is not necessary to build a CAD pipeline for cervical abnormality.

Ablation Study. In this section, we experiment to demonstrate the effectiveness of all the proposed parts in our pipeline. We divide all 5384 samples into five independent parts for five-fold cross-validation, and the results are shown in Table 2. Here, CG means the classification passes only the coarse-grained stage. As can be seen, its performance is low, in that the resized images sacrifices the resolution and thus perform poorly for image-based classification. FG refers to classifying in the fine-grained stage. It is worth noting that without the attention scores provided by the coarse-grained stage, we have no way of knowing which local images might contain suspicious positive cells. Thus, we use random selection to experiment for FG only, as exhaustively checking all local images is computationally forbidden. As can be seen, the classification result is the lowest because it lacks enough access to the key image content in WSIs.

Table 2. Ablation study of our proposed methods. CG indicates coarse-grained classification, FG indicates fine-grained classification, PT indicates pooling transformer, and CL indicates contrastive learning.

Configuration				Metric (%)			
CG	FG	PT	CL	ACC	Precision	Recall	F1 Score
✔	–	–	–	74.96 ± 1.23	71.39 ± 1.21	82.49 ± 1.39	75.34 ± 1.66
–	✔	–	–	73.11 ± 2.10	70.48 ± 1.45	81.59 ± 1.45	74.09 ± 1.49
✔	✔	–	–	79.72 ± 1.30	74.64 ± 1.34	84.45 ± 1.95	78.48 ± 1.23
✔	✔	✔	–	81.34 ± 1.56	77.36 ± 1.23	84.79 ± 0.98	80.72 ± 0.93
✔	✔	✔	✔	83.84 ± 1.56	78.36 ± 1.23	85.22 ± 0.98	82.12 ± 0.93

By combining the two stages for attention guided selection, it is effective to improve the classification performance compared to the two previous experiments. Here, for the cases of CG, FG and CG+FG, an original transformer network without clustering-based pooling is used. In addition, as shown in the last two rows of the table, both pooling transformer (PT) and unsupervised pretraining (CL) contribute to our pipeline. Ultimately, we combine them together to achieve the best performance.

Sample Numbers and Inference Time. In order to further demonstrate the huge potential of our method, we also perform an ablation study on the number of samples used for training and compare the time consuming of the different methods. For the experiment of sample numbers, We compare the best fully supervised detection-based method (Retinanet+GAT [28]) with ours under the sample numbers of 500, 1000, 2000, and 5384. As shown by Table 3 and

Table 3. Alation study on sample number between Retinanet+GAT [28] and ours (%).

Number	Method	Accuracy↑	Precision↑	Recall↑	F1-Score↑
500	RetinaNet+GAT	77.59 ± 1.33	71.32 ± 1.39	75.23 ± 1.03	73.85 ± 1.45
	Ours	70.34 ± 2.58	64.49 ± 1.98	71.49 ± 2.49	68.88 ± 2.36
1000	RetinaNet+GAT	79.85 ± 1.70	75.60 ± 1.26	80.06 ± 0.96	77.96 ± 1.50
	Ours	74.47 ± 1.29	70.23 ± 1.21	76.45 ± 0.73	73.96 ± 0.78
2000	RetinaNet+GAT	81.59 ± 1.09	78.12 ± 1.73	82.37 ± 1.55	80.79 ± 1.38
	Ours	78.10 ± 1.32	73.88 ± 1.38	79.93 ± 1.06	77.39 ± 1.08
5384	RetinaNet+GAT	83.74 ± 1.35	80.38 ± 1.43	84.58 ± 1.48	81.93 ± 1.02
	Ours	83.84 ± 1.56	78.36 ± 1.23	85.22 ± 0.98	82.12 ± 0.93

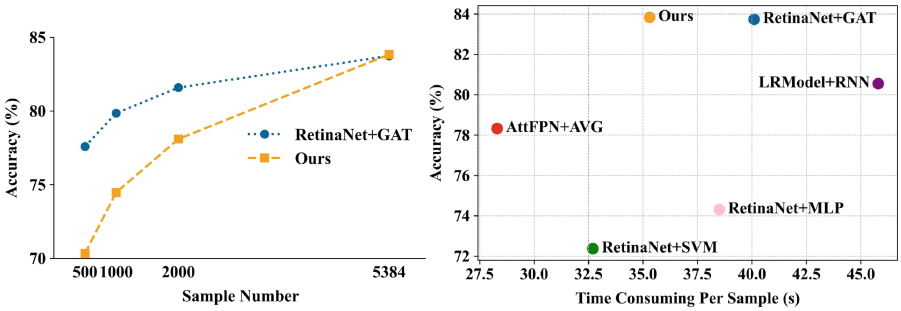

Fig. 3. Visualization of accuracy and inference time consuming for different methods.

left of Fig. 3, the traditional detection-based method has quickly encountered a saturation bottleneck as the amount of data increases. Although our method initially has poorer performance, it has shown an impressive growth trend. And at our current maximum data number (5384), the proposed pipeline has already exceeded the performance of the detection-based method. The above results also demonstrate that our new pipeline method has greater potential, even though it requires no cell-level image annotation. For inference time consuming, as shown in right of Fig. 3, our method has shorter inference time and a good balance between accuracy and inference time.

4 Conclusion and Discussion

In this paper, we propose a novel two-stage detection-free pipeline for WSI classification of cervical abnormality. Our method does not rely on detection models and eliminates the need for expensive cell-level data annotation. By leveraging just sample-level diagnosis labels, we achieve results that are competitive with fully supervised detection-based methods. Through the use of the proposed pooling transformer and unsupervised pre-training, our method makes full use

of information within WSIs, resulting in improved efficiency in the use of pathological images. Importantly, our method offers even greater advantages with increasing amounts of data. And also, by utilizing attention weights, we can calculate attention scores to visually represent the importance of each image in the sample, making it easier for doctors to make judgments. Relevant visualization results can be found on our project homepage. Admittedly, our method has some limitations, such as slow training. Accelerating the training of massive data can be our next optimization direction.

References

1. Cao, L., et al.: A novel attention-guided convolutional network for the detection of abnormal cervical cells in cervical cancer screening. Med. Image Anal. **73**, 102197 (2021)
2. Chen, B., et al.: PSViT: better vision transformer via token pooling and attention sharing. arXiv preprint arXiv:2108.03428 (2021)
3. Chen, X., Fan, H., Girshick, R., He, K.: Improved baselines with momentum contrastive learning. arXiv preprint arXiv:2003.04297 (2020)
4. Chen, X., Xie, S., He, K.: An empirical study of training self-supervised vision transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 9640–9649 (2021)
5. Cheng, S., et al.: Robust whole slide image analysis for cervical cancer screening using deep learning. Nat. Commun. **12**(1), 1–10 (2021)
6. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: a large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255. IEEE (2009)
7. Dosovitskiy, A., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
8. Frey, B.J., Dueck, D.: Clustering by passing messages between data points. Science **315**(5814), 972–976 (2007)
9. Gultekin, M., Ramirez, P.T., Broutet, N., Hutubessy, R.: World health organization call for action to eliminate cervical cancer globally. Int. J. Gynecol. Cancer **30**(4), 426–427 (2020)
10. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16000–16009 (2022)
11. 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)
12. Koss, L.G.: The papanicolaou test for cervical cancer detection: a triumph and a tragedy. Jama **261**(5), 737–743 (1989)
13. Lin, T.-Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2980–2988 (2017)
14. Marin, D., Chang, J.H.R., Ranjan, A., Prabhu, A., Rastegari, M., Tuzel, O.: Token pooling in vision transformers. arXiv preprint arXiv:2110.03860 (2021)
15. Meng, Z., Zhao, Z., Li, B., Fei, S., Guo, L.: A cervical histopathology dataset for computer aided diagnosis of precancerous lesions. IEEE Trans. Med. Imaging **40**(6), 1531–1541 (2021)

16. Nayar, R., Wilbur, D.C.: The Bethesda System for Reporting Cervical Cytology: Definitions, Criteria, and Explanatory Notes. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-11074-5
17. Patel, M.M., Pandya, A.N., Modi, J.: Cervical pap smear study and its utility in cancer screening, to specify the strategy for cervical cancer control. National J. Commun. Med. **2**(01), 49–51 (2011)
18. Qu, L., Luo, X., Liu, S., Wang, M., Song, Z.: DGMIL: distribution guided multiple instance learning for whole slide image classification. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) MICCAI 2022. LNCS, vol. 13432, pp. 24–34. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16434-7_3
19. Radford, A., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning, pp. 8748–8763. PMLR (2021)
20. Redmon, J., Farhadi, A.: Yolov3: an incremental improvement. arXiv preprint arXiv:1804.02767 (2018)
21. Robbins, H., Monro, S.: A stochastic approximation method. Ann. Math. Stat. 400–407 (1951)
22. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, vol. 30 (2017)
23. Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., Bengio, Y.: Graph attention networks. arXiv preprint arXiv:1710.10903 (2017)
24. Wei, Z., Cheng, S., Liu, X., Zeng, S.: An efficient cervical whole slide image analysis framework based on multi-scale semantic and spatial deep features. arXiv preprint arXiv:2106.15113 (2021)
25. Wu, Z., Xiong, Y., Yu, S.X., Lin, D.: Unsupervised feature learning via non-parametric instance discrimination. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3733–3742 (2018)
26. Ye, M., Zhang, X., Yuen, P.C., Chang, S.-F.: Unsupervised embedding learning via invariant and spreading instance feature. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6210–6219 (2019)
27. Zaremba, W., Sutskever, I., Vinyals, O.: Recurrent neural network regularization. arXiv preprint arXiv:1409.2329 (2014)
28. Zhang, X., et al.: Whole slide cervical cancer screening using graph attention network and supervised contrastive learning. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) MICCAI 2022. LNCS, vol. 13432, pp. 202–211. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16434-7_20
29. Zhou, M., Zhang, L., Xiaping, D., Ouyang, X., Zhang, X., Shen, Q., Luo, D., Fan, X., Wang, Q.: Hierarchical pathology screening for cervical abnormality. Comput. Med. Imaging Graph. **89**, 101892 (2021)