Refining Cervical Cell Classification with Cytological Knowledge and Optimal Attribute Descriptor Matching

Manman Fei[1], Zhenrong Shen[1], Mengjun Liu[2], Zhiyun Song[1], Yusong Sun[1], Xu Han[1], Zelin Liu[1], Haotian Jiang[1], Lu Bai[3], Qian Wang[2(✉)], and Lichi Zhang[1(✉)]

[1] School of Biomedical Engineering, Shanghai Jiao Tong University, Shanghai, China
lichizhang@sjtu.edu.cn
[2] School of Biomedical Engineering, ShanghaiTech University, Shanghai, China
qianwang@shanghaitech.edu.cn
[3] School of Artificial Intelligence, Beijing Normal University, Beijing, China

Abstract. Cervical cancer remains a significant global health concern, emphasizing the need for effective diagnostic methods. Despite advancements in Vision Language Models, challenges persist in incorporating cytological knowledge, ensuring data relevance, and maintaining accuracy when aggregating visual information. Current methods often struggle to handle fine-grained morphological details and the complex relationships between images and textual knowledge. In this paper, we present a novel framework for cervical cell classification that combines attribute descriptors with cytological knowledge for enhanced morphology recognition. Our approach leverages the Vision Large Language Model to generate descriptions for each cervical image and pretrain image and text encoders, improving both image understanding and cytological context. We introduce Attribute Descriptors Extraction using LLMs and Retrieval-Augmented Generation to generate detailed descriptors that capture important cytological features while minimizing irrelevant information. Additionally, we propose Optimal Attribute Descriptors Matching to dynamically align textual descriptors with image features, enhancing prediction accuracy, interpretability, and cytological relevance. Experimental results demonstrate the superior performance and generalizability of our method with varying amounts of labeled data. The code is publicly available at https://github.com/feimanman/CervicalCellClassifier

Keywords: Cervical Cell Classification · Vision Language Models · Optimal Transport · Prompt Learning

1 Introduction

Cervical cancer is a significant health risk for women worldwide. Clinical pathology screening using liquid-based cytology is recommended for the early diagnosis of cervical precancerous lesions, which effectively prevents the progression to invasive cancer [21]. In pathology image analysis, the interpretation of

cervical cells is crucial for accurate diagnosis. However, traditional microscopic analysis is inefficient and labor-intensive, struggling to meet the growing demand for regular screening. Computer-aided diagnosis using deep neural networks [9,16,11,5,6,22,10,12] has shown the potential to significantly enhance the diagnostic efficiency of pathologists. According to the Bethesda system (TBS), cervical squamous cells are classified into five subtypes: HSIL, ASC-H, LSIL, ASC-US, and NILM. Ghoneim et al. [13] utilized convolutional neural networks (CNN) to extract cell features and applied extreme learning machine-based classifiers for cell classification. Yu et al. [25] introduced spatial pyramid pooling and inception modules into CNNs to handle the classification of images of varying sizes. While effective, these image-based techniques require large datasets and often lack interpretability.

Recently, Vision-Language Models have shown remarkable performance in various computer vision tasks [20]. Models like CLIP [20] align image features with textual descriptions through contrastive learning, they generally focus on broad class labels rather than the fine morphological details required for medical image analysis. To address this limitation, soft prompt tuning has become prominent in enabling vision-language models to efficiently adapt to downstream tasks [29,28,3]. For example, CoOp [29] introduces soft prompts, achieving improvements at the cost of robustness, while CoCoOp [28] enhances adaptability by conditioning prompts on individual images, albeit with higher computational demands. Despite these advancements, these methods typically enhance high-level semantics in a coarse-grained manner, leading to holistic alignment across modalities. This can hinder the model's ability to distinguish among classes that share similar visual attributes. To overcome this, recent works like ArGue [24] and LLaMP [27] use large language models (LLMs) to enrich text prompts with fine-grained class descriptions, thereby improving classification performance. However, these methods require additional procedures to ensure the relevance of LLM outputs.

While these attempts significantly enhance classification performance, several issues remain to be addressed: Firstly, methods attempting to use textual prompts to highlight critical areas in pathology images [17,30,26] often underperform when handling cytological knowledge for cervical cell classification. Secondly, relying solely on LLMs can result in the generation of irrelevant attributes, as these models may produce hallucinated outputs, introducing noise into the descriptors. Finally, approaches aggregating descriptors into global categories to match visual features may be ineffective if not all images within a class possess relevant attribute descriptors.

To address these challenges, we propose a novel framework for cervical cell classification that integrates structured attribute descriptors with clinically relevant textual knowledge for fine-grained morphology recognition. Our main contributions include: 1) We utilize the Vision Large Language Model (VLLM) to generate descriptions for each cervical image and pretrain the image and text encoders, effectively learning representation that enhances both image understanding and cytological context; 2) We introduce Attribute Descriptors Ex-

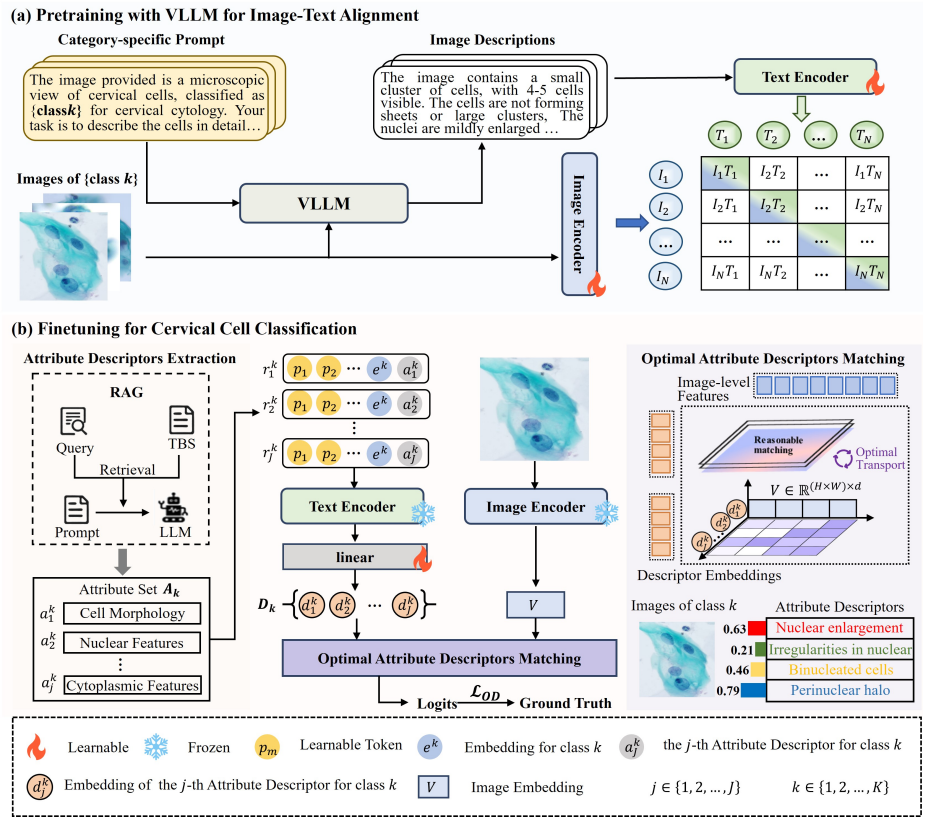

Fig. 1. (a) Pretraining with VLLM for Image-Text Alignment. (b) Fine-tuning for Cervical Cell Classification, involving attribute descriptor extraction from medical documents, followed by optimal attribute descriptor matching to refine the alignment between visual and textual information.

traction by employing LLMs and Retrieval-Augmented Generation (RAG) technology [18] to generate fine-grained descriptors that capture key cytological attributes, minimizing irrelevant information.; 3) We propose Optimal Attribute Descriptors Matching to dynamically align textual descriptors with image features, enhancing prediction accuracy, interpretability, and cytological relevance; 4) Experimental results across datasets with varying labeled data demonstrate the superior performance and generalizability of our approach.

2 Methods

As illustrated in Fig. 1, the proposed framework consists of two training stages. In the first stage, the model establishes a robust representation by associating visual features with detailed textual descriptions, enhancing both image under-

standing and cytological context. In the second stage, the model is finetuned to specialize in distinguishing the subtle morphological characteristics of cervical cells, thereby enhancing classification accuracy and clinical relevance.

2.1 Pretraining with VLLM for Image-Text Alignment

We leverage a pretrained VLLM to analyze cervical cell images and generate corresponding descriptions as shown in Fig. 1 (a). Our method use category-specific prompts to ensure the generated descriptions are highly relevant to the morphological characteristics of the cervical cells. Specifically, for each image I belonging to a specific class k, we design a category-specific prompt that elicits the most relevant description based on the visual characteristics of that category: "The image provided is a microscopic view of cervical cells, classified as class $\{k\}$ for cervical cytology. Your task is to describe the cells in detail". This prompt is input into the VLLM to generate descriptions T_n that corresponds to a specific visual feature of the cell, such as shape, size, texture, or structural characteristics. We integrate the generated descriptions into the CLIP framework for image-text representation learning. The loss function is formulated as follows:

$$\mathcal{L}_{\text{pretrain}} = -\sum_{(I,T_n)} \log \frac{\exp(cos(I,T_n)/\tau)}{\sum_{n'} \exp(cos(I,T_{n'})/\tau)}, \tag{1}$$

where $cos(I,T_n)$ is the cosine similarity between the image and text pair, τ is a temperature scaling factor, and n' represents negative samples, i.e., mismatched image-text pairs. It enhances the model's comprehension of cervical cell characteristics, and facilitates better generalization across various categories.

2.2 Fine-tuning for Cervical Cell Classification

Attribute Descriptors Extraction We classify cervical cancer cells using the CLIP framework. The process involves calculating matching scores via cosine similarity between image and category embeddings. While this method works well in many cases, it often relies solely on the understanding of category names, which may result in a less distinctive semantic space. To address this, we introduce Attribute Descriptors Extraction, where each cervical cell type is represented by fine-grained attribute descriptors. To ensure that our cervical cell classification model is trained on clinically validated morphological attributes, we leverage Retrieval-Augmented Generation (RAG) to extract structured knowledge from the Bethesda system (TBS) [19], which dynamically constructs the retrieval query to extract the most relevant cytological descriptions.

Formally, we define a query for class k as "Retrieve the key morphological features of class $\{k\}$ according to TBS and pathology guidelines." to retrieve relevant information. This query retrieves relevant passages from the database. Then, we prompt the LLM with the retrieved information to generate a structured attribute set, where the prompt is designed as "Based on the retrieved medical literature, list the key morphological attributes of class $\{k\}$". The LLM

then generates an attribute set $A_k = \{a_1^k, a_2^k, ..., a_j^k\}$. Each a_j corresponds to an attribute descriptor. Through RAG, the generated attribute set is aligned with cytological knowledge.

Optimal Attribute Descriptors Matching In image classification tasks that require recognizing subtle morphological distinctions, it is essential to match image features with relevant attribute descriptors for precise analysis. Optimal Attribute Descriptors Matching addresses this need by ensuring that each image is paired with the most pertinent attribute descriptors. We represent each attribute descriptor for class k using $r_j^k = \{p_1, p_2, \ldots, p_M, e^k, a_j^k\}$, where $\{p_m\}_{m=1}^M$ are learnable tokens, e^k is the word embedding for the class, and a_j^k is the embedding for the j-th attribute descriptor. This composite representation r_j^k is processed through a pretrained text encoder and a linear layer to generate the soft embedding d_j^k. The full set of descriptor embeddings for class k is denoted as $D_k = \left\{d_1^k, d_2^k, \ldots, d_J^k\right\}$, where J is the number of descriptors per class, and d is the embedding dimension.

Previous methods typically form a global descriptor embedding to compute similarity: $S_k^{\text{pool}} = \cos(V, D_k)$, where $V \in \mathbb{R}^{(H\times W)\times d}$ represents the image-level features, $\cos(\cdot,\cdot)$ is the cosine similarity. However, this approach encounters difficulties in efficiently matching descriptors to image features. To address this issue, we propose using Optimal Descriptor Solver (OD Solver), initially proposed by Chen et al. [7], which adapts optimal transport theory to solve this image-text matching problem as an optimal matching flow. After obtaining the image-level features V and the descriptor-level embedding D_k for each class, we define the cost matrix C_k for each class k. The cost matrix measures the dissimilarity between image features and attribute descriptors: $C_k = 1 - \cos(V, D_k)$.

We introduce the entropy-regularized Optimal Transport (OT) problem, which aims to find the optimal matching between the image features and the attribute descriptors. Assuming we have two sets of discrete empirical distributions:

$$\mu = \sum_{i=1}^{H\times W} \alpha_i \delta_{x_i}, \quad \nu = \sum_{j=1}^{J} \beta_j \delta_{y_j}, \tag{2}$$

where α_i and β_j are the probability distribution summing to 1, δ denotes the Dirac function. In this setting, we can adapt Kantorovich OT formulation [15] and form the optimal transport problem as:

$$P^* = \arg\min_{P\in\mathbb{R}^{(H\times W)\times J}} \sum_{i=1}^{(H\times W)} \sum_{j=1}^{J} P_{ij} C_{ij} - \lambda H(P) \tag{3}$$

$$\text{s.t.} \quad Pe = \mu, \quad P^\top e = \nu.$$

We can obtain the optimal matching flow between the image and descriptors $P^* \in \mathbb{R}^{(H\times W)\times J}$. Here, $H(P)$ is the entropy regularization term, λ is the regularization hyperparameter. The constraints $Pe = \mu$ and $P^\top e = \nu$ ensure that the

total mass flowing out of the image features and the total mass flowing into the attribute descriptors are both preserved and equal to their respective marginal distributions, μ and ν.

This optimization problem can be solved efficiently using the Sinkhorn algorithm [8]. It provides the optimal transport plan P^* that minimizes the total matching cost. Once the optimal transport matrix P^* is computed, we can obtain the matching score between the image and descriptors by calculating the Frobenius inner product between the transport matrix and the similarity matrix:

$$S_k^{\mathrm{OT}} = \sum_{i=1}^{(H\times W)} \sum_{j=1}^{J} P_{ij}^* \cos(V_i, D_k). \tag{4}$$

The matching score quantifies the alignment between the image features and the attribute descriptors based on the optimal transport flow. Finally, we fuse the overall matching score obtained in both the Euclidean space and Wasserstein space to compute the final logits for each class:

$$S_k^{\mathrm{OD}} = \frac{1}{2}\left(S_k^{\mathrm{pool}} + S_k^{\mathrm{OT}}\right). \tag{5}$$

This fusion combines the matching scores in both spaces to produce the final logits, which are used for the classification decision.

2.3 Optimization

In our cervical cell classification framework, we first perform a pretraining phase by a VLLM model. The pretraining loss, as defined in Eq. 1, is a contrastive loss that encourages the image and text encoders to produce aligned representations. During fine-tuning, we use the OD Solver for classification logits. Considering the logits in Eq. 5, the softmax-normalized similarity scores can be expressed as:

$$p_i^{OD} = \frac{1}{K}\sum_{k=1}^{K} \frac{\exp(S_k^i/\tau)}{\sum_{b=1}^{B}\exp(S_k^b/\tau)}, \tag{6}$$

where τ is the temperature hyperparameter, B is the mini-batch size. To optimize the model, we use Cross Entropy(CE) Loss to minimize the difference between predicted and ground-truth q. The loss for the fine-tuning phase is defined as: $\mathcal{L}_{OD} = \mathrm{CE}(p^{\mathrm{OD}}, q)$. The total training loss is the sum of the pretraining and fine-tuning losses.

$$\mathcal{L}_{\mathrm{total}} = \mathcal{L}_{\mathrm{pretrain}} + \mathcal{L}_{OD}. \tag{7}$$

In the first stage, we perform pretraining to adapt CLIP [20] from the natural image domain to the cervical cell pathology domain, achieving coarse-grained alignment of image and text features. In the second stage, we introduce the Attribute Set and leverage the CoOP [29] method to fine-tune CLIP for fine-grained classification, ensuring precise alignment between the extracted features and the Attribute Set, including learnable tokens.

Table 1. Performance comparison with state-of-the-art methods.

Methods	Data used		Performance Metrics(%)				
	Labeled	Text	ACC	Recall	Precision	F1-score	AUC
ResNet50 [14]	100%	×	70.85	70.59	69.93	70.07	91.73
ViT [9]	100%	×	**72.10**	**71.37**	**71.11**	**71.02**	**93.05**
CLIP [20]	100%	✓	68.39	67.83	67.34	67.45	90.80
PathCLIP [23]	25%	✓	62.56	61.71	62.94	61.61	88.59
CoOP [29]	25%	✓	62.91	62.77	62.38	62.52	87.98
CoCoOp [28]	25%	✓	63.04	62.91	62.78	62.48	89.07
LASP [3]	25%	✓	65.80	65.94	64.88	65.16	90.76
Ours	25%	✓	**69.30**	**68.16**	**68.67**	**68.35**	**91.71**
PathCLIP [23]	50%	✓	64.38	63.60	64.67	63.78	90.15
CoOP [29]	50%	✓	64.16	63.85	63.30	63.15	90.10
CoCoOp [28]	50%	✓	65.20	65.53	64.79	64.88	90.36
LASP [3]	50%	✓	67.92	67.30	66.87	67.02	91.01
Ours	50%	✓	**71.42**	**71.90**	**70.74**	**71.13**	**92.66**

Table 2. Quantitative results for the ablation study.

Pretraining with VLLM	Attribute Descriptors	Optimal Matching	Performance Metrics(%)				
			ACC	Recall	Precision	F1-score	AUC
×	×	×	65.93	64.82	66.34	65.30	90.27
✓	×	×	66.80	66.83	66.09	66.07	90.95
✓	✓	×	67.27	66.86	66.97	66.90	91.08
✓	✓	✓	**71.42**	**71.90**	**70.74**	**71.13**	**92.66**

3 Experimental Results

3.1 Dataset and Experimental Setup

Dataset We used the publicly available HiCervix dataset [4], a cervical cytology dataset with 29 annotated categories. We selected five categories for classification: ASC-US, ASC-H, LSIL, HSIL, and NILM. The division of the training set, validation set, and testing set is based on the methodology described in [4].

Implementation Details In the first stage, Pretraining with VLLM for Image-Text Alignment, we used the Qwen2.5-VL-7B [1] model for analyzing cervical images and generating descriptions. The ResNet50 backbone weights from CLIP were utilized, with an initial learning rate of 1e-4, Adam optimizer, weight decay of 1e-4, and a batch size of 64. For the fine-tuning phase, we selected GPT-3 [2] as our LLM and kept the Adam optimizer and weight decay the same but reduced the learning rate to 1e-6. The evaluation metrics for this study included accuracy, recall, precision, F1-score, and macro AUC.

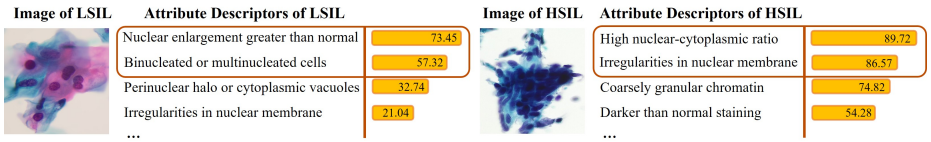

Fig. 2. We demonstrate several attributes inside each class and the number within the yellow bar indicates its optimal matching similarity to images.

3.2 Evaluation of Cervical Cell Classification

Comparison with SOTA Methods To validate the effectiveness of our method in scenarios with limited labeled data, we set up the experiments as shown in the Table 1. We compared our method with prompt learning approaches: CoOp [29], CoCoOp [28], and LASP [3]. Additionally, we also compared our method with PathCLIP [23], a CLIP model specifically designed for pathology. As shown in Table 1, in the scenario with 50% labeled data and 25% labeled data, our method outperforms several state-of-the-art models across multiple metrics. With 50% labeled data, our model achieved F1-score (71.13%) and AUC (92.66%), outperforming models like PathCLIP, which is specifically designed for pathology tasks. The strength of our approach lies in its ability to effectively integrate both visual and textual information to boost classification performance. These results highlight that our method excels with limited labeled data, offering better generalization and outstanding performance in cervical cell classification. Furthermore, the results of our method in the 50% labeled scenario are similar to those obtained using fully labeled data, demonstrating the effectiveness of combining image-text alignment and prompt learning for superior classification results. Additionally, to showcase the interpretability of our proposed method, as shown in Fig. 2, we are able to assign images to their most similar attributes.

Ablation Study In the ablation study, we conducted a comprehensive evaluation to quantify the contribution of each individual component, using 50% labeled data, as shown in Table 2. The baseline configuration, which is based on the CLIP model. Incorporating VLLM pretraining enhanced the model's feature extraction capabilities, resulting in more semantically enriched representations and improving both accuracy and AUC. The addition of attribute alignment enabled the model to focus on clinically relevant morphological features, thereby refining its understanding and further improving performance. The attributes descriptors for each category was aggregated into a unified descriptor. For each category, we calculate the cosine similarity between the image feature embeddings and the aggregated attribute descriptor to achieve matching. The accuracy reached 67.27% and the AUC increased to 91.08%. Finally, the combination of all components, including optimal descriptor matching, led to the highest performance.

4 Conclusion

In this work, we propose a novel framework for cervical cell classification that integrates structured attribute descriptors with clinically relevant textual knowledge for fine-grained morphology recognition. Utilizing a VLLM, we generate detailed descriptions for cervical images and pretrain image and text encoders to enhance the model's understanding of both visual and cytological contexts. Additionally, we introduce an Attribute Descriptors Extraction using LLM and RAG technology to produce fine-grained, clinically validated descriptors. Our method further incorporates Optimal Attribute Descriptors Matching, dynamically aligning textual descriptors with image features to ensure precise morphological attribute matching, thereby enhancing interpretability and clinical relevance. Experimental results show that our framework significantly improves classification performance, especially in scenarios with limited labeled data.

Acknowledgments. This work was supported by the National Natural Science Foundation of China (Grant No. 62471288).

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

References

1. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al.: Qwen2. 5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)
2. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. Advances in neural information processing systems **33**, 1877–1901 (2020)
3. Bulat, A., Tzimiropoulos, G.: Lasp: Text-to-text optimization for language-aware soft prompting of vision & language models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 23232–23241 (2023)
4. Cai, D., Chen, J., Zhao, J., Xue, Y., Yang, S., Yuan, W., Feng, M., Weng, H., Liu, S., Peng, Y., et al.: Hicervix: An extensive hierarchical dataset and benchmark for cervical cytology classification. IEEE transactions on medical imaging (2024)
5. Cao, M., Fei, M., Cai, J., Liu, L., Zhang, L., Wang, Q.: Detection-free pipeline for cervical cancer screening of whole slide images. In: International conference on medical image computing and computer-assisted intervention. pp. 243–252. Springer (2023)
6. Cao, M., Fei, M., Xiong, H., Zhang, X., Fan, X., Zhang, L., Wang, Q.: Patch-to-sample reasoning for cervical cancer screening of whole slide image. IEEE Transactions on Artificial Intelligence **5**(6), 2779–2789 (2023)
7. Chen, T., Yu, H., Yang, Z., Li, Z., Sun, W., Chen, C.: Ost: Refining text knowledge with optimal spatio-temporal descriptor for general video recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18888–18898 (2024)
8. Cuturi, M.: Sinkhorn distances: Lightspeed computation of optimal transport. Advances in neural information processing systems **26** (2013)

9. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
10. Fei, M., Shen, Z., Song, Z., Wang, X., Cao, M., Yao, L., Zhao, X., Wang, Q., Zhang, L.: Distillation of multi-class cervical lesion cell detection via synthesis-aided pretraining and patch-level feature alignment. Neural Networks **178**, 106405 (2024)
11. Fei, M., Zhang, X., Cao, M., Shen, Z., Zhao, X., Song, Z., Wang, Q., Zhang, L.: Robust cervical abnormal cell detection via distillation from local-scale consistency refinement. In: International conference on medical image computing and computer-assisted intervention. pp. 652–661. Springer (2023)
12. Fei, M., Zhang, X., Chen, D., Song, Z., Wang, Q., Zhang, L.: Whole slide cervical cancer classification via graph attention networks and contrastive learning. Neurocomputing **613**, 128787 (2025)
13. Ghoneim, A., Muhammad, G., Hossain, M.S.: Cervical cancer classification using convolutional neural networks and extreme learning machines. Future Generation Computer Systems **102**, 643–649 (2020)
14. 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)
15. Kantorovich, L.V.: On the translocation of masses. Journal of mathematical sciences **133**(4) (2006)
16. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11976–11986 (2022)
17. Lu, M.Y., Chen, B., Zhang, A., Williamson, D.F., Chen, R.J., Ding, T., Le, L.P., Chuang, Y.S., Mahmood, F.: Visual language pretrained multiple instance zero-shot transfer for histopathology images. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 19764–19775 (2023)
18. Ma, X., Gong, Y., He, P., Zhao, H., Duan, N.: Query rewriting in retrieval-augmented large language models. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. pp. 5303–5315 (2023)
19. Nayar, R., Wilbur, D.C.: The Bethesda system for reporting cervical cytology: definitions, criteria, and explanatory notes. Springer (2015)
20. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
21. Saslow, D., Solomon, D., Lawson, H.W., Killackey, M., Kulasingam, S.L., Cain, J., Garcia, F.A., Moriarty, A.T., Waxman, A.G., Wilbur, D.C., et al.: American cancer society, american society for colposcopy and cervical pathology, and american society for clinical pathology screening guidelines for the prevention and early detection of cervical cancer. American journal of clinical pathology **137**(4), 516–542 (2012)
22. Shen, Z., Fei, M., Wang, X., Cai, J., Wang, S., Zhang, L., Wang, Q.: Two-stage cytopathological image synthesis for augmenting cervical abnormality screening. arXiv preprint arXiv:2402.14707 (2024)
23. Sun, Y., Zhu, C., Zheng, S., Zhang, K., Sun, L., Shui, Z., Zhang, Y., Li, H., Yang, L.: Pathasst: A generative foundation ai assistant towards artificial general intelligence of pathology. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 5034–5042 (2024)

24. Tian, X., Zou, S., Yang, Z., Zhang, J.: Argue: Attribute-guided prompt tuning for vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 28578–28587 (2024)
25. Yu, S., Feng, X., Wang, B., Dun, H., Zhang, S., Zhang, R., Huang, X.: Automatic classification of cervical cells using deep learning method. IEEE Access **9**, 32559–32568 (2021)
26. Zhang, Y., Gao, J., Zhou, M., Wang, X., Qiao, Y., Zhang, S., Wang, D.: Text-guided foundation model adaptation for pathological image classification. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 272–282. Springer (2023)
27. Zheng, Z., Wei, J., Hu, X., Zhu, H., Nevatia, R.: Large language models are good prompt learners for low-shot image classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 28453–28462 (2024)
28. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Conditional prompt learning for vision-language models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16816–16825 (2022)
29. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Learning to prompt for vision-language models. International Journal of Computer Vision **130**(9), 2337–2348 (2022)
30. Zuo, J., Hong, J., Zhang, F., Yu, C., Zhou, H., Gao, C., Sang, N., Wang, J.: Plip: Language-image pre-training for person representation learning. arXiv preprint arXiv:2305.08386 (2023)