Cyto-SSL: A Self-Supervised Pretraining Framework for Cytology Foundation Model

Yiming Zhang[1*], Rui Yan[2,3*], Xiaohua Wan[1†], Yifan Zhao[1], Shuang Feng[1], Zhetao Xu[1], Ying Wang[4†], Fa Zhang[1†], Bin Hu[1†]

[1]School of Medical Technology, Beijing Institute of Technology
[2]School of Biomedical Engineering, University of Science and Technology of China
[3]Suzhou Institute for Advanced Research, University of Science and Technology of China
[4]Department of Pathology, Beijing Obstetrics and Gynecology Hospital, Capital Medical University
{wanxiaohua, zhangfa, bh}@bit.edu.cn, wangying_blk@ccmu.edu.cn

Abstract

Cytological images originate from exfoliated cells, collected via liquid-based slides and digitized into whole slide images (WSIs). Unlike histological WSIs that exhibit continuous and well-structured tissue, cytological WSIs are sparse in spatial distribution and unstructured in cellular relationships. Typically, the nucleus serves as the primary diagnostic feature, while surrounding cytoplasmic information plays a supportive role. These unique characteristics limit the development of effective foundation models and hinder the transferability of histology-based models for cytopathology. To address this, we propose **Cyto-SSL**, the first self-supervised pretraining framework for cytological images. It introduces *Nuclei-Centered Perturbation*, which highlights individual nuclei by perturbing non-nuclear regions. We also design an *SR-Transformer* module, which complements this by using sparse attention to concentrate on diagnostically relevant scattered cells, while iRPE helps model to capture local spatial relationships and avoids unnecessary attention to irrelevant global structures. Experimental results show that Cyto-SSL enhances performance across diverse cytological datasets and multiple instance learning methods. On a WSI-level dataset, it achieved 95.67% accuracy and outperformed ImageNet-pretrained ResNet-50 by 11.33%, highlighting its enhanced capability for cytological analysis. Additionally, Cyto-SSL modules are plug-and-play, easily integrated into other pretraining frameworks, yielding a 2.6% accuracy gain across different SSL methods.

Introduction

Cytological images are derived from exfoliated cells, which are collected using liquid-based slides and then digitally scanned to create whole slide images (WSIs) (Davey et al. 2006; Lee et al. 2011). These images are crucial for early cancer detection, offering a non-invasive and cost-effective method to identify abnormal cells (Fitzgerald et al. 2022; Li et al. 2020), particularly in breast, cervical, and lung cancers (Garud et al. 2017; Zhang et al. 2022b; Teramoto et al. 2017). Despite the remarkable advancements of foundation models in histopathology, progress in the cytopathology domain has lagged behind.

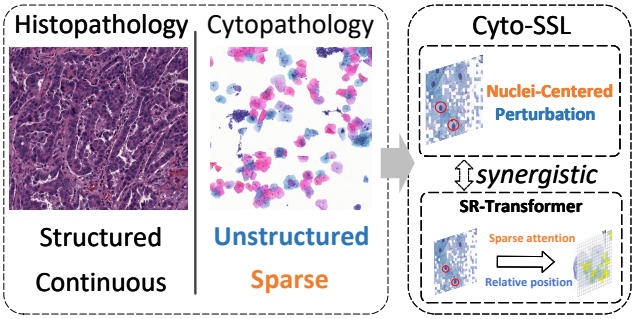

Figure 1: Comparison between histopathology and cytopathology images, along with a schematic illustration of the proposed Cyto-SSL method.

In cytopathological diagnosis, the nucleus is the central morphological feature for clinical assessment. The surrounding cytoplasm plays a supporting role, mainly helping to assess the nucleus-to-cytoplasm ratio. However, existing self-supervised pretraining methods struggle to effectively capture these critical features due to two fundamental challenges (Figure 1) of cytological images: First, cells are spatially sparse, leading to a high proportion of background and blank regions (Awan et al. 2021). Second, cellular relationships are semantically unstructured, with limited intercellular continuity and organization. Due to these unique characteristics, as well as the limited availability of cytological datasets (Jiang et al. 2023), it is difficult to develop effective methods for cytopathology. Currently, some studies (Cheng et al. 2021; Wang et al. 2024) use ResNet pretrained on ImageNet for feature extraction and classification. However, due to the significant differences between natural images and cytopathology, these methods perform poorly. Other approaches have applied Variational Positive-Unlabeled (VPU) learning (Chen et al. 2020a; Zhao et al. 2024), which is specifically optimized for cytopathology. Despite its adaptation to cytological features, the method's reliance on CNN architectures still leads to suboptimal performance, particularly when dealing with the sparse and un-

*These authors contributed equally.
†Corresponding authors.

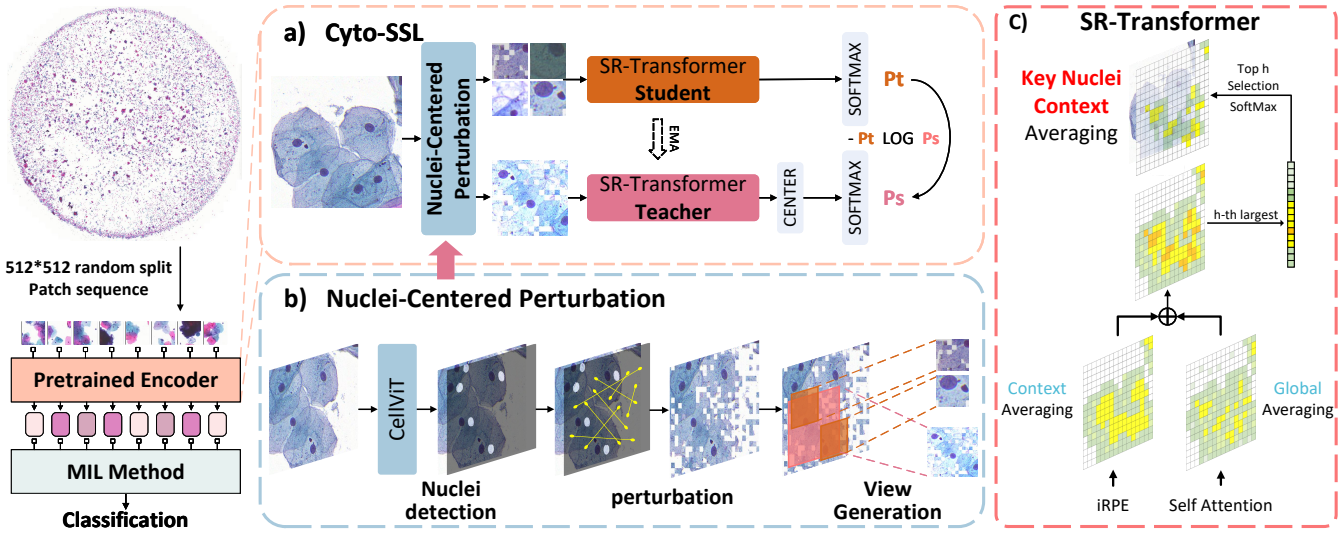

Figure 2: Overview of the proposed Cyto-SSL framework. a) The Cyto-SSL pretraining framework adopts a contrastive learning pipeline with a teacher-student architecture. b) Nuclei-Centered Perturbation generates nucleus-centered multi-scale views by perturbing the surrounding environment. c) Detailed architecture of the SR-Transformer, which integrates sparse self-attention and relative position encoding for enhanced nuclear feature extraction.

structured cytological images. Therefore, there is a need for further development of methods specifically tailored for cytopathology.

In contrast, several histopathological foundation models provide high-quality feature representations, which enhance the performance of subsequent multiple instance learning (MIL) methods (Xu et al. 2024; Chen et al. 2024; Lu et al. 2024). Foundation models for histopathology often use self-supervised training frameworks like DINO (Caron et al. 2021), MoCo (He et al. 2020), and SimCLR (Chen et al. 2020b). These methods learn global features, which work well for histopathological images with continuous and structured tissue. However, this approach is not suitable for cytopathology. Cytological images have sparse cell distribution and lack coherent global structure. Although early efforts like CytoFM (Ivezic et al. 2025) have been made for cytopathology, it does not tackle the sparsity and unstructured morphology and current histopathological models do not transfer effectively to cytological WSI data (Huang et al. 2025). Therefore, there is a need to design a dedicated training framework that can tackle these challenges and train a foundational model for cytopathology from scratch.

In this paper, we propose **Cyto-SSL**, a novel self-supervised pretraining framework designed for cytological foundation model. The contributions of this paper are summarized as follows:

- We propose the first pretraining framework specifically designed for cytopathological images, providing high-quality feature representations tailored for downstream tasks.
- We introduce *Nuclei-Centered Perturbation*, a nuclear-guided view generation strategy designed to direct model focus toward diagnostically salient nuclear regions, while attenuating the influence of surrounding contextual information.
- We develop an *SR-Transformer* module that complements the perturbation strategy by employing sparse self-attention to capture sparse nuclei and image relative positional encoding (iRPE) to model local cellular relationships without relying on global structures.
- We validate Cyto-SSL on both private and public datasets, achieving state-of-the-art performance across multiple MIL methods.

Method

In this section, we introduce Cyto-SSL, a self-supervised learning framework specifically designed for pretraining on cytology WSIs. The goal of Cyto-SSL is to improve the ViT model's ability to recognize diagnostically important nuclei, while addressing the challenges of spatial sparsity and semantic looseness in cytological images. As shown in Figure 2, the following subsections describe the overall training process and key components of the Cyto-SSL framework. The core of our method are two novel components tailored for cytological characteristics: the *Nuclei-Centered Perturbation* module that generates diagnostic-aware views by focusing on nuclear regions, and the *SR-Transformer* that enhances attention on critical nuclei and captures local spatial dependencies.

Overall Architecture of Cyto-SSL

Our Cyto-SSL framework adopts a teacher-student self-distillation architecture for self-supervised learning, as shown in Figure 2a. Self-distillation is well-suited for cytological images, as it eliminates the need for explicit negative samples and avoids semantic ambiguity. By aligning

representations from different views of the same region, the model learns to focus on diagnostically relevant nuclei, improving feature consistency in sparse and unstructured settings. Cyto-SSL follows a dual-network setup, where the student network f_s learns from the teacher network f_t. Both networks receive different augmented views of the same input image x, denoted x_s and x_t, respectively. The student model learns from the teacher model by minimizing the cross-entropy loss between the teacher's probability distribution and the student's logits on the class token:

$$\mathcal{L}_{\text{Cyto-SSL}} = -\sum_{i=1}^{K} p_t^{(i)}(f_t(x_t)) \log p_s^{(i)}(f_s(x_s)), \quad (1)$$

where p_t and p_s denote the softmax-normalized output probabilities from the teacher and student networks, respectively, and K is the number of prototype dimensions.

After the student model is trained, the teacher network is updated using an exponential moving average (EMA) of the student's weights:

$$\theta_t \leftarrow \tau\theta_t + (1-\tau)\theta_s, \quad (2)$$

where θ_t and θ_s are the parameters of the teacher and student networks, and τ is the momentum coefficient (typically close to 1).

In addition to the architecture, we introduce Nuclei-Centered Perturbation and SR-Transformer as two key components. First, Nuclei-Centered Perturbation, as shown in Figure 2b, generates views centered on nuclei, encouraging the model to attend to diagnostically meaningful cells while mitigating the impact of spatial sparsity and semantic looseness. Second, SR-Transformer modules, as shown in Figure 2c, are incorporated into both networks. By combining sparse self-attention and image-relative positional encoding (iRPE), it enhances the model's ability to focus on scattered diagnostic nuclei and capture local cellular relationships without relying on global structure.

Nuclei-Centered Perturbation

Nuclei-Centered Perturbation is a core component of the Cyto-SSL framework. By explicitly guiding the model to focus on diagnostically relevant nuclear regions and the local structural context, while suppressing irrelevant background structures, it addresses the unique structural characteristics of cytology images. Cytological images exhibit spatial sparsity, as cells are sparsely distributed, and semantic looseness, since diagnostically relevant cells often lack strong contextual relationships with their surrounding cells. In many self-supervised learning methods, random view sampling tends to generate patches dominated by irrelevant content. These views dilute the training signal and hinder effective feature learning. To address these challenges, we introduce Nuclei-Centered Perturbation, a view generation strategy that focuses sampling around nuclei to reduce the effects of both spatial sparsity and semantic looseness. A pseudo-code implementation is presented in Algorithm 1. The process includes three main steps:

Algorithm 1: Nuclei-Centered Perturbation for View Generation in Cytology

Input: Cytology image I
Parameters: Nucleus radius r, scales s_g, s_l, views per type n_g, n_l
Output: View set $\mathcal{V}$

$\mathcal{P} \leftarrow \text{DETECTNUCLEI}(I)$
$I' \leftarrow \text{PERTURB}(I, \mathcal{P}, r)$
Initialize $\mathcal{V} \leftarrow \emptyset$
for $i = 1$ to n_g **do**
  $p \leftarrow \text{SAMPLEPOINT}(\mathcal{P})$
  $v \leftarrow \text{CROPAUG}(I', p, s_g)$
  $\mathcal{V} \leftarrow \mathcal{V} \cup \{v\}$
end for
for $i = 1$ to n_l **do**
  $p \leftarrow \text{SAMPLEPOINT}(\mathcal{P})$
  $v \leftarrow \text{CROPAUG}(I', p, s_l)$
  $\mathcal{V} \leftarrow \mathcal{V} \cup \{v\}$
end for
return $\mathcal{V}$

Step 1: Nuclei Detection. To localize nuclei without requiring precise segmentation annotations, we adopt a deep learning-based nuclear segmentation model. While various models such as CellViT (Hörst et al. 2024) and Cellpose (Stringer et al. 2021) are suitable, we employ CellViT due to its strong recognition capability and transformer-based architecture. It is trained on public cytological datasets and generalizes well to diverse data. During Cyto-SSL training, CellViT's weights remain frozen to ensure efficient computation and stable feature localization.

Step 2: Perturbation. After locating the nuclei, we apply perturbations to the image. The core principle is to preserve the key diagnostic features, as the nucleus and surrounding cytoplasm are crucial for cytopathological analysis. To this end, we randomly retain part of the nucleus while perturbing the surrounding areas, including the cytoplasm and neighboring cells. Specifically, we randomly shuffle pixels outside a defined radius r from the nucleus to disrupt irrelevant background structures and enhance focus on diagnostic regions. The radius is set to $r = 50$ pixels, which corresponds to a meaningful cytoplasmic context at $40\times$ magnification. Within this range, morphological, color, and textural features critical for cytological diagnosis are typically preserved. This ensures that the model retains relevant perinuclear information while reducing distraction from distant background.

Step 3: Cytology View Generation. Finally, we apply random transformations such as scaling, rotation, and color jittering to the nuclear-centered patches in order to generate diverse augmented views. These transformations simulate variations in imaging conditions and cellular morphology, improving generalization. The sampling scales are adapted to match the view policy of each self-supervised learning framework. For instance, in SimCLR, both views are generated at similar scales to enable contrastive feature align-

ment, whereas in DINO, different scale ranges are used to promote multi-scale feature learning. This framework-aware augmentation design ensures that Nuclei-Centered Perturbation produces compatible and effective views for different paradigms, including SimCLR, MoCo, and BYOL.

SR-Transformer

SR-Transformer is a key component of the Cyto-SSL framework, designed to enhance cytological feature extraction by integrating sparse self-attention and image relative position encoding (iRPE).

In standard self-attention, the attention score matrix P is computed as:

$$P = \frac{QK^T}{\sqrt{d}}, \tag{3}$$

where Q, K, and V denote the query, key, and value matrices, respectively, and d is the feature dimension. The self-attention output is:

$$\text{Att}(Q, K, V) = \text{Softmax}(P) \cdot V. \tag{4}$$

However, this uniform treatment of all token pairs has two drawbacks in cytology: (1) It fails to emphasize sparse but critical diagnostic nuclei; (2) Absolute position encoding may causes misleading long-range dependencies from irrelevant cells. To address these issues, SR-Transformer introduces two enhancements: sparse self-attention and image relative position encoding.

Sparse Self-Attention. Sparse self-attention restricts computation to the top-h most relevant tokens per query, suppressing non-informative signals and highlighting diagnostic nuclei (Zhao et al. 2019; Yan et al. 2023). A binary mask M is applied to the attention score matrix P as follows:

$$M(P, h)_{ij} = \begin{cases} P_{ij}, & \text{if } P_{ij} \geq t_i, \\ -\infty, & \text{otherwise}, \end{cases} \tag{5}$$

where t_i is the h-th largest value in the i-th row. The model retains the most important attention contributions while discarding irrelevant information or noise. Since h is typically small, such as 16, this not only focuses attention on critical regions but also simplifies the model and reduces computational burden. The sparse attention output is:

$$\text{SparseAtt}(Q, K, V, h) = \text{Softmax}(M(P, h)) \cdot V. \tag{6}$$

This sparse attention mechanism improves model performance by limiting attention to the most relevant tokens, which is crucial in cytology where diagnostic nuclei are sparse and limited. This ensures that attention is directed toward the diagnostic nuclei and relevant regions, enhancing feature representation in sparse cytological images.

Image Relative Position Encoding. To better capture spatial context in cytology, we employ iRPE to encode the relative positions between tokens. The relative position matrix R is computed as:

$$R = Q(r^K)^T + K(r^Q)^T, \tag{7}$$

where r^Q and r^K are learnable positional vectors. The modified attention score becomes:

$$P = \frac{QK^T + R}{\sqrt{d}}. \tag{8}$$

We also introduce a learnable bias r^V to the value matrix:

$$\text{SR-Att}(Q, K, V, h) = \text{Softmax}(M(P, h)) \cdot (V + r^V). \tag{9}$$

This formulation improves sensitivity to local structures while preserving spatial awareness. Furthermore, following findings from iRPE research (Wu et al. 2021), we retain the absolute positional encoding alongside iRPE, which has been shown to benefit downstream tasks such as object detection. This design choice ensures broader adaptability.

In summary, SR-Transformer improves the analysis of cytological images by (1) emphasizing diagnostically relevant nuclei through sparse self-attention, and (2) capturing local spatial cues via relative position encoding.

Experiments and Results

Datasets and Evaluation Metrics

We conducted experiments on three cytological image datasets: (1) a private Cervical Cytologic WSI dataset from collaborating hospital, consisting of 760 WSI samples scanned at 40× magnification with a resolution of approximately 80,000×80,000 pixels, with negative and positive labels; (2) FNAC 2019 (Saikia et al. 2019), a public dataset with 212 breast cytology images at a resolution of 2,048×1,536 pixels, labeled as benign and malignant; and (3) NIH-NLM Thin Blood Smears Pf (Yu et al. 2020), a public dataset with 965 blood smear images from 193 malaria patients, each with a resolution of 5,312×2,988 pixels, with negative and positive labels. The Cervical Cytologic WSI dataset was used for pretraining, while the other datasets were used for fine-tuning and evaluation to test the generalizability of the model. We evaluate model performance using AUC, ACC, and Recall.

Implementation Details

We followed a typical foundation model pipeline widely used in histopathology: pretraining, feature extraction, and MIL-based classification. All experiments were conducted on a workstation equipped with Intel Core i9-14900K CPUs and NVIDIA RTX 4090 GPUs.

In the pretraining phase, we adopted ResNet-50, DINOV2 (initialized with ImageNet weights), Prov-Gigapath, and ViT-tiny as backbone networks, pretrained using Cyto-SSL on the entire private Cervical Cytologic WSI dataset, which includes 2.5 million 40× magnification patches. Cyto-SSL was also integrated into SimCLR, MoCo, and DINO, with Nuclei-Centered Perturbation adapted to each framework's view scale policy. Training used the Adam optimizer (initial learning rate of 5e-4) with cosine annealing. Considering Prov-Gigapath was pretrained on histopathology data, we fine-tuned it using the LoRA strategy.

In the feature extraction phase, for the Cervical Cytologic WSI dataset, we sampled 100 non-overlapping patches

Encoder Model	MIL method	Cervical Cytologic WSI dataset			FNAC 2019			NIH-NLM-Thin Blood Smears Pf		
		AUC	ACC	Recall	AUC	ACC	Recall	AUC	ACC	Recall
ResNet-50	ABMIL	0.829	83.04	54.10	0.997	98.44	97.06	0.867	83.33	96.94
Prov-GigaPath		0.970	92.50	85.71	0.995	92.19	1.00	0.869	89.17	94.16
DINO V2		0.825	86.15	69.36	0.975	90.62	1.00	0.879	87.50	92.40
Ours		**0.974**	**94.37**	**88.71**	1.00	1.00	1.00	**0.912**	**90.83**	**97.08**
ResNet-50	TransMIL	0.900	83.91	44.26	1.00	1.00	1.00	0.930	87.50	90.64
Prov-GigaPath		**0.940**	90.90	80.64	1.00	1.00	1.00	0.928	**92.08**	**95.32**
DINO V2		0.902	87.45	74.19	1.00	1.00	1.00	**0.932**	90.00	91.81
Ours		0.915	**93.94**	**83.87**	1.00	1.00	1.00	0.911	89.17	93.57
ResNet-50	DTFD-MIL	0.914	87.39	82.38	1.00	98.25	95.65	0.919	91.35	94.88
Prov-GigaPath		0.942	91.34	79.03	1.00	99.56	98.91	**0.946**	**93.33**	**97.08**
DINO V2		0.901	86.15	**90.32**	1.00	99.12	98.91	0.908	90.31	95.18
Ours		**0.977**	**92.21**	**90.32**	1.00	1.00	1.00	0.942	90.31	94.30
ResNet-50	CLAM-SB	0.859	83.62	59.68	1.00	1.00	1.00	0.931	90.00	**99.42**
Prov-GigaPath		0.921	87.45	70.97	1.00	1.00	1.00	0.938	**92.08**	97.66
DINO V2		0.799	80.95	61.29	1.00	1.00	1.00	0.942	**92.08**	97.66
Ours		**0.968**	**91.34**	**90.32**	0.988	98.44	97.06	**0.944**	90.42	94.15
ResNet-50	DSMIL	0.894	87.87	77.41	1.00	1.00	1.00	0.874	86.66	91.22
Prov-GigaPath		**0.956**	91.32	**93.54**	1.00	1.00	1.00	0.928	90.41	94.15
DINO V2		0.889	88.42	72.58	1.00	1.00	1.00	0.823	84.58	93.56
Ours		0.951	**92.64**	88.70	1.00	1.00	1.00	**0.941**	**91.25**	**94.73**
CNN(VPU)	LESS	0.928	87.87	70.96	0.985	98.43	1.00	0.786	81.66	**98.24**
Prov-GigaPath		0.928	88.31	70.96	1.00	1.00	1.00	0.872	**90.83**	97.07
DINO V2		0.818	80.95	64.51	1.00	1.00	1.00	0.886	88.33	94.73
Ours		**0.966**	**95.67**	**88.70**	1.00	1.00	1.00	**0.939**	90.41	96.49

Table 1: Evaluation of the transferability of the ViT-tiny encoder pretrained with Cyto-SSL across multiple MIL methods and cytology datasets. The highest performance is marked in **bold**, and the second highest in underlined. Compared to ImageNet-pretrained ResNet, DINOV2 and Prov-Gigapath, Cyto-SSL consistently provides superior feature representations. Notably, our pretrained model achieves the best performance when combined with the LESS framework.

(512×512 pixels) per WSI. For the FNAC and NIH-NLM Thin Blood Smears Pf datasets, we extracted patches (256×256 pixels) using a sliding window.

In the MIL classification phase, we evaluated ABMIL (Ilse, Tomczak, and Welling 2018), TransMIL (Shao et al. 2021), DTFD-MIL (Zhang et al. 2022a), CLAM-SB (Lu et al. 2021), DSMIL (Li, Li, and Eliceiri 2021), and LESS (Zhao et al. 2024), all configured in accordance with their official implementations. Models were tested on three datasets, split into 70% training and 30% testing.

Incorporating foundation model into MIL methods

To validate the representation capability of the ViT-tiny encoder trained with Cyto-SSL, we replaced the original feature extractors in several state-of-the-art MIL methods, while keeping all other settings unchanged. While most methods originally used ResNet-50 pretrained on ImageNet or VPU-pretrained CNNs (e.g., in LESS), we directly swapped in our encoder for fair comparison.

As shown in Table 1, Cyto-SSL consistently outperformed baselines across three cytological datasets. On the proprietary Cervical Cytologic WSI dataset, it significantly improved AUC and accuracy. For instance, when combined with ABMIL, it achieved 0.974 AUC and 94.37% accuracy, outperforming ResNet-50 by 0.145 AUC and 11.33% accuracy. With LESS, it reached 0.966 AUC and 95.67% accuracy, exceeding the VPU-CNN baseline by 0.038 and 7.80%, respectively. Similar gains were observed with TransMIL, CLAM-SB, and DTFD-MIL, showing that Cyto-SSL provides more discriminative cytological features. In addition, when compared with DINOV2 and Prov-Gigapath, fine-tuned via LoRA on cytological data, Cyto-SSL still achieved higher performance, demonstrating that training a foundation model from scratch on cytopathology yields better representations than transferring from histopathology.

Interestingly, ABMIL consistently outperformed complex frameworks such as DTFD-MIL and TransMIL, contrary to trends in histopathology. This suggests that the sparse and unstructured cell distribution in cytology challenges the assumptions underlying histology-based MIL methods, thereby highlighting the importance of adapting to cytology-specific characteristics.

Consistent improvements were also observed on the FNAC 2019 and NIH-NLM Thin Blood Smears Pf datasets.

Despite lower resolution and limited scale, our encoder maintained strong generalization. On FNAC, it reached 100% accuracy in just 5 epochs, while ResNet-50 and VPU-based models required 20–100 epochs to converge. These results further confirm the robustness of Cyto-SSL across diverse cytological settings.

Ablation Study

To evaluate the contribution of each key component in Cyto-SSL, we conducted an ablation study focusing on their impact on the feature quality of pretrained models. Specifically, we tested the effect of the Nuclei-Centered Perturbation, sparse self-attention, and image relative position encoding (iRPE) within the DINO framework. The pretrained models were evaluated with two representative MIL classifiers, ABMIL and LESS. The results are summarized in Table 2.

	MIL	AUC	ACC	Recall
DINO		0.938	91.77	79.03
NC Perturbation		0.959	93.07	85.48
Sparse self attention	ABMIL	0.964	93.51	85.48
iRPE		0.958	92.21	85.48
Ours		0.974	94.37	85.48
DINO		0.963	92.20	83.87
NC Perturbation		0.967	93.50	87.09
Sparse self attention	LESS	0.952	93.07	88.70
iRPE		0.965	93.07	87.09
Ours		0.966	95.67	88.70

Table 2: Ablation study on various components of Cyto-SSL: Nuclei-Centered (NC) Perturbation, sparse self-attention, and iRPE. These components were separately added to the self-distillation architecture, with ABMIL and LESS used as downstream MIL classifiers.

Our findings show that incorporating the Nuclei-Centered Perturbation consistently improved feature quality across different MIL methods, highlighting the importance of targeted nuclear-centric view generation in cytological image representation. The effects of sparse self-attention and iRPE, however, were more nuanced. These modules provided noticeable performance gains when paired with ABMIL, but showed only marginal improvements with LESS. The better performance of ABMIL can be attributed to its attention mechanism, which focuses on the most diagnostically relevant regions in sparse cytological images. By emphasizing important nuclei, ABMIL avoids being distracted by irrelevant background or non-diagnostic cells. In contrast, LESS, which uses cross-attention, still maintains attention across the entire image, including non-diagnostic areas, which reduces its ability to focus on the sparse diagnostic cells.

Nevertheless, when all three components were integrated into the complete Cyto-SSL framework, the models consistently achieved the highest performance. This demonstrates that the combination of Nuclei-Centered Perturbation and SR-Transformer leads to superior cytological feature representations. This further highlights their complementary nature: one guides the model to focus on diagnostic nuclei through view design, while the other enhances representation by modeling sparsity and local structure.

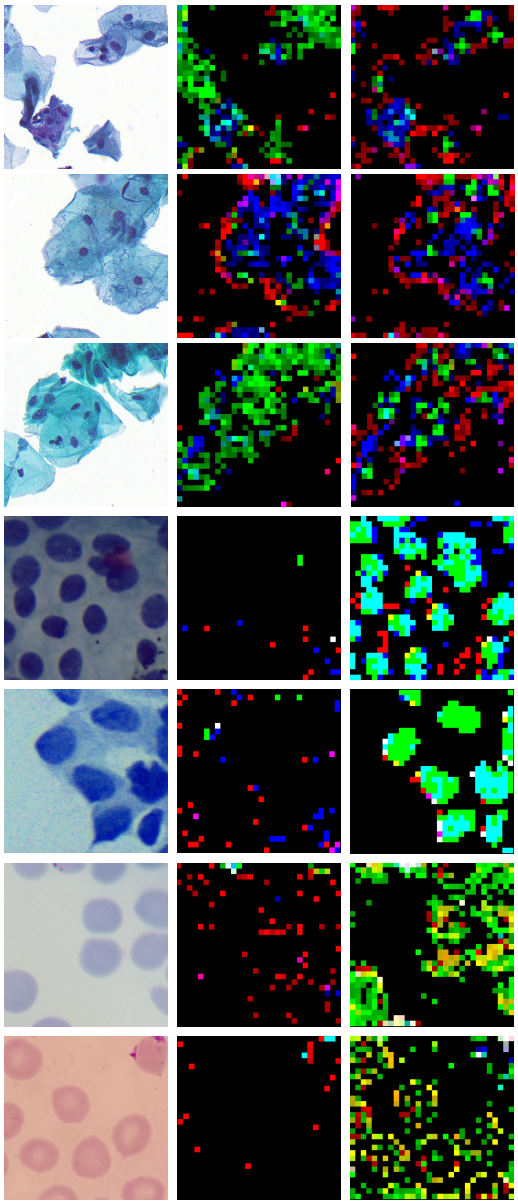

Figure 3: Attention visualization from Cyto-SSL. Left: original cytology patches. Middle: DINO's colored attention maps. Right: Cyto-SSL's colored attention maps-green (nuclei), red (cytoplasm), blue (artifacts). Rows 1–3, 4–5, and 6–7 correspond to the Cervical Cytologic dataset, FNAC 2019 dataset, and NIH-NLM Thin Blood Smears Pf dataset, respectively.

Probing the self-attention map

To further investigate the impact of Cyto-SSL, we visualized the last-layer multi-head attention maps of ViT-tiny models from three datasets. The results, shown in Figure 3, highlight

a clear contrast between models trained with and without Cyto-SSL.

In models trained with Cyto-SSL, attention heads focused on diagnostically relevant areas. Green heads focused on cell nuclei, red heads highlighted cytoplasm, and blue heads responded to artifacts. The model successfully distinguished between relevant regions and noise such as background clutter and neutrophils. In contrast, the DINO baseline without Cyto-SSL struggled to accurately locate the nuclei, with attention dispersed across irrelevant regions. This reduced its ability to differentiate diagnostic feature from sparse and unstructured cytological images.

Despite not being pretrained on the FNAC and Malaria datasets, the models still focused attention on nuclei, demonstrating that Cyto-SSL enhances model generalization across different cytological datasets. These findings confirm that Cyto-SSL improves attention allocation and feature learning, leading to better classification performance.

Incorporating Cyto-SSL into different SSL methods

In this experiment, we evaluated the effectiveness of Cyto-SSL as a plug-and-play module across multiple self-supervised learning (SSL) frameworks. Three popular SSL methods were selected: SimCLR, MoCo, and DINO. Each framework was used to pretrain a ViT-tiny model on our Cervical Cytologic dataset.

As shown in Table 3, Cyto-SSL demonstrated consistent and substantial performance improvements across all four SSL frameworks. For example, in the SimCLR setting, Cyto-SSL enhanced accuracy by 1.73%. In the MoCo framework, Cyto-SSL boosted AUC and ACC by 0.037 and 0.86% respectively. DINO, a widely used method for histopathological image pretraining, achieved the best overall performance when combined with Cyto-SSL, yielding a 0.036 increase in AUC and a 2.6% improvement in accuracy.

These results show that Cyto-SSL works effectively as a plug-and-play module across various SSL frameworks. Its compatibility with different pretraining architectures and consistent performance improvements highlight its generalizability. Applied to cytological image pretraining, Cyto-SSL significantly boosts model performance, providing a robust solution for enhancing feature representation across SSL methods.

Discussion

The proposed Cyto-SSL framework effectively improves feature extraction in cytological image analysis, consistently enhancing performance across diverse datasets and MIL methods. Its gains stem from better focus on diagnostic nuclei from sparse and unstructured cytological image, achieved through sparse attention and iRPE. Attention visualizations confirm that Cyto-SSL-trained models effectively concentrate on sparse and semantically critical nuclei regions, while baseline models exhibit dispersed attention across spatially and semantically irrelevant areas, failing to address the unique inherent in cytological images.

SSL-method	Fine-tuning		
	AUC	ACC	Recall
SimCLR	0.932	88.31	79.03
SimCLR+Cyto-SSL	0.915	90.04	79.03
MoCo	0.896	87.88	69.35
MoCo+Cyto-SSL	0.933	88.74	77.42
DINO	0.938	91.77	79.03
DINO+Cyto-SSL	0.974	94.37	85.48

Table 3: Performance comparison of different self-supervised learning (SSL) frameworks with the proposed Cyto-SSL on Cervical Cytologic WSI dataset.

Our experiments show that MIL methods perform differently in cytology due to structural differences with histology. While methods such as DTFD-MIL and TransMIL that are initially designed for dense histological tissues exhibit suboptimal performance, ABMIL achieves better performance by selecting sparse diagnostic signals. LESS, a cytology-specific method, achieves state-of-the-art performance, confirming Cyto-SSL as a robust feature extractor for cytology that works well with various MIL frameworks. ABMIL's strong performance also highlights its potential as a baseline for evaluating cytology encoders.

However, our pretraining data is currently limited to a single-center dataset. Future work could expand the dataset to include multi-center and diverse cytopathological images, and use larger model architectures like ViT-S and ViT-B. These improvements should enhance the model's performance and generalization, potentially enabling exploration of scaling laws in cytopathology. Additionally, despite the structural differences between histopathology and cytopathology, both are part of the broader pathology image category. This suggests potential benefits of joint pretraining on both modalities for foundation model development. Moreover, integrating pathological text reports or multi-omics data (Yan et al. 2025) into pretraining can boost multi-modal capabilities.

Conclusion

In this work, we introduced Cyto-SSL, a self-supervised pretraining framework designed for cytological images. Given the substantial differences between cytopathological and histopathological images, especially the sparse and unstructured nature of cytological images, we incorporate Nuclei-Centered Perturbation and the SR-Transformer, which combines sparse self-attention with relative positional encoding. Our experiments across three cytological datasets and multiple MIL methods show that Cyto-SSL significantly improves performance, achieving state-of-the-art results. Attention map visualizations confirm that it helps models focus on diagnostically meaningful regions. Our proposed cytopathology pretraining approach paves the way for developing cytopathology foundation models, ultimately advancing pathology artificial intelligence and precision medicine.

Acknowledgments

This study was funded by the National Natural Science Foundation of China (62402473), Beijing Natural Science Foundation (L252175), the National Natural Science Foundation of China (No. W2511070, 32241027, 62472034, 62227807, 62272326), in part by the National Key R&D Program of China (No. 2019YFA0706200, 2019YFA0706200).

References

Awan, R.; Benes, K.; Azam, A.; Song, T.-H.; Shaban, M.; Verrill, C.; Tsang, Y. W.; Snead, D.; Minhas, F.; and Rajpoot, N. 2021. Deep learning based digital cell profiles for risk stratification of urine cytology images. *Cytometry Part A*, 99(7): 732–742.

Caron, M.; Touvron, H.; Misra, I.; Jégou, H.; Mairal, J.; Bojanowski, P.; and Joulin, A. 2021. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9650–9660.

Chen, H.; Liu, F.; Wang, Y.; Zhao, L.; and Wu, H. 2020a. A variational approach for learning from positive and unlabeled data. *Advances in Neural Information Processing Systems*, 33: 14844–14854.

Chen, R. J.; Ding, T.; Lu, M. Y.; Williamson, D. F.; Jaume, G.; Song, A. H.; Chen, B.; Zhang, A.; Shao, D.; Shaban, M.; et al. 2024. Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 30(3): 850–862.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020b. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, 1597–1607. PmLR.

Cheng, S.; Liu, S.; Yu, J.; Rao, G.; Xiao, Y.; Han, W.; Zhu, W.; Lv, X.; Li, N.; Cai, J.; et al. 2021. Robust whole slide image analysis for cervical cancer screening using deep learning. *Nature Communications*, 12(1): 5639.

Davey, E.; Barratt, A.; Irwig, L.; Chan, S. F.; Macaskill, P.; Mannes, P.; and Saville, A. M. 2006. Effect of study design and quality on unsatisfactory rates, cytology classifications, and accuracy in liquid-based versus conventional cervical cytology: a systematic review. *The Lancet*, 367(9505): 122–132.

Fitzgerald, R. C.; Antoniou, A. C.; Fruk, L.; and Rosenfeld, N. 2022. The future of early cancer detection. *Nature Medicine*, 28(4): 666–677.

Garud, H.; Karri, S. P. K.; Sheet, D.; Chatterjee, J.; Mahadevappa, M.; Ray, A. K.; Ghosh, A.; and Maity, A. K. 2017. High-magnification multi-views based classification of breast fine needle aspiration cytology cell samples using fusion of decisions from deep convolutional networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 76–81.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9729–9738.

Hörst, F.; Rempe, M.; Heine, L.; Seibold, C.; Keyl, J.; Baldini, G.; Ugurel, S.; Siveke, J.; Grünwald, B.; Egger, J.; et al. 2024. Cellvit: Vision transformers for precise cell segmentation and classification. *Medical Image Analysis*, 94: 103143.

Huang, J.; Li, G.; Kan, S.; Liu, J.; and Liang, Y. 2025. An efficient framework based on large foundation model for cervical cytopathology whole slide image screening. *Biomedical Signal Processing and Control*, 107: 107859.

Ilse, M.; Tomczak, J.; and Welling, M. 2018. Attention-based deep multiple instance learning. In *International Conference on Machine Learning*, 2127–2136. PMLR.

Ivezic, V.; Radhachandran, A.; Redekop, E.; Athreya, S.; Lee, D.; Sant, V.; Arnold, C.; and Speier, W. 2025. CytoFM: The first cytology foundation model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4749–4757.

Jiang, P.; Li, X.; Shen, H.; Chen, Y.; Wang, L.; Chen, H.; Feng, J.; and Liu, J. 2023. A survey on deep learning-based cervical cytology screening: from cell identification to whole slide image analysis.

Lee, R. E.; McClintock, D. S.; Laver, N. M.; and Yagi, Y. 2011. Evaluation and optimization for liquid-based preparation cytology in whole slide imaging. *Journal of Pathology Informatics*, 2(1): 46.

Li, B.; Li, Y.; and Eliceiri, K. W. 2021. 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*, 14318–14328.

Li, J.; Guan, X.; Fan, Z.; Ching, L.-M.; Li, Y.; Wang, X.; Cao, W.-M.; and Liu, D.-X. 2020. Non-invasive biomarkers for early detection of breast cancer. *Cancers*, 12(10): 2767.

Lu, M. Y.; Chen, B.; Williamson, D. F.; Chen, R. J.; Liang, I.; Ding, T.; Jaume, G.; Odintsov, I.; Le, L. P.; Gerber, G.; et al. 2024. A visual-language foundation model for computational pathology. *Nature Medicine*, 30(3): 863–874.

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

Saikia, A. R.; Bora, K.; Mahanta, L. B.; and Das, A. K. 2019. Comparative assessment of CNN architectures for classification of breast FNAC images. *Tissue and Cell*, 57: 8–14.

Shao, Z.; Bian, H.; Chen, Y.; Wang, Y.; Zhang, J.; Ji, X.; et al. 2021. Transmil: Transformer based correlated multiple instance learning for whole slide image classification. *Advances in Neural Information Processing Systems*, 34: 2136–2147.

Stringer, C.; Wang, T.; Michaelos, M.; and Pachitariu, M. 2021. Cellpose: a generalist algorithm for cellular segmentation. *Nature Methods*, 18(1): 100–106.

Teramoto, A.; Tsukamoto, T.; Kiriyama, Y.; and Fujita, H. 2017. Automated classification of lung cancer types from cytological images using deep convolutional neural networks. *BioMed Research International*, 2017(1): 4067832.

Wang, J.; Yu, Y.; Tan, Y.; Wan, H.; Zheng, N.; He, Z.; Mao, L.; Ren, W.; Chen, K.; Lin, Z.; et al. 2024. Artificial intelligence enables precision diagnosis of cervical cytology grades and cervical cancer. *Nature Communications*, 15(1): 4369.

Wu, K.; Peng, H.; Chen, M.; Fu, J.; and Chao, H. 2021. Rethinking and improving relative position encoding for vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10033–10041.

Xu, H.; Usuyama, N.; Bagga, J.; Zhang, S.; Rao, R.; Naumann, T.; Wong, C.; Gero, Z.; González, J.; Gu, Y.; et al. 2024. A whole-slide foundation model for digital pathology from real-world data. *Nature*, 630(8015): 181–188.

Yan, R.; Lv, Z.; Yang, Z.; Lin, S.; Zheng, C.; and Zhang, F. 2023. Sparse and hierarchical transformer for survival analysis on whole slide images. *IEEE Journal of Biomedical and Health Informatics*, 28(1): 7–18.

Yan, R.; Zhang, X.; Jiang, Z.; Wang, B.; Bian, X.; Ren, F.; and Zhou, S. K. 2025. Pathway-aware multimodal transformer (PAMT): Integrating pathological image and gene expression for interpretable cancer survival analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Yu, H.; Yang, F.; Rajaraman, S.; Ersoy, I.; Moallem, G.; Poostchi, M.; Palaniappan, K.; Antani, S.; Maude, R. J.; and Jaeger, S. 2020. Malaria Screener: a smartphone application for automated malaria screening. *BMC Infectious Diseases*, 20(1): 825.

Zhang, H.; Meng, Y.; Zhao, Y.; Qiao, Y.; Yang, X.; Coupland, S. E.; and Zheng, Y. 2022a. Dtfd-mil: Double-tier feature distillation multiple instance learning for histopathology whole slide image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18802–18812.

Zhang, X.; Cao, M.; Wang, S.; Sun, J.; Fan, X.; Wang, Q.; and Zhang, L. 2022b. Whole slide cervical cancer screening using graph attention network and supervised contrastive learning. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 202–211. Springer.

Zhao, B.; Deng, W.; Li, Z. H. H.; Zhou, C.; Gao, Z.; Wang, G.; and Li, X. 2024. LESS: Label-efficient multi-scale learning for cytological whole slide image screening. *Medical Image Analysis*, 94: 103109.

Zhao, G.; Lin, J.; Zhang, Z.; Ren, X.; Su, Q.; and Sun, X. 2019. Explicit sparse transformer: Concentrated attention through explicit selection. *arXiv preprint arXiv:1912.11637*.