Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

# Self-supervised learning-based cervical cytology for the triage of HPV-positive women in resource-limited settings and low-data regime

Thomas Stegmüller [a,*], Christian Abbet [a], Behzad Bozorgtabar [a,c], Holly Clarke [b], Patrick Petignat [b], Pierre Vassilakos [b], Jean-Philippe Thiran [a,c]

[a] *Ecole Polytechnique Fédérale de Lausanne, Lausanne, 1015, Switzerland*
[b] *Hôpitaux Universitaires de Genève, Genève, 1205, Switzerland*
[c] *Centre Hospitalier Universitaire Vaudois, Lausanne, 1011, Switzerland*

ARTICLE INFO

*Keywords:*
Digital cytology
WSIs classification
Self-supervised learning
Pasting augmentation

ABSTRACT

Screening Papanicolaou test samples has proven to be highly effective in reducing cervical cancer-related mortality. However, the lack of trained cytopathologists hinders its widespread implementation in low-resource settings. Deep learning-assisted telecytology diagnosis emerges as an appealing alternative, but it requires the collection of large annotated training datasets, which is costly and time-consuming. In this paper, we demonstrate that the abundance of unlabeled images that can be extracted from Pap smear test whole slide images presents a fertile ground for self-supervised learning methods, yielding performance improvements compared to off-the-shelf pre-trained models for various downstream tasks. In particular, we propose **C**ervical **C**ell **C**opy-**P**asting ($\mathtt{C^3P}$) as an effective augmentation method, which enables knowledge transfer from public and labeled single-cell datasets to unlabeled tiles. Not only does $\mathtt{C^3P}$ outperforms naive transfer from single-cell images, but we also demonstrate its advantageous integration into multiple instance learning methods. Importantly, all our experiments are conducted on our introduced *in-house* dataset comprising liquid-based cytology Pap smear images obtained using low-cost technologies. This aligns with our long-term objective of deep learning-assisted telecytology for diagnosis in low-resource settings.

## 1. Introduction

Cervical cancer is considered nearly completely preventable but continues to be a leading cause of cancer mortality. In 2020, about 342 000 women died from this disease, most of them in developing countries where cytology-based screening programs to detect and treat precancerous lesions are not available or affordable [1].

In the knowledge that human papillomavirus (HPV) is the etiological factor that drives cervical cancer development, secondary prevention with HPV testing has, in recent years, become the preferred screening method in many high-income settings. It is recommended by the WHO for women aged > 30 years in low-and-middle-income countries (LMICs) [2]. Its high sensitivity and negative predictive value in detecting cervical intraepithelial neoplasia grade 2 or worse (≥CIN2) allow extended screening intervals. Recently, the development of fully automated diagnostic devices providing rapid HPV testing of self-obtained vaginal samples has offered a great opportunity to improve the effectiveness of cervical cancer prevention in low-resource contexts [3].

However, a single HPV test has limited specificity and can lead to unnecessary workup and overtreatment. Therefore, a triage strategy is required for HPV-positive women to mitigate this difficulty. Cytology is generally proposed as it is an effective method for triaging HPV-positive women [4], but in low-resource settings, various logistic and operational reasons prevent successful cytology implementation. Amongst other barriers, cytological triage can be time-consuming, and in countries that use cytology as a triage method, results are typically unavailable on the same day as sample collection. In lower-income settings, loss to follow-up means that this becomes a seriously limiting problem. In these settings, therefore, rapid tests that give same-day results and lead to decisions about treatment are preferred.

A solution for countries with limited resources could be affordable digital imaging technology for real-time remote cytologic diagnosis by specialists [5]. Using this scheme, the preparation and digitization of cervical smears from HPV-positive women would be performed on-site during the same visit using a "test-triage-and-treat" approach (3T-approach) [6]. This process eliminates the need for in-house cytopathologists and might allow for reliable, cost-effective triage of

[*] Corresponding author.
*E-mail address:* thomas.stegmuller@epfl.ch (T. Stegmüller).

https://doi.org/10.1016/j.compbiomed.2023.107809
Received 19 June 2023; Received in revised form 24 November 2023; Accepted 1 December 2023
Available online 6 December 2023

HPV-positive women. Furthermore, to facilitate the visual analysis of Pap slides and reduce the screening time, deep learning-based algorithms could be used to obtain a rapid and accurate cytological diagnosis allowing a "same-day treatment".

The emergence of affordable and portable high-resolution scanners, such as the Grundium Ocus®40, along with low-cost slide preparation procedures like SurePath™, creates a favorable environment for this endeavor. When it comes to the learning algorithm, the main expenses are associated with the annotation process and the level of expertise it demands. Nevertheless, acquiring a large and well-curated annotated dataset proves challenging and time-consuming. Therefore, an important stepping stone towards the long-term objective of deep learning-assisted cytology diagnostics is to lower the barrier imposed by annotation requirements. Towards that goal, we investigate the application of self-supervised learning (SSL) methods to effectively utilize the abundance of unlabeled images freely available from whole slide images (WSIs) of Pap smear tests. More generally, we analyze and report some of the successes and shortcomings of deep learning for cytology images. Specifically, we unveil the following aspects of deep learning-based Pap smear cytology:

- We thoroughly evaluate the ability of models pre-trained with self-supervised learning to learn meaningful visual representations of cytology images for various downstream tasks. In particular, the resulting representations show superior discriminability and generalizability;
- Our experiments reveal that representations learned with publicly available single cervical cell datasets, *e.g.*, Herlev [7], or Sipakmed [8], do not generalize well to different modalities such as images representing multiple cells. To mitigate this issue, we propose a data augmentation strategy tailored for cytology images dubbed **C**ervical **C**ell **C**opy-**P**asting ($\mathtt{C}^3\mathtt{P}$). Furthermore, we demonstrate the effectiveness of $\mathtt{C}^3\mathtt{P}$ for learning generalizable representations from single-cell datasets;
- We experimentally observe that multiple instance learning (MIL), the commonly used strategy for obtaining WSI-level representations and predictions, does not fully exploit the inherent properties of Pap smear cytology slides. Consequently, we introduce a set of simple yet effective modifications, *e.g.*, processing only the top-k most suspicious instances, to better align MIL methods with Pap smear test images;
- We present a medium-sized liquid-based cytology Pap smear test images dataset from HPV-positive women. The slides of this dataset are prepared with the SurePath™ procedure, which results in a small cell-deposit area. This shortens the time of digitization and yields smaller WSIs files. This is ideal for our telecytologic-based same-day "test-triage-and-treat" long-term objective. The presented dataset is particularly challenging as all samples are from HPV-positive women, and negative slides typically portray signs of infections, which complicates the diagnosis.

## 2. Related work

Whole slide images have been adopted in surgical pathology, where there is evidence that the diagnostic performance of digital microscopy is equivalent to light microscopy. However, in cytopathology practice, the uptake of WSIs has been slower, as cytological preparations frequently display thick cell groups. This phenomenon, which necessitates multiple scanning planes to allow proper analysis of the slides, has been a technological barrier preventing digital cytology from widespread adoption. Today, whole slide microscope scanners integrate z-stacking, overcoming the technological constraints of the three-dimensional nature of cytological preparations [9]. The increase in file size resulting from scanning multiple focal planes can be counterbalanced by using liquid-based cytology preparations, which concentrate the cytological material in a limited region of the glass [9]. These technological advances in whole slide imaging and portable devices for primary diagnosis have been accompanied by increasing efforts to evaluate, validate, and regulate their usage [10–12], and it is recommended that future studies are conducted in accordance with the updated guidelines proposed by the US College of American Pathologists (CAP) [10].

The aforementioned technological advances and the availability of affordable and transportable digital microscopes in the market bring forth numerous opportunities for regulated innovation. The added possibility of incorporating computer-assisted and/or automated diagnostics makes this an even more exciting prospect. For most cancer-related diagnostics, histopathology-based assessment is considered the "gold standard". This explains why histology garners far more attention from the machine learning community, *e.g.*, Abbet et al. [13], Stegmüller et al. [14], Bozorgtabar et al. [15] than cytology. Recently though, cytopathology has gained more traction and recognition, as it offers a non-invasive and inexpensive diagnostic tool suited to resource-constrained countries. Consequently, there have been significant advancements in machine-learning approaches applied to cytology.

Most of these advancements focus on cell-level tasks, *e.g.*, classification [16,17], detection [18,19] or segmentation [20]. These innovations aim to improve the efficiency and accuracy of cervical cancer screening and other forms of cancer diagnosis from cytological samples. The classification of whole slide Pap smear test images remains significantly less studied, despite its practical application being the most promising. Notable works on the topic include [21], who combined low- and high-resolution stages for the identification/localization of suspicious lesions and their classification. The high-resolution stage relies on a recurrent neural network (RNN)-based classification model to predict the WSI-level scores. Another study, [22], leveraged a YOLO-based [23] approach to generate cell/tile-level predictions in the first stage and a transformer model for the aggregation and WSI-level classification in the second stage. Similarly, [24] proposed the integration of an attention module to detect abnormal cells in large patches and computed the abnormality probability of a given patch as the average of its constituent cells. The WSI-level score is obtained by averaging the abnormality of its patches. Recently, [25] proposed a three-stage pipeline for lung cancer cytopathological WSIs classification. Their approach integrates a transformer-based model to extract fine-grained lesion features, which are then aggregated into intermediate patch-level features, and coarse-grained features for final WSI-level classification.

## 3. Datasets

**In-house dataset.** We present our in-house dataset composed of a cohort of 307 Pap smear slides. The Pap tests were performed after the occurrence of a positive primary HPV test in Cameroon. The prevalence of cytology-positive slides is approximately 20%, translating to 69 positive and 238 negative slides. The preparation of the slides follows the SurePath™ procedure. This choice is aligned with our long-term objective: same-day "test-triage-and-treat" of HPV-positive women in a resource-limited setting. Indeed, this preparation yields a small cell-deposit area, which shortens the scanning time and reduces the size of the digitized slides. Additionally, the SurePath™ procedure exists in a manual and low-cost version to further ease its adoption in a low-income setting. Along the same line, the slides are digitized with the Grundium Ocus®40 scanner: a portable and affordable solution. The WSIs are acquired with a 12 megapixels image sensor, a 40× objective, and Z-stacking (3 focal planes spaced by 1μm).

After digitization, cell-level annotations are obtained using QuPath [26], resulting in a total of 1228 annotated positive cells. The annotations are used to create a dataset of 1228 positive cell images and as many positive 320 × 320 pixels tiles (cell+context, see Fig. 1) both at 20× magnification (0.50 μm/px). Alternatively, we create an unlabeled tile dataset by finding the cell-deposit area with standard image processing techniques and subsequently sampling 320 × 320 tiles

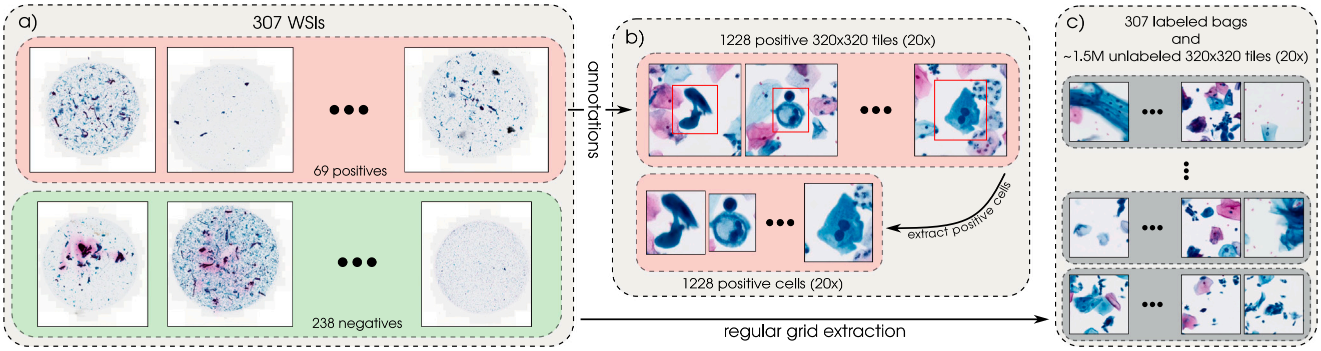

**Fig. 1. Overview of the in-house dataset.** (a) The $N$ WSIs are labeled as positives ($N_p$ samples) or negatives ($N_n$ samples). (b) 1228 cell-level annotations are used to extract isolated positive cells and 320 × 320 pixels tiles, both at 20× magnification. (c) The tiles of each WSI are extracted based on a regularly spaced grid, yielding ∼1.5M unlabeled tiles (320 × 320 pixels at 20× magnification) distributed over 307 slide bags.

on a regular grid without overlap, yielding approximately 1.5M images, subdivided into 307 bags encompassing an average of 5200 tiles each. The bags are used for WSI-level classification (see Section 4.3), whereas the individual tiles serve as a substrate for the SSL pre-training (see Section 4.1). We use a stratified 4-fold split approach to partition the slides into training, validation, and test subsets. The slide-level splits are common to all experiments, including the SSL pre-training.

**Herlev.** A liquid-based cytology Pap smear tests images dataset [7] encompassing 917 labeled cell images. It contains 242 cytology-negative images of annotated cell types of *superficial squamous epithelial* (NS), *intermediate squamous epithelial* (NI), and *Columnar epithelial* (NC). The 675 cytology-positives images are annotated as lesion cells of *mild squamous non-keratinizing dysplasia* (LD), *moderate squamous non-keratinizing dysplasia* (MD), *severe squamous non-keratinizing dysplasia* (SD), and *squamous cell carcinoma in situ intermediate* (CIS), respectively.

**Sipakmed.** A dataset of cervical squamous cells [8] from Pap smear images, which comprises 4049 labeled cell images. The 2411 cytology-negative images are further categorized in *metaplastic* (M), *superficial-intermediate* (SI), and *parabasal* (P). Similarly, images are annotated as *Koilocytotic* (K) and *dyskeratotic* (D) are represented among the 1638 cytology-positive images.

## 4. Method & experiments

Collecting extensive and meticulously curated annotated data is time-consuming and expensive. Consequently, we investigate self-supervised learning methods and approaches that require few labeled samples. More precisely, in Section 4.1, we provide empirical evidence that self-supervised learning approaches can be successfully leveraged to learn meaningful representations of multiple-cell images, *i.e.*, unlabeled tiles. Our proposed cell augmentation method $\mathtt{C^3P}$ is discussed and extensively tested in Section 4.2. Finally, in Section 4.3, we provide and discuss simple, yet effective tools tailored to existing MIL methods for cytology Pap smear test WSIs.

### *4.1. How well do self-supervised models transfer to cytology images?*

While self-supervised learning methods have gained attention for diverse downstream tasks in histopathology images, these approaches have received little attention for cytology-related tasks due to sparse evidence for their use on Pap smear cytology images. Most state-of-the-art SSL methods [27–30] for image-level representations learning rely on maximizing the similarity of an image's representation under information preserving transformations. Crucially, one of these transformations is a spatial crop, which is at risk of losing its information-preserving property on cytology images, partly due to the preparation of the slides that break long-range spatial dependencies. However, each digitized cytology slide can yield thousands of unlabeled images, which indicates that cytology could potentially be a playground where SSL methods thrive.

**Self-supervised pre-training.** We pre-train our models, using DINO [30] as the self-supervised learning framework. This choice is motivated by its strong nearest neighbor classifier capability and excellent performance across different backbones. DINO relies on a pair of Siamese teacher-student networks and falls in the category of self-distillation methods. The underpinning principle of the method is to train the student network to mimic the teacher's output distribution when both models are fed with distinct views of the same input image. DINO leverages both *global views* and *local views*. The former typically spans a larger image region and captures image-level dependencies, while the latter occupies a fraction of the image and yields localized features. By leveraging views at different scales, local-to-global consistency can be distilled from the teacher to the student network. Compared to contrastive learning approaches [27,31], self-distillation methods [30,32] must explicitly avoid the collapse of the learned representation to trivial solutions. In particular, DINO only updates the teacher network's weights with an exponential moving average (EMA) of those of the student network. Additionally, the entropy of the teacher's output distribution is constrained with sharpening and centering tricks.

We experiment with two types of architecture, ResNet-50 [33] and vision transformer (ViT), ViT-S/16 [34], for which we use the recommended hyperparameters available on the official repository. The arguments only differ from the recommendations for the batch size and the number of local crops. The batch size is set to fill the available GPU memory, *i.e.*, `batch_size` = 256 for a ResNet-50 and `batch_size` = 192 for a ViT-S/16. We do not use local crops as they can result in ambiguous positive pairs for non-object-centric datasets, as is the case here. For each architecture, we train one model per stratified split (see Section 3) for 300 epochs; hence we obtain four pre-trained models for each architecture.

In all the following experiments, we compare the quality of the learned visual representations under the above-described setting to the ones obtained under a supervised pre-training on ImageNet-1k. For the ResNet-50 architecture, we use the weights provided by PyTorch [35], whereas, for the ViT-S/16, we rely on the weights of [36] (trained without distillation).

**Cell-level classification.** After model pre-training, we probe the quality of the learned features on a cell-level classification task. We opt for a k-NN classifier to limit the manual intervention to the minimum, thereby obtaining results that reflect the learned representations' quality. We use two publicly available cervical cell Pap smear datasets: the Herlev dataset [7] and the Sipakmed dataset [8]. These datasets are randomly split in train/validation with a 75/25 partition. We report the mean and standard deviation of the class-wise and weighted $F_1$ scores over 4 independent runs for each pre-trained model, *i.e.*, a total of 16 for the models pre-trained under the self-supervised framework

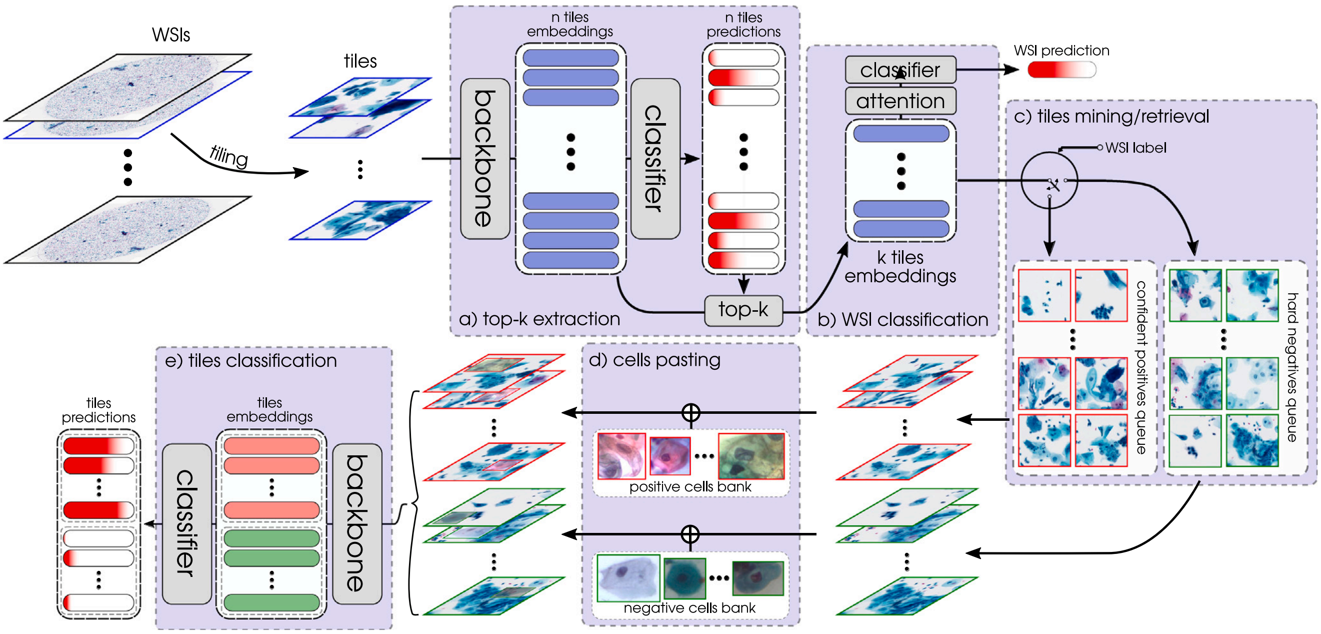

**Fig. 2. Overview of the proposed MIL-based method for classifying Pap smear WSIs.** (a) A positivity score is obtained **independently** for each tile of the input WSI, and the embeddings of the tiles having the top-k highest scores are extracted. (b) The top-k embeddings attend to one another to produce the slide-level representation, where the positivity score is obtained using the **same classifier** as for the independent tiles predictions. (c) The tiles corresponding to the top-k scores are stored as *confident positives* or *hard negatives* queues, depending on the slide-level label. (d) Positive and negative cells are pasted upon randomly sampled *confident positives* and *hard negatives*, respectively. (e) A score for each pasted tile is obtained using the *same backbone and classifier*. The model is conjointly trained to correctly classify WSIs and pasted tiles.

**Table 1**
Cell-level classification results on Herlev. We report the class-wise and weighted $F_1$ scores of a k-NN classifier. The features are extracted by a ViT-S/16 or a ResNet-50 pre-trained under a supervised pre-training on ImageNet or a self-supervised pre-training on our in-house unlabeled tiles dataset using DINO. The highest mean score for a given class and backbone are highlighted in **bold**.

| | *positives* | | | | *negatives* | | | *average* | | |
|---|---|---|---|---|---|---|---|---|---|---|
| *backbone* | CIS | LD | MD | SD | NC | NI | NS | positives | negatives | weighted $F_1$ |
| ResNet-50 | 55.0 ± 4.5 | 64.3 ± 5.3 | **48.6 ± 4.8** | 51.3 ± 3.7 | **53.8 ± 4.1** | 87.0 ± 7.5 | 89.6 ± 6.3 | **93.2 ± 0.6** | **80.1 ± 1.7** | 60.2 ± 3.2 |
| ResNet-50 | **59.0 ± 7.1** | **66.2 ± 3.7** | **48.6 ± 6.8** | **53.0 ± 4.9** | 50.9 ± 8.0 | **88.8 ± 5.2** | **92.4 ± 2.8** | 92.3 ± 1.5 | 78.6 ± 3.8 | **61.7 ± 3.2** |
| ViT-S/16 | 55.1 ± 5.7 | 59.5 ± 3.2 | 38.6 ± 8.2 | 48.7 ± 3.1 | 49.2 ± 4.6 | 75.8 ± 5.1 | 83.2 ± 7.2 | 92.4 ± 1.1 | 76.8 ± 2.4 | 55.2 ± 2.0 |
| ViT-S/16 | **62.8 ± 4.4** | **66.8 ± 4.5** | **48.2 ± 5.1** | **53.5 ± 5.5** | **58.8 ± 7.5** | **83.4 ± 1.9** | **89.7 ± 3.2** | **93.1 ± 0.9** | **80.8 ± 2.9** | **62.6 ± 2.8** |

**Table 2**
Cell-level classification results on Sipakmed. We report the class-wise and weighted $F_1$ scores of a k-NN classifier. The features are extracted by a ViT-S/16 or a ResNet-50 pre-trained under a supervised pre-training on ImageNet or a self-supervised pre-training on our in-house unlabeled tiles dataset using DINO. The highest mean score for a given class and backbone are highlighted in **bold**.

| | *positives* | | *negatives* | | | *average* | | |
|---|---|---|---|---|---|---|---|---|
| *backbone* | D | K | M | P | SI | positives | negatives | weighted $F_1$ |
| ResNet-50 | **94.5 ± 1.4** | 85.0 ± 0.8 | 89.2 ± 1.2 | 94.4 ± 1.4 | 97.2 ± 0.4 | 94.5 ± 0.4 | 96.4 ± 0.3 | 92.1 ± 0.6 |
| ResNet-50 | 93.7 ± 0.9 | **87.7 ± 1.7** | **91.2 ± 1.3** | **97.2 ± 0.9** | **98.6 ± 0.6** | **95.5 ± 0.6** | **96.9 ± 0.4** | **93.7 ± 0.7** |
| ViT-S/16 | 89.6 ± 1.8 | 82.6 ± 2.8 | 85.2 ± 2.3 | 94.2 ± 0.6 | 95.3 ± 0.7 | 93.4 ± 1.1 | 95.5 ± 0.8 | 89.3 ± 1.3 |
| ViT-S/16 | **94.8 ± 0.8** | **88.1 ± 1.3** | **91.1 ± 1.2** | **98.0 ± 0.5** | **98.5 ± 0.3** | **95.6 ± 0.6** | **97.0 ± 0.4** | **94.1 ± 0.4** |

(see Section 4.1) and 4 runs for the supervised ones. The number of neighbors $k$ is selected to maximize the weighted $F_1$ score.

In Tables 1 and 2, we observe that despite being pre-trained without any labels and not on isolated cells, the models resulting from DINO's pre-training are on-par or better than the ones pre-trained on ImageNet-1k, which are competitive baselines and the *de facto* choice for most practitioners. It further appears that the Herlev dataset is more challenging, especially with fine-grained class labels. However, the representations are good enough to differentiate negative cells from positive ones.

**Tile-level classification.** The representations learned via self-supervised learning are also evaluated on a tile-level classification task. As for the cell-level classification task, we rely on a k-NN approach. To that end, we prepare a labeled tiles dataset composed of our in-house 1228 positive tiles (see Section 3) and as many tiles randomly sampled from negative slides. The k-NN classifier is fitted on 75% of the resulting dataset and tested against the remaining 25%. We use the same evaluation setting as for the above-described cell-level classification task.

We observe in Table 3 that the self-supervised pre-training yields a significant boost in performance when the pre-training and target datasets are well aligned. Overall, it is remarkable that the self-supervised models pre-trained with DINO transfer well to cytology images, considering that DINO was originally tailored for object-centric datasets. Furthermore, the quality of the classification obtained with a k-NN classifier only seems to imply that the SSL models do not encode multiple cells as a single pattern, as it would not allow for the matching of positive tiles. We postulate that this is a consequence of the random cropping operation.

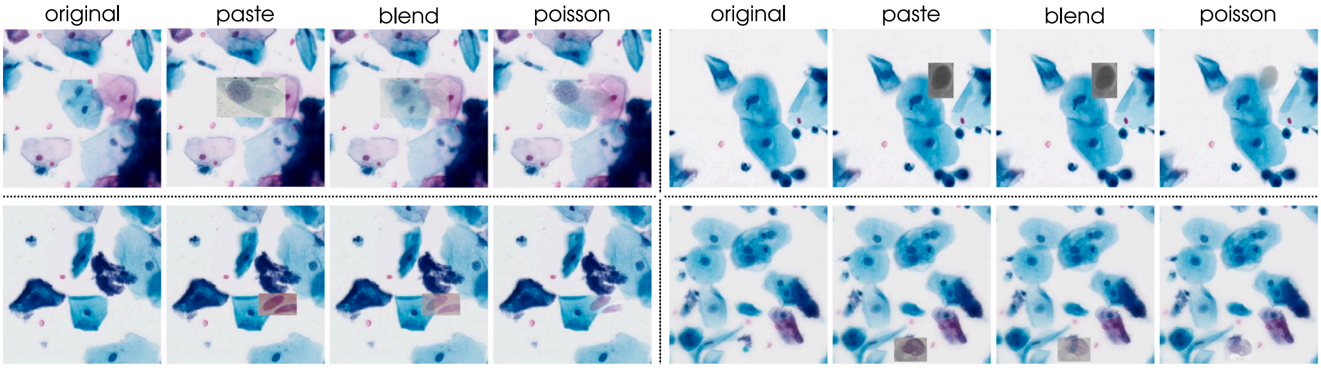

**Fig. 3.** Visualization of the different pasting approaches on randomly sampled tiles from the in-house dataset and random pasted cells from both Herlev and Sipakmed datasets.

**Table 3**
Tiles-level evaluation results of the frozen models on the in-house set of tiles. A k-NN classifier is fitted on 75% of the samples and evaluated on the remaining 25%. We report the class-wise and weighted $F_1$ scores of 4 independent runs. The features are extracted by a ViT-S/16 or a ResNet-50 pre-trained under a supervised pre-training on ImageNet or a self-supervised pre-training on our internal dataset using DINO. The highest mean score for a given class and backbone is highlighted in **bold**.

| *backbone* | SSL | positives | negatives | weighted $F_1$ |
|---|---|---|---|---|
| ResNet-50 | ✗ | 81.6 ± 0.3 | 79.1 ± 0.5 | 80.4 ± 0.2 |
| ResNet-50 | ✓ | **95.3 ± 0.6** (+13.7) | **95.1 ± 0.7** (+16.0) | **95.2 ± 0.6** (+14.8) |
| ViT-S/16 | ✗ | 77.8 ± 0.6 | 72.1 ± 1.8 | 74.9 ± 1.0 |
| ViT-S/16 | ✓ | **96.6 ± 0.4** (+18.8) | **96.4 ± 0.5** (+24.3) | **96.5 ± 0.4** (+21.6) |

**Table 4**
Transfer learning results from Herlev and Sipakmed to our in-house labeled tiles dataset. A k-NN classifier is fitted on the binary version of the Herlev (H) and Sipakmed (S) datasets and is evaluated on the in-house set of labeled tiles. The features are extracted by a ViT-S/16 or a ResNet-50 pre-trained under a supervised pre-training on ImageNet-1k or a self-supervised pre-training on our in-house unlabeled tiles dataset using DINO. The highest mean score for a given source dataset and backbone is highlighted in **bold**.

| | | *Herlev* → *tiles* | | *Sipakmed* → *tiles* | |
|---|---|---|---|---|---|
| *backbone* | SSL | negatives | positives | negatives | positives |
| ResNet-50 | ✗ | 33.3 ± 5.7 | 0.0 ± 0.0 | **68.9 ± 2.9** | 0.0 ± 0.0 |
| ResNet-50 | ✓ | **63.7 ± 1.6** | **33.8 ± 5.9** | 65.3 ± 1.2 | **18.4 ± 4.8** |
| ViT-S/16 | ✗ | 39.5 ± 3.9 | 0.0 ± 0.0 | **74.8 ± 2.8** | 0.0 ± 0.0 |
| ViT-S/16 | ✓ | **50.4 ± 1.4** | **41.3 ± 2.6** | 61.3 ± 1.8 | **30.2 ± 6.9** |

**Table 5**
Transfer learning results from Herlev and Sipakmed to our in-house labeled tiles dataset using C$^3$P-`paste`. A k-NN classifier is fitted on the pasted cells from the Herlev (H) and Sipakmed (S) datasets, and evaluated on the in-house set of positive cells/tiles. The features are extracted by a ViT-S/16 or a ResNet-50 pre-trained under a supervised pre-training on ImageNet-1k or a self-supervised pre-training on our in-house unlabeled tiles dataset using DINO. The highest mean score for a given source dataset and backbone is highlighted in **bold**.

| | | *Herlev* → *tiles* | | *Sipakmed* → *tiles* | |
|---|---|---|---|---|---|
| *backbone* | SSL | negatives | positives | negatives | positives |
| ResNet-50 | ✗ | 29.8 ± 2.5 | 65.0 ± 0.4 | 46.0 ± 1.5 | 65.3 ± 0.7 |
| ResNet-50 | ✓ | **41.0 ± 1.3** | **70.7 ± 0.5** | **63.4 ± 1.0** | **74.5 ± 0.8** |
| ViT-S/16 | ✗ | 27.1 ± 0.8 | 66.0 ± 0.6 | 52.9 ± 0.8 | 67.4 ± 1.3 |
| ViT-S/16 | ✓ | **54.8 ± 1.4** | **75.3 ± 0.4** | **77.9 ± 0.5** | **83.4 ± 0.3** |

### 4.2. *Cervical Cell Copy-Pasting:* C$^3$P

In Section 4.1, we discuss the applicability of self-supervised learning to cytology images and report evidence of its effectiveness. As much as self-supervised learning is an adequate approach that can yield semantically coherent clusters of image representations, it does not permit the labeling of the aforementioned clusters. Therefore, we investigate if this labeling operation can be performed using publicly available datasets. The major obstacle to achieving this objective is that most public datasets are at the cell level, whereas a lot of cytology tasks, *e.g.*, whole slide image classification, require patch/tile level representations and annotations. Consequently, we first show that naively using models trained on cell-level datasets does not transfer well to tile-level downstream tasks. We then propose a simple yet effective method based on samples mixing to overcome this issue.

Methods that involve sample mixing, *e.g.*, mixup [37] or CutMix [38], have been introduced to improve the generalization capabilities of neural networks. They achieve this by generating new image-label pairs through the combination of existing labeled samples. Here, we explore the mixing of unlabeled tiles and labeled single-cell images to produce labeled tiles.

**Cells to tiles transfer learning.** We first evaluate the capability of a classifier trained on open-source cell-level datasets for the tile-level classification at test time. To that end, a k-NN classifier is fitted on the Herlev or Sipakmed datasets using only binary labels, *i.e.*, negative or positive, and subsequently evaluated on our *in-house* set of labeled tiles. For each pre-trained model, we report the class-wise $F_1$ score averaged over 4 independent runs, which only use 75% of the training set each. When the model is pre-trained in a self-supervised manner, the scores are further averaged over the pre-training splits (see Section 4.1). The number of neighbors $k$ is selected to maximize the $F_1$ score of the positive class.

The results reported in Table 4 clearly show that a direct transfer learning from cells to tiles with a k-NN classifier performs poorly. More precisely, it can be observed that the models pre-trained in a supervised setting cannot detect the discriminant signal from the positive tiles. It is unclear whether this failure is a consequence of the shift in modality, *i.e.*, single-cell images to multi-cell images, or due to the small capacity of the classifier, the backbone, or another domain discrepancy between the source and target datasets. Overall, the self-supervised pre-trained models generalize better on this task.

**Cells to tiles transfer learning with pasting.** In order to understand why a k-NN classifier is unable to transfer learning from single-cell images to tiles. We repeat the same experiment with one crucial addition. As a pre-processing step, we use the proposed augmentation *i.e.*, we paste all the cells from Herlev or Sipakmed upon randomly sampled tiles from negative slides, referred to as canvases. The label of the pasted cell is attributed to the resulting pasted tile. In this first pasting scenario, we use the most straightforward pasting technique, which is referred to as the `paste` strategy.

`paste`: The strategy relies on a two-step procedure to paste a cell on a tile: (i) the pasting location of the cell is uniformly sampled among all the positions that would allow the cell to fit entirely in the tile, and (ii) the pixels of the tile in the pasting site are replaced by those of the cell. As such, this strategy is closest to CutMix [38].

As can be seen in Table 5, the proposed augmentation significantly improves the ability of the classifier to detect positive cells in tiles. This

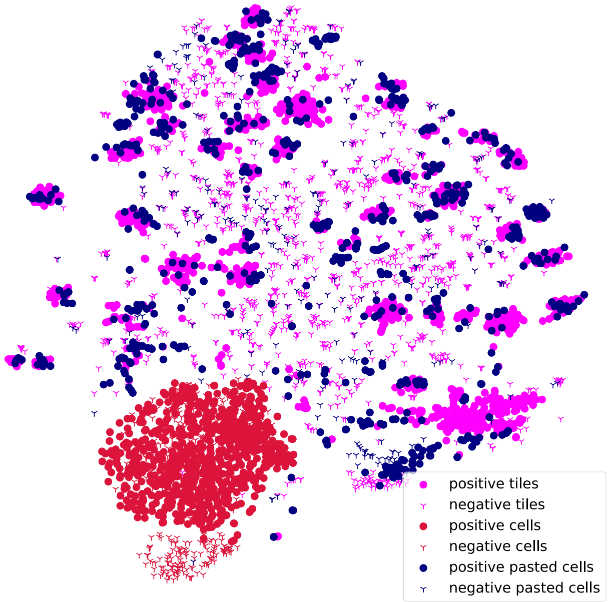

**Fig. 4. The t-SNE projection** obtained from a ViT-S/16 encoder of cells from Herlev, our *in-house* labeled tiles, and Herlev cells augmented with C$^3$P-Poisson.

**Table 6**
Ablation results on pasting method. A classifier is trained on cells from Herlev or Sipakmed with C$^3$P and various pasting techniques and subsequently evaluated on the in-house labeled tiles. We report the class-wise and weighted $F_1$ scores. The highest mean score for a given source dataset, class, and backbone is in **bold**. The selected pasting technique is highlighted.

| | | *Herlev* | | *Sipakmed* | |
|---|---|---|---|---|---|
| *pasting* | *backbone* | negatives | positives | negatives | positives |
| paste | ResNet-50 | 76.3 ± 3.3 | 76.8 ± 5.6 | **74.2 ± 6.0** | **77.2 ± 5.3** |
| blend | ResNet-50 | 72.9 ± 3.9 | 75.2 ± 6.9 | 67.0 ± 1.3 | 70.0 ± 6.4 |
| Poisson | ResNet-50 | **80.4 ± 2.2** | **77.8 ± 5.9** | 73.3 ± 4.6 | 71.3 ± 2.1 |
| paste | ViT-S/16 | **73.2 ± 7.5** | **77.1 ± 8.9** | 45.5 ± 3.3 | 70.5 ± 4.7 |
| blend | ViT-S/16 | 67.6 ± 3.1 | 67.6 ± 3.6 | 51.8 ± 6.2 | 69.7 ± 8.4 |
| Poisson | ViT-S/16 | 71.2 ± 9.4 | 76.7 ± 8.5 | **72.6 ± 8.3** | **77.0 ± 8.9** |

is not trivial considering the large distribution shift between the cells of Helev/Sipakmed and the ones represented in our *in-house* tiles, the small capacity of the classifier, and that tiles resulting from paste do not look natural. The t-SNE [39] mapping depicted in Fig. 4 shows that the cells and labeled tiles representation are mapped to different regions of the space.

*Pasting technique:* In Table 5, we showed that the proposed pasting method could significantly improve the transferability from public single-cell datasets to tiles representing multiple cells. Although the pasting method (paste) used to generate the results depicted in Table 5 works, it is coarse and does not produce natural-looking images. As such, it can result in the model focusing exclusively on the pasted regions throughout training, hence performing poorly at test time. Therefore, we investigate if this scenario occurs and if better alternatives exist. In addition to paste, we test two other alternatives referred to as blend and Poisson. Examples of samples obtained with the different pasting methods are depicted in Fig. 3.

blend: The only difference w.r.t. paste is that, instead of replacing the pixels of the canvas with those of the cell, the pixels of the pasting site result from a convex combination of those of the cell and canvas:

$$x_{\texttt{blend}} = (1 - \lambda_{\text{paste}}) \cdot x_{\text{cell}} + \lambda_{\text{paste}} \cdot x_{\text{canvas}} \tag{1}$$

where $\lambda_{\text{paste}}$ is sampled uniformly at random from the interval $[0, 1]$. Due to the transparency of the pasting operation, the resulting images look more natural as it mimics the effect of overlapping cells and the border of the cell image is less visible. This mixing technique is closest to mixup [37].

**Table 7**
Ablation experiments for pasting probability. A classifier is trained on cells from Herlev or Sipakmed with various probabilities of applying C$^3$P-Poisson on negative (-) and positive (+) tiles. The classifier is then evaluated on the in-house labeled tiles. We report the class-wise and weighted $F_1$ scores. The highest mean score for a given source dataset, class, and backbone is in **bold**. The selected pasting method is highlighted.

| | | *Herlev* → *tiles* | | *Sipakmed* → *tiles* | |
|---|---|---|---|---|---|
| *pasting [%]* (- \| +) | *backbone* | negatives | positives | negatives | positives |
| 0 \| 100 | ResNet-50 | 76.6 ± 3.3 | 71.5 ± 8.2 | 83.3 ± 5.5 | 81.3 ± 8.0 |
| 50 \| 100 | ResNet-50 | **81.6 ± 2.7** | 75.0 ± 5.2 | **84.2 ± 4.2** | **82.2 ± 3.9** |
| 100 \| 100 | ResNet-50 | 80.4 ± 2.2 | **77.8 ± 5.9** | 73.3 ± 4.6 | 71.3 ± 2.1 |
| 0 \| 100 | ViT-S/16 | 78.4 ± 4.8 | 71.7 ± 8.7 | 25.2 ± 16.0 | 65.2 ± 8.5 |
| 50 \| 100 | ViT-S/16 | **83.1 ± 2.3** | **81.4 ± 2.1** | **80.4 ± 4.7** | 75.7 ± 11.5 |
| 100 \| 100 | ViT-S/16 | 71.2 ± 9.4 | 76.7 ± 8.5 | 72.6 ± 8.3 | **77.0 ± 8.9** |

Poisson: The main pitfall of the blend strategy is that it can only conceal the boundaries of the pasting site by concealing the cell, which is undesirable. Poisson blending [40] was proposed to mitigate that issue. The blending operation is formulated as an optimization problem, which aims to compute the values of the pixels in the pasting site to preserve the gradients of the source/cell image while matching the pixel intensities of the target/canvas image at the boundaries.

We train a linear classifier on top of the pre-trained models with different pasting operations. In this experiment, 1000 unlabeled tiles are used as canvases for each class (negative/positive), and labeled tiles are obtained online by pasting a randomly selected labeled cell upon one of the canvases. Notably, positive cells are pasted upon unlabeled tiles from positive slides and reciprocally for negative cells. After training, the classifier is evaluated on the *in-house* labeled tiles. We report the class-wise $F_1$ score averaged over 4 independent runs per pre-trained weights. The scores are further averaged over the pre-training splits (see Section 4.1). For this experiment (and the ones that follow), we only use models pre-trained in a self-supervised manner as they have shown to be on par or better than their supervised counterparts.

Table 6 shows that the blend approach yields worsen results compared to paste. We postulate that this is a consequence of $\lambda_{\text{paste}}$ either being too low and the resulting images not looking more natural than the ones produced with paste, or it being too high and the pasted content being barely visible. Moreover, we observe that the Poisson technique performs similarly to paste for all backbone/dataset combinations, except for the ViT-S/16 + Sipakmed scenario, in which case it is the only pasting technique that yields decent results for the classification of negative tiles.

Fig. 4 reflects that positive cells augmented with C$^3$P-Poisson appear to be close to groups of positive tiles, demonstrating the improved alignment obtained with our augmentation strategy compared to the paste strategy.

*Pasting probability:* So far, we have applied the pasting operation in a perfectly symmetric manner, *i.e.*, it is systematically applied independently of the cell's label and that of the slide from which the canvas is extracted. Nonetheless, our setting is inherently asymmetric: on one side, we know with certainty that tiles extracted from negative slides are all negatives; on the other side, little can be said with regard to the label of tiles extracted from positive slides. Furthermore, by systematically using C$^3$P, we are encouraging the model to only consider the pasting site which is undesirable. We propose to exploit the asymmetry of the setting and not systematically use C$^3$P on negative tiles. This further allows the model to learn from real negative examples without the risk of feeding mislabeled samples to the model. Therefore, we replicate the experiment of Table 6, but this time, C$^3$P-Poisson is applied on the unlabeled tiles from negative slides with a given probability (see Table 7).

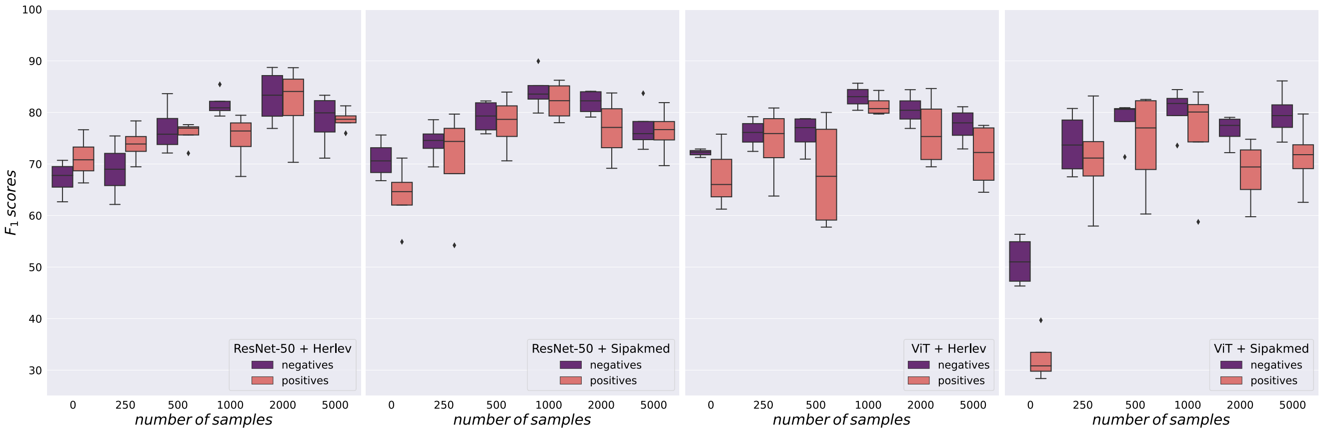

**Fig. 5. Box plots depicting the class-wise $F_1$ scores against the number of unlabeled tiles used as canvases for the pasting augmentation.** The performance achieved without the proposed augmentation can be observed at the zero of the x-axes.

**Table 8**
Evaluation results of the cells-pasting augmentation method with transfer learning from Herlev or Sipakmed to our in-house tiles dataset. A classifier is trained on the cells dataset without and with C$^3$P-Poisson. We report the class-wise and weighted $F_1$ scores. The highest mean score for a given backbone, class, and source dataset is highlighted in **bold**.

| | *Herlev → in-house tiles* | | *Sipakmed → in-house tiles* | |
|---|---|---|---|---|
| *backbone* | negatives | positives | negatives | positives |
| ResNet-50 | 67.2 ± 3.5 | 71.1 ± 4.4 | 70.9 ± 3.9 | 63.8 ± 6.7 |
| ResNet-5( | **83.1** ± **5.5** (+15.9 | **81.8** ± **8.1** (+10.7 | **84.2** ± **4.2** (+13.3 | **82.2** ± **3.2** (+18.4) |
| ViT-S/16 | 72.2 ± 0.8 | 67.7 ± 7.4 | 52.2 ± 5.0 | 32.4 ± 5.0 |
| ViT-S/16 | **83.1** ± **2.3** (+10.9 | **81.4** ± **2.1** (+13.7 | **80.4** ± **4.7** (+28.2 | **75.7** ± **11.5** (+43.3) |

Although never applying C$^3$P to negative tiles is the scenario in which the model processes the most realistic samples, we observe in Table 7 that it can be harmful. This observation is unsurprising, considering that in this situation, the positive label is perfectly correlated with the pasting operation. In fact, it is surprising that we do not see even worse. We argue that this is in part due to the ability of C$^3$P-Poisson to fool the model. The models trained using C$^3$P with a probability of 0.5 seem to perform favorably compared to the ones using it systematically. It is noteworthy that in that setting, the positive label is correlated with the action of pasting.

*How many canvases are required?:* To answer this question, we repeat the experiment of Table 6, with a 0.5 probability of applying C$^3$P-Poisson and a varying number of canvases per class.

Fig. 5 depicts the class-wise $F_1$ scores for each available backbone/dataset combination. It appears clear that, up until ≈ 2000 canvases, increasing the number of canvases favorably impacts the classifier's performance. After that point, the model tends to overfit the pasted cells, which occur more often and independently of the canvases, which in turn translates to a decreased downstream performance.

C$^3$P *results:* Our extended experiments reveal that C$^3$P offers a well-grounded augmentation strategy to bridge the gap between publicly available single-cell and unlabeled tiles datasets. We further show that the proposed augmentation yields significant improvement compared to the approach of naively transferring from a classifier trained on single-cell datasets. In Table 8, we also show that our approach outperforms the naive transferring methods by a large margin with a classifier trained with C$^3$P-Poisson, a pasting probability of 0.5, and the optimal number of canvases (see Fig. 5).

### 4.3. Aligning MIL to cytology images

In Sections 4.1 and 4.2, we showcase the benefits of self-supervised learning for cell-level and tile-level classification tasks on cytology images and proposed an augmentation strategy C$^3$P to make the most out of publicly available single-cell datasets. Combined together, this offers the opportunity to design Pap smear WSIs classification modules requiring few labels. More precisely, we harness the power of self-supervised learning and our augmentation strategy C$^3$P to better align well-established MIL methods for Pap smear WSIs classification.

**Problem formulation.** As a primer, we briefly revisit the underlying concepts and assumptions of the multiple instance learning framework. In a binary MIL setting, the objective is to correctly predict the label $Y \in \{0, 1\}$ of an input bag of instances $\boldsymbol{X} = \{\boldsymbol{x}_1, \dots, \boldsymbol{x}_n\}$, where $n$ is allowed to vary from one bag to the other. The instance-level labels $\{y_i\}_{i=1}^n \in \{0, 1\}$ are assumed to exist but to be unknown throughout the training phase. As such, the MIL objective can be formulated as the detection of positive instances ($y = 1$) within the bags, i.e.:

$$Y = \begin{cases} 1, & \text{iff } \sum_i y_i > 0, \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

As pointed out in AbMIL [41], the above bag labeling function is permutation invariant w.r.t. the instance labels, hence so must be the predictions $\hat{Y} = S(\boldsymbol{X})$, where $S$ is the bag scoring function. In the context of cytology, WSIs are the bags and their constituent tiles are the instances. One can observe that the permutation invariance assumption is particularly well-grounded in that setting. The overall diagnosis is based on the presence of abnormal cells within the entire slide rather than the specific arrangement or order of those cells. Furthermore, as a consequence of the slide preparation, the arrangement of the cells on the slides exhibits little to no ordering or positional dependency.

In most MIL methods, the slide-level representation $\mathbf{z}$ is obtained as a weighted sum/convex combination of the $n$ instance-level representations $H = \{\mathbf{h}_1, \dots, \mathbf{h}_n\}$:

$$\mathbf{z} = \sum_i \alpha_i \mathbf{h}_i \tag{3}$$

where $\alpha_i$ is a scalar that modulates the contribution of the $i$th instance to the overall representation. The slide's score is obtained by feeding the slide-level representation to a classifier $g$:

$$\hat{Y} = g(\mathbf{z}) \tag{4}$$

As the instance-level representations $\mathbf{h}_i$ and the slide-level representations $\mathbf{z}$ span the same space, we argue that instance-level predictions can be obtained with the same classifier:

$$\hat{y}_i = g\left(\mathbf{h}_i\right) \tag{5}$$

**MIL experimental setup.** We experiment with 3 different MIL methods, namely AbMIL [41], TransMIL [42], and CLAM [43]. We directly use their official implementations. We remove the positional encoding from TransMIL as it brings little information in our setting, as discussed

**Table 9**
Evaluation of the MIL-based methods before and after adding our augmentation C$^3$P and *top-k selection* strategy for Pap-smear test WSIs classification on our *in-house* dataset.

| | | | | *AuC scores* | |
|---|---|---|---|---|---|
| *top-k* \| C$^3$P | *method* | *backbone* | $\lambda_{\text{loc}}$ | slide-level | tile-level |
| ✗\| ✗ | AbMIL | ViT-S/16 | – | 59.0 ± 11.2 | 65.9 ± 11.6 |
| ✓\| ✗ | AbMIL | ViT-S/16 | – | **76.8** ± **3.3** | **86.9** ± **2.2** |
| ✓\| ✓ | AbMIL | ViT-S/16 | 0.1 | 76.5 ± 2.9 (+17.5) | 86.1 ± 2.2 (+20.2) |
| ✗\| ✗ | AbMIL | ResNet-50 | – | 61.1 ± 11.2 | 61.6 ± 12.5 |
| ✓\| ✗ | AbMIL | ResNet-50 | – | 70.9 ± 8.0 | 74.6 ± 20.8 |
| ✓\| ✓ | AbMIL | ResNet-50 | 1.0 | **72.8** ± **2.2** (+11.7) | **80.6** ± **4.3** (+19.0) |
| ✗\| ✗ | TransMIL | ViT-S/16 | – | 58.1 ± 5.4 | 53.3 ± 9.9 |
| ✓\| ✗ | TransMIL | ViT-S/16 | – | 59.8 ± 14.6 | 53.7 ± 12.4 |
| ✓\| ✓ | TransMIL | ViT-S/16 | 0.1 | **72.1** ± **8.4** (+14.0) | **67.7** ± **11.8** (+14.4) |
| ✗\| ✗ | TransMIL | ResNet-50 | – | 49.9 ± 5.2 | 46.7 ± 10.7 |
| ✓\| ✗ | TransMIL | ResNet-50 | – | 46.1 ± 8.7 | 50.1 ± 12.9 |
| ✓\| ✓ | TransMIL | ResNet-50 | 1.0 | **69.4** ± **14.5** (+19.5) | **71.6** ± **15.1** (+24.9) |
| ✗\| ✗ | CLAM | ViT-S/16 | – | 61.3 ± 6.2 | 69.4 ± 8.9 |
| ✓\| ✗ | CLAM | ViT-S/16 | – | 73.8 ± 4.4 | **84.1** ± **3.6** |
| ✓\| ✓ | CLAM | ViT-S/16 | 0.5 | **74.8** ± **3.0** (+13.5) | 77.0 ± 2.9 (+7.6) |
| ✗\| ✗ | CLAM | ResNet-50 | – | 64.3 ± 11.1 | 61.4 ± 11.3 |
| ✓\| ✗ | CLAM | ResNet-50 | – | 68.8 ± 14.4 | 68.4 ± 7.2 |
| ✓\| ✓ | CLAM | ResNet-50 | 1.0 | **77.5** ± **3.4** (+13.2) | **79.0** ± **4.0** (+17.6) |

above. Each MIL method is tested with both types of backbones (ViT-S/16 and ResNet-50), which are initialized with the weights obtained from DINO's pre-training [30]. In all experiments, the weights of the backbone are kept frozen. Considering the availability of 4 pre-training weights per backbone (see Section 4.1), the first one is used to determine the best hyperparameters, and the three remaining ones are reserved for evaluation purposes. For each setting, we report the average and standard deviation of the slide-level and instance-level AUC scores. The instance-level score is computed using our *in-house* positive tiles and randomly sampled tiles from negative slides (both extracted from the test tiles).

**Results discussion for MIL-based methods.** In the first scenario, we experiment with the MIL methods using their default implementations. It can be observed in Table 9 that this setting is suboptimal for all MIL method/backbone combinations. This is intriguing, considering that the backbone demonstrated strong performances (see Table 3) at the tile level and that the chosen MIL methods are well-established baselines. We argue that this is a consequence of the particularity of Pap smear test images. Of note, features correlated with negativity are present almost everywhere, even in positive tiles, and features correlated with positivity are scarce. Together, this makes for a particularly challenging setting to capture the positive signal in the slide-level representation, $\mathbf{z}$.

*Top-k selection:* To mitigate the aforementioned issue, we propose only processing the top-k most suspicious tiles in each slide using the same backbone and classifier as for the slide-level predictions. We see that as the backbone is frozen, the tiles' representations can be pre-computed, which makes the identification of the top-k tiles not compute-intensive. In all following experiments, we use $k = 8$ top-k tiles and a `batch_size` = 16.

Table 9 shows that adding a top-k module yields significant improvements for all MIL methods except for TransMIL. When using the top-k is beneficial for the slide-level predictions, we observe that it also benefits the tiles-level predictions, which is unsurprising considering that the slide-level representation $\mathbf{z}$ is most likely closer to that of tiles when it results from a weighted-sum over 8 tiles representations than over the entire bag (see Eq. (3)). Similarly, the poor tile-level performance of TransMIL is a potential explanation for its ineffectiveness at the slide level. Indeed, if the model cannot detect the positive tiles, the overall representation does not reflect the nature of the slide well.

*Tile-level objective:* To improve the ability of the model to identify suspicious tiles, we propose to integrate a tile-level loss into the overall training objective:

$$\mathcal{L} = \mathcal{L}_{\text{slide}} + \lambda_{\text{tile}} \cdot \mathcal{L}_{\text{tile}} \tag{6}$$

where $\lambda_{\text{tile}}$ denotes the tile loss coefficient. We use a `batch_size` of 8 for the tiles and the optimal value of $\lambda_{\text{tile}}$ is determined independently for each backbone/method combination. Moreover, since we aim for a method using only slide-level labels, we explore the possibility of benefiting from C$^3$P. As depicted in Fig. 2, *hard negative* and *confident positive* tiles are collected throughout training and used as canvases where negative and positive cells can be pasted upon, respectively. We refer to *hard negatives/confident positives* as the 10 tiles having the highest positivity score in each negative/positive slide, respectively. The method used for pasting is C$^3$P-`Poisson`, and we rely on cells from both Herlev and Sipakmed.

As shown in Table 9, incorporating a localized objective alongside C$^3$P yields significant improvements in the tile-level predictions of TransMIL. Consequently, this facilitates the detection of suspicious tiles and ultimately enhances the accuracy of slide-level predictions. It is worth noting that C$^3$P is not exclusively advantageous for TransMIL, as it proves to be beneficial at the slide level in all scenarios except for ViT-S/16 + AbMIL. We posit that a tile-level loss is implicitly enforced when employing a top-k selection approach. In other words, the representation of the top-k selected tiles is encouraged to be aligned with the slide label. Hence, when the top-k selection is accurate, C$^3$P becomes less relevant as "real" labeled tiles are available. Nonetheless, this scenario seldom occurs, especially when the backbone is a ResNet-50, whose features have 2048 dimensions (> 5× more than for a ViT-S/16), which increases the number of parameters of the MIL module. This may explains why settings relying on a ResNet-50 as backbone tend to benefit more from C$^3$P.

## 5. Conclusion

This paper intervenes at a particularly opportune moment in the realm of telecytology innovation. The introduction of affordable slide scanners, such as the Grundium Ocus®40, coupled with cost-effective slide preparation methods like SurePath™, presents unique opportunities for advancing remote cytology diagnostics. To contribute to this ambitious endeavor, we present a medium-sized dataset of Pap test WSIs from HPV-positive women, collected in a resource-constrained setting in Cameroon.

Additionally, our experimental findings highlight the successful application of self-supervised learning to reduce the annotation burden, with the resulting representations outperforming *off-the-shelf* pre-trained models across various downstream tasks. Additionally, we have introduced C$^3$P, an augmentation strategy, which effectively transfers knowledge from public single-cell datasets to unlabeled tiles. C$^3$P proves to be beneficial not only for tile-level classification but also for slide-level classification. Regarding the WSIs classification, our experiments reveal that MIL methods may overlook crucial characteristics of Pap smear images. These limitations can be accounted for by introducing simple modifications that prove to be beneficial. Overall, classifying Pap smear WSIs relying solely on slide-level labels remains challenging, particularly in our scenario where all samples are from HPV-positive women, which adds an additional layer of complexity.

**Limitations.** Our experiments are conducted on only one self-supervised learning method, namely DINO [30] due to its strong performance on the k-NN evaluation benchmark and compatibility with various backbones. We argue that the main reason why SSL methods could be inadequate for cytology images is that the objective may enforce consistency between semantically unrelated views. Nevertheless, this potential pitfall results from the spatial cropping strategy, which is common to most self-distillation and contrastive methods. Therefore, our conclusions, based on DINO, are likely also applicable to other methods. Alternatively, larger vision transformer backbones, *e.g.*, ViT-B/16, would be worth investigating, yet lighter architectures, such as ViT-S/16 and ResNet-50, remain better suited in the low-data regime.

## CRediT authorship contribution statement

**Thomas Stegmüller:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Visualization, Writing – original draft, Writing – review & editing, Software. **Christian Abbet:** Formal analysis, Investigation, Software. **Behzad Bozorgtabar:** Conceptualization, Funding acquisition, Supervision, Validation, Writing – original draft, Writing – review & editing. **Holly Clarke:** Data curation, Writing – original draft, Writing – review & editing. **Patrick Petignat:** Funding acquisition, Resources, Supervision. **Pierre Vassilakos:** Data curation, Funding acquisition, Investigation, Methodology, Supervision, Writing – original draft, Writing – review & editing. **Jean-Philippe Thiran:** Funding acquisition, Resources, Supervision.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Declaration of generative ai and ai-assisted technologies in the writing process

During the preparation of this work, the author(s) used ChatGPT in order to improve readability. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

## Acknowledgment

This work is supported by the Personalized Health and Related Technologies (PHRT), Switzerland, grant number 2021/344.

## References

[1] H. Sung, J. Ferlay, R.L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, F. Bray, Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries, CA: Cancer J. Clinicians 71 (3) (2021) 209–249.

[2] W.H. Organization, et al., WHO Guideline for Screening and Treatment of Cervical Pre-Cancer Lesions for Cervical Cancer Prevention, World Health Organization, 2021.

[3] R. Saidu, L. Kuhn, A. Tergas, R. Boa, J. Moodley, C. Svanholm-Barrie, D. Persing, S. Campbell, W.-Y. Tsai, T.C. Wright, et al., Performance of Xpert HPV on self-collected vaginal samples for cervical cancer screening among women in South Africa, J. Lower Genital Tract Disease 25 (1) (2021) 15.

[4] L. von Karsa, M. Arbyn, H. De Vuyst, J. Dillner, L. Dillner, S. Franceschi, J. Patnick, G. Ronco, N. Segnan, E. Suonio, et al., European guidelines for quality assurance in cervical cancer screening. summary of the supplements on HPV screening and vaccination, Papillomavirus Res. 1 (2015) 22–31.

[5] P. Vassilakos, H. Clarke, M. Murtas, T. Stegmüller, A. Wisniak, F. Akhoundova, Z. Sando, G.E. Orock, J. Sormani, J.-P. Thiran, et al., Telecytologic diagnosis of cervical smears for triage of self-sampled human papillomavirus–positive women in a resource-limited setting: concept development before implementation, J. Am. Soc. Cytopathol. (2023).

[6] J. Levy, M. De Preux, B. Kenfack, J. Sormani, R. Catarino, E.F. Tincho, C. Frund, J.T. Fouogue, P. Vassilakos, P. Petignat, Implementing the 3T-approach for cervical cancer screening in Cameroon: Preliminary results on program performance, Cancer Med. 9 (19) (2020) 7293–7300.

[7] J. Jantzen, J. Norup, G. Dounias, B. Bjerregaard, Pap-smear benchmark data for pattern classification, in: Nature inspired Smart Information Systems, NiSIS 2005, 2005, pp. 1–9.

[8] M.E. Plissiti, P. Dimitrakopoulos, G. Sfikas, C. Nikou, O. Krikoni, A. Charchanti, Sipakmed: A new dataset for feature and image based classification of normal and pathological cervical cells in pap smear images, in: 2018 25th IEEE International Conference on Image Processing, ICIP, 2018, pp. 3144–3148, http://dx.doi.org/10.1109/ICIP.2018.8451588.

[9] A. Eccher, I. Girolami, Current state of whole slide imaging use in cytopathology: Pros and pitfalls, Cytopathology 31 (5) (2020) 372–378.

[10] L. Pantanowitz, J.H. Sinard, W.H. Henricks, L.A. Fatheree, A.B. Carter, L. Contis, B.A. Beckwith, A.J. Evans, A. Lal, A.V. Parwani, Validating whole slide imaging for diagnostic purposes in pathology: guideline from the college of American pathologists pathology and laboratory quality center, Arch. Pathol. Lab. Med. 137 (12) (2013) 1710–1722.

[11] I. Kholová, G. Negri, M. Nasioutziki, L. Ventura, A. Capitanio, M. Bongiovanni, P.A. Cross, C. Bourgain, H. Edvardsson, R. Granados, et al., Inter-and intraobserver agreement in whole-slide digital ThinPrep samples of low-grade squamous lesions of the cervix uteri with known high-risk HPV status: A multicentric international study, Cancer Cytopathol. 130 (12) (2022) 939–948.

[12] N. Santonicco, S. Marletta, L. Pantanowitz, G. Fadda, G. Troncone, M. Brunelli, C. Ghimenton, P. Antonini, G. Paolino, I. Girolami, et al., Impact of mobile devices on cancer diagnosis in cytology, Diagn. Cytopathol. 50 (1) (2022) 34–45.

[13] C. Abbet, L. Studer, A. Fischer, H. Dawson, I. Zlobec, B. Bozorgtabar, J.-P. Thiran, Self-rule to multi-adapt: Generalized multi-source feature learning using unsupervised domain adaptation for colorectal cancer tissue detection, Med. Image Anal. (ISSN: 1361-8415) 79 (2022) 102473, http://dx.doi.org/10.1016/j.media.2022.102473, URL: https://www.sciencedirect.com/science/article/pii/S1361841522001207.

[14] T. Stegmüller, B. Bozorgtabar, A. Spahr, J.-P. Thiran, Scorenet: Learning non-uniform attention and augmentation for transformer-based histopathological image classification, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2023, pp. 6170–6179.

[15] B. Bozorgtabar, G. Vray, D. Mahapatra, J.-P. Thiran, SOoD: Self-supervised out-of-distribution detection under domain shift for multi-class colorectal cancer tissue types, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 3324–3333.

[16] H. Lin, Y. Hu, S. Chen, J. Yao, L. Zhang, Fine-grained classification of cervical cells using morphological and appearance based convolutional neural networks, IEEE Access 7 (2019) 71541–71549.

[17] T. Albuquerque, R. Cruz, J.S. Cardoso, Ordinal losses for classification of cervical cancer risk, PeerJ Comput. Sci. 7 (2021) e457.

[18] Y. Liang, C. Pan, W. Sun, Q. Liu, Y. Du, Global context-aware cervical cell detection with soft scale anchor matching, Comput. Methods Programs Biomed. 204 (2021) 106061.

[19] X. Li, Q. Li, et al., Detection and classification of cervical exfoliated cells based on faster R-CNN, in: 2019 IEEE 11th International Conference on Advanced Infocomm Technology, ICAIT, IEEE, 2019, pp. 52–57.

[20] E. Hussain, L.B. Mahanta, C.R. Das, M. Choudhury, M. Chowdhury, A shape context fully convolutional neural network for segmentation and classification of cervical nuclei in pap smear images, Artif. Intell. Med. 107 (2020) 101897.

[21] S. Cheng, S. Liu, J. Yu, G. Rao, Y. Xiao, W. Han, W. Zhu, X. Lv, N. Li, J. Cai, et al., Robust whole slide image analysis for cervical cancer screening using deep learning, Nature Commun. 12 (1) (2021) 1–10.

[22] Z. Wei, S. Cheng, X. Liu, S. Zeng, An efficient cervical whole slide image analysis framework based on multi-scale semantic and spatial deep features, 2021, arXiv preprint arXiv:2106.15113.

[23] J. Redmon, A. Farhadi, Yolov3: An incremental improvement, 2018, arXiv preprint arXiv:1804.02767.

[24] L. Cao, J. Yang, Z. Rong, L. Li, B. Xia, C. You, G. Lou, L. Jiang, C. Du, H. Meng, et al., A novel attention-guided convolutional network for the detection of abnormal cervical cells in cervical cancer screening, Med. Image Anal. 73 (2021) 102197.

[25] G. Li, Q. Liu, H. Liu, Y. Liang, A novel transformer-based pipeline for lung cytopathological whole slide image classification, in: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, IEEE, 2023, pp. 1–5.

[26] P. Bankhead, M.B. Loughrey, J.A. Fernández, Y. Dombrowski, D.G. McArt, P.D. Dunne, S. McQuaid, R.T. Gray, L.J. Murray, H.G. Coleman, et al., QuPath: Open source software for digital pathology image analysis, Sci. Rep. 7 (1) (2017) 1–7.

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

[28] J. Zhou, C. Wei, H. Wang, W. Shen, C. Xie, A. Yuille, T. Kong, iBOT: Image BERT pre-training with online tokenizer, 2021, arXiv preprint arXiv:2111.07832.

[29] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, A. Joulin, Unsupervised learning of visual features by contrasting cluster assignments, Adv. Neural Inf. Process. Syst. 33 (2020) 9912–9924.

[30] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, A. Joulin, Emerging properties in self-supervised vision transformers, 2021, arXiv preprint arXiv:2104.14294.

[31] T. Chen, S. Kornblith, M. Norouzi, G. Hinton, A simple framework for contrastive learning of visual representations, in: International Conference on Machine Learning, PMLR, 2020, pp. 1597–1607.

[32] J.B. Grill, F. Strub, F. Altché, C. Tallec, P.H. Richemond, E. Buchatskaya, C. Doersch, B.A. Pires, Z.D. Guo, M.G. Azar, et al., Bootstrap your own latent: A new approach to self-supervised learning, 2020, arXiv preprint arXiv:2006.07733.

[33] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.

[34] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16x16 words: Transformers for image recognition at scale, 2020, arXiv preprint arXiv:2010.11929.

[35] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al., Pytorch: An imperative style, high-performance deep learning library, Adv. Neural Inf. Process. Syst. 32 (2019) 8026–8037.

[36] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, H. Jegou, Training data-efficient image transformers & distillation through attention, in: M. Meila, T. Zhang (Eds.), Proceedings of the 38th International Conference on Machine Learning, in: Proceedings of Machine Learning Research, vol. 139, PMLR, 2021, pp. 10347–10357, URL: https://proceedings.mlr.press/v139/touvron21a.html.

[37] H. Zhang, M. Cisse, Y.N. Dauphin, D. Lopez-Paz, mixup: Beyond empirical risk minimization, Int. Conf. Learn. Representations (2018) URL: https://openreview.net/forum?id=r1Ddp1-Rb.

[38] S. Yun, D. Han, S.J. Oh, S. Chun, J. Choe, Y. Yoo, Cutmix: Regularization strategy to train strong classifiers with localizable features, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 6023–6032.

[39] L. Van der Maaten, G. Hinton, Visualizing data using t-SNE, J. Mach. Learn. Res. 9 (11) (2008).

[40] P. Pérez, M. Gangnet, A. Blake, Poisson image editing, in: ACM SIGGRAPH 2003 Papers, 2003, pp. 313–318.

[41] M. Ilse, J. Tomczak, M. Welling, Attention-based deep multiple instance learning, in: International Conference on Machine Learning, PMLR, 2018, pp. 2127–2136.

[42] Z. Shao, H. Bian, Y. Chen, Y. Wang, J. Zhang, X. Ji, Y. Zhang, TransMIL: Transformer based correlated multiple instance learning for whole slide image classication, 2021, arXiv preprint arXiv:2106.00908.

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