Automatic grading of cervical biopsies by combining full and self-supervision

Mélanie Lubrano di Scandalea[1,2], Tristan Lazard[2,3,4], Guillaume Balezo[1], Yaëlle Bellahsen-Harrar[5], Cécile Badoual[5], Sylvain Berlemont[1], and Thomas Walter[2,3,4]

[1] *KEEN EYE, 74 Rue du Faubourg Saint Antoine, Paris, 75012, France*
[2] *Centre for Computational Biology (CBIO), MINES ParisTech, PSL University, 60 Boulevard Saint Michel, 75272 Paris Cedex 06, France*
[3] *Institut Curie, 75248 Paris Cedex, France*
[4] *INSERM, U900, 75248 Paris Cedex, France*
[5] *Department of Pathology, Hôpital Européen Georges-Pompidou, APHP, Paris, France*

January 14, 2022

Abstract

In computational pathology, the application of Deep Learning to the analysis of Whole Slide Images (WSI) has provided results of unprecedented quality. Due to their enormous size, WSIs have to be split into small images (tiles) which are first encoded and whose representations are then agglomerated in order to solve prediction tasks, such as prognosis or treatment response. The choice of the encoding strategy plays a key role in such algorithms. Current approaches include the use of encodings trained on unrelated data sources, full supervision or self-supervision. In particular, self-supervised learning (SSL) offers a great opportunity to exploit all the unlabelled data available. However, it often requires large computational resources and can be challenging to train. On the other end of the spectrum, fully-supervised methods make use of valuable prior knowledge about the data but involve a costly amount of expert time.

This paper proposes a framework to reconcile SSL and full supervision and measures the trade-off between long SSL training and annotation effort, showing that a combination of both has the potential to substantially increase performance. On a recently organized challenge on grading Cervical Biopsies, we show that our mixed supervision scheme reaches high performance (weighted accuracy (WA): 0.945), outperforming both SSL (WA: 0.927) and transfer learning from ImageNet (WA: 0.877). We further provide insights and guidelines to train a clinically impactful classifier with a limited expert and/or computational workload budget. We expect that the combination of full and self-supervision is an interesting strategy for many tasks in computational pathology and will be widely adopted by the field.

1 Introduction

Recent advances in slide digitization have led to increased interest in Artificial Intelligence (AI) applications for histopathology. The development of AI models could help reduce pathologists' workloads, limit subjectivity and help contributing to medical discoveries. Deep learning models can now match pathologist performance for many tasks: diagnostic, detection of mitoses [Veta et al., 2015], prediction of gene mutations [Coudray et al., 2018, Kather et al., 2020] or genetic signatures [Kather et al., 2020, Diao et al., 2021, Lazard et al., 2021], cancer subtyping [Coudray et al., 2018] and more.

One of the applications, automated diagnosis from Whole Slide Images (WSIs), induces two main challenges: first, WSIs are very high-resolution and, because of memory constraint, cannot be fed directly into traditional neural networks. Second, expert annotations are laborious to attain, costly and prone to subjectivity. The most popular methods today rely on Multiple Instance Learning (MIL), which frames the problem as a bag classification task. WSIs are split into small workable images (tiles), which are processed separately. Features from each of the individual tiles are extracted and then aggregated to classify the WSI.

The extraction of these tiles' specific representation is crucial to the downstream WSI classification task. One common approach consists of initializing the feature extractor with pre-trained weights on ImageNet, a natural image dataset. This technique allows one to extract generic features that are powerful, but that do not lie within the histopathological domain. Different strategies have been developed to extract these tile encodings taking advantage of the available data and their respective level of supervision.

A first strategy aims to learn tile features with full supervision [Ehteshami Bejnordi et al., 2017]. To create a supervised dataset, one or several experts manually review tiles and sort them into meaningful classes (preferably related to the downstream task of classifying the WSIs). Even though experts' annotations can bring powerful prior knowledge to the model, this technique often requires large quantities of annotations.

A second strategy consists of learning tile representations through self-supervision. It leverages the unannotated data by training a convolutional neural network on a pretext task. It has proven its efficacy [Saillard et al., 2021, Lu et al., 2021] and even its superiority to the fully supervised scheme [Dehaene et al., 2020]. However, this approach has a non-negligible computational cost, as training necessitates around 1000 hours of computation on a standard GPU [Dehaene et al., 2020]. Moreover, it is not guaranteed that the obtained encodings are most relevant for the prediction task we are trying to solve.

Techniques from both sides of the supervision spectrum have proven to bring important benefits for relevant feature extraction. Combining them could allow us to benefit from the best of both worlds. In this work, in addition to proposing a joint-optimization process mixing self, full and weak supervision (Figure 1), we measure the trade-off in performance between the number of annotations and the computational cost of training a self-supervised model. We thus provide guidelines to train a clinically impactful classifier with a limited budget in expert and/or computational workload.

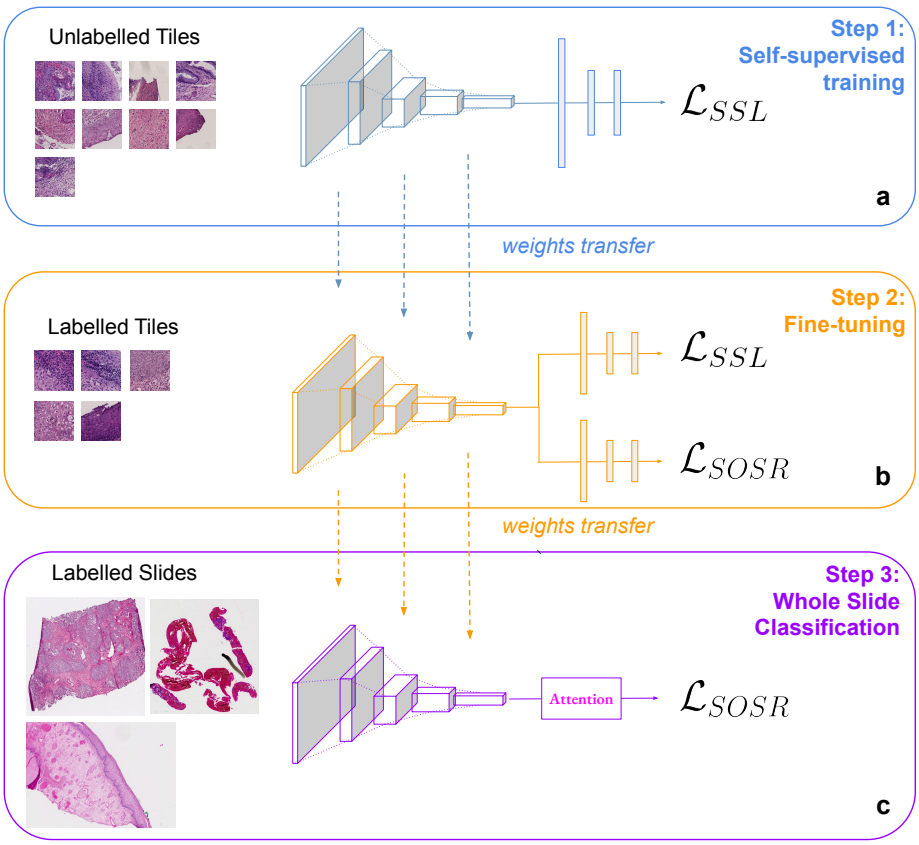

Figure 1: **Mixed Supervision Process**: **a)**A self-supervised model (SimCLR) is trained on unlabelled tiles extracted from the slides. Feature extractor and contrastive layer weights are transferred to the joint-optimization architecture **b)** Joint-optimization model is trained on the labeled tiles of the dataset. The feature extractor weights are transferred to the WS classification model. **c)** WS classification model is trained on the 1015 whole slide images.

2 Related Work

Mixed Supervision Medical data is often limited. For this reason, one might want to take advantage of all the available data even if annotations might not be homogeneous and even though they might be difficult to exploit because multiple levels of supervision are available. For instance, whole slide images are often associated with one global label (weak supervision), they can contain millions of unlabelled tiles (no supervision), but, as a pathologist reviews the slides and performs a diagnostic, it is almost effortless for them to mark the region of interest that signs the corresponding

diagnostic (strong supervision). AI applications have usually been dichotomized between supervised and unsupervised methods, spoiling the potential of combining several types of annotations. For this reason, mixing supervision for medical images analysis has gained interest in past years [Huang et al., 2020, Li et al., 2018, 2021].
For instance, in [Mlynarski et al., 2019] the author showed that combining global labels and local annotations by training in a multi-task setting, the capacities of the model to segment brain tumors on Magnetic Resonance Images were improved.
In [Tourniaire et al., 2021], the author introduced a mixed supervision framework for metastasis detection building on the CLAM [Lu et al., 2021] architecture. CLAM is a variant of the popular attention based MIL [Ilse et al., 2018] with 2 extensions: first, in order to make the method applicable in a multi-class setting, class-specific attention scores are learned and applied. Second, the last layer of the tile encoding network is trained to also predict the top and bottom attention scores, thus mimicking tile-level annotations. In [Tourniaire et al., 2021], the authors highlight the limitations of this instance-classification approach and propose to leverage a low number of fully annotated slides to train the attention mechanism. In a second step, they propose to turn to a standard MIL training (using only slide-level annotations). Even with few annotated slides, this approach allows to boost classification performance. However, there are also some limitations. First, the method relies on exhaustive annotation of selected slides: for the annotated slides, all the key regions are annotated pixel-wise. Second, due to the CLAM architecture, the approach only fine-tunes a single dense layer downstream the pre-trained feature extractor. Third, the algorithm has been designed for an application case in which the slide and tile labels coincide (tumour presence). This however is not always the case: when predicting genetic signatures, grades or treatment responses, it is unclear how tile and slide level annotations relate to each other. In this article, we propose to overcome these limitations. We propose to combine self-supervised learning with supervision prior to training the MIL network. We thus start from more powerful encodings, that are not only capable of solving the pretext task of self-supervised learning, but also the medical classification task that comes with the annotated tiles. Consequently, this method does not require full-slide annotations, optimizes the full tile encoding network and does not come with any constraint regarding the relationship between tile and slide level annotations.

3 Materials and Method

3.1 Dataset and Problem Setting

The Tissue Net Challenge [DrivenData] organized in 2020, the *Société Française de Pathologie (SFP)* and the *Health Data Hub* aimed at developing methods to automatically grade lesions of the uterine cervix in four classes according to their severity. The training dataset for the challenge was made up of biopsy samples from female uterine cervix, focusing on squamous lesions (Figure 2). These lesions are often benign but can also be qualified as low grade or high grade depending on the

risk of invasion of the underlying conjunctive tissue and evolution into carcinomas. The grade of the lesions depends on the proportion of squamous epithelium affected by dysplastic criteria. Low-grade squamous intraepithelial lesions (LSIL) are defined as having a dysplastic criteria involving less than one third of the thickness of the epithelium. High-grade squamous intraepithelial lesions (HSIL) indicate a greater proportion of the epithelium composed of undifferentiated basal cells with abnormalities. Carcinoma is diagnosed when abnormal epithelial cells invade the underlying conjunctive tissue. The class of a WSI was determined by the highest lesion's grade present on it.

3.2 Fully Supervised Dataset

5926 annotated Regions of Interest (ROIs) of fixed size 300x300 micrometers were provided. Each ROI had roughly the same size as a tile at 10x magnification and were labeled by the severity of the lesion it contained: "Normal" (0) if tissue was normal, (1) LSIL or (2) HSIL if it presented precancerous lesions that could have malignant potential and (3) invasive squamous carcinoma (Table 1).

Classes	Number of Slides	Number of Tiles
0 (Normal)	270	1923
1 (Low Grade)	288	1405
2 (High Grade)	238	1368
3 (Carcinoma)	219	1230
Total	1015	5926

Table 1: Dataset Summary

3.3 Weakly Supervised Dataset

The dataset was composed of 1015 WSIs acquired from 20 different centers in France at an average resolution of 0.234 +/- 0.0086 mpp (40X). The slide resolution varied slightly due to the multicentric

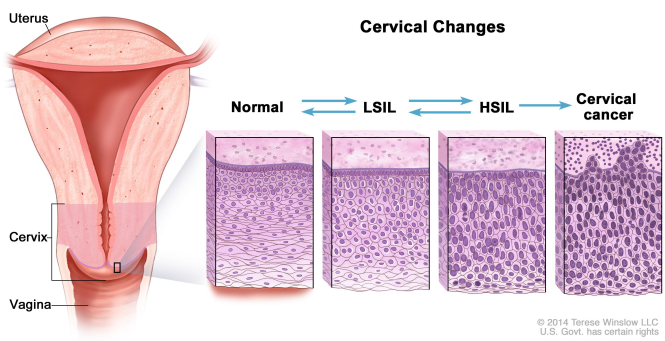

Figure 2: **Illustration of Uterine cervix dysplasia** - [National Cancer Institute, 2011]

provenance of the data. The class of the WSI corresponded to the class of the most severe lesions it contained (grade from 0 to 3 also). All the native WSI formats were converted to pyramidal TIFF (Tagged Image File Format). Both the WSI-level and tile-level labels have been attributed by a consortium of expert pathologists (Table 1).

3.4 Misclassification Costs

Misclassification errors do not lead to equally serious consequences (i.e predicting a benign lesion if it is cancerous is more serious than predicting a LSIL instead of a HSIL). Accordingly, a panel of pathologists established a grading of each of these errors i.e they attributed to each pair of possible outcome $(i, j) \in \{0, 1, 2, 3\}^2$ a severity score $0 \leqslant C_{i,j} \leqslant 1$ (Table 2)
The metric used in the challenge to evaluate and rank the submissions is computed from the average of these misclassification costs.
More precisely, if we name $P(S)$ the prediction of a slide S labelled $l(S)$, the challenge metric M_{WA} is:

$$M_{WA} = \frac{1}{N} \sum_{S} (1 - C_{l(S),P(S)}) \tag{1}$$

with N the number of samples.
The problem is thus framed as a cost-sensitive classification problem, and, to our knowledge, all the winning solutions took awareness of this cost in their training procedure.

Ground Truth	**Benign (pred)**	**Low-grade (pred)**	**High-grade (pred)**	**Carcinoma (pred)**
Benign	0.0	0.1	0.7	1.0
Low-grade	0.1	0.0	0.3	0.7
High-grade	0.7	0.3	0.0	0.3
Carcinoma	1.0	0.7	0.3	0.0

Table 2: Weighted Accuracy Error Table - Error table to ponderate misclassification according to their gap with the ground truth

4 Proposed Architecture

4.1 Multiple Instance Learning and Attention

In Multiple Instance Learning, we are given sets of samples $B_k = \{x_i | i = 1 \dots N_k\}$, also called bags. The annotation y_k we are given refers only to the bags and not the individual samples. We assume however, that such tile-level labels exist in principle, but that we just do not have access to them.

The strategy is to first map each tile x_i to its encoding z_i, which is then mapped to a scalar value a_i, often referred to as attention score. The tile representations z_i and attention scores a_i are then agglomerated to build the slide representation s_k which is then further processed by a neural network. The agglomeration can be based on tile selection [Campanella et al., 2019, Courtiol et al., 2020], or on an attention mechanism [Ilse et al., 2018], which is today the most widely used strategy.

4.2 Self-Supervised Learning

Self-supervised learning provides a framework to train neural networks without human supervision. The main goal of self-supervised learning is to learn to extract efficient features with inputs and labels derived from the data itself using a pretext task. Many self-supervised approaches are based on contrastive learning in the feature space. SimCLR, a simple framework relying on data augmentation was introduced in [Chen et al., 2020]. Powerful feature representations are learned by maximizing agreement between differently augmented views of the same data point via a contrastive loss applied in the feature space.

An image is transformed through random data augmentations into two new images. They are then embedded using the feature extractor. The two features vectors (z_i and z_j) are mapped with a projection head (dense layers) to obtain final vectors h_i and h_j . The feature extractor and projection head are trained to maximize agreement using the contrastive loss. Positive pairs consist of the two augmented views of the same image, the other $2(n-1)$ views play the role of negative samples. The loss function (NT-Xent) for a positive pair (i, j) is defined as:

$$\mathcal{L}_{SSL} = -log \frac{exp(sim(h_i, h_j)/\tau}{\sum_{k=1}^{2n} \mathbf{1}_{k \neq i} exp(sim(h_i, h_j)\ \tau} \tag{2}$$

Where $sim(u, v) = \frac{u^T v}{||u|| \cdot ||v||}$, the cosine similarity, $\mathbf{1}_{k \neq i \in (0,1)}$ determines if $k \neq i$ and τ is a parameter. After convergence, the projection head is discarded and the pretrained feature extractor can be used for subsequent tasks.

4.3 Cost-Sensitive Training

Instead of the traditional cross-entropy loss we used a cost-aware classification loss, the Smooth-One-Sided Regression Loss $\mathcal{L}_{SOSR}$. First introduced to train SVMs in [Tu and Lin, 2010], this objective function was smoothed and adapted for backpropagation in deep networks in [Chung et al., 2016]. When using this loss, the network is trained to predict the class-specific risk rather than a posterior probability; the decision function chooses the class minimizing this risk.
The SOSR loss is defined as follows:

$$\mathcal{L}_{SOSR} = \sum_i \sum_j ln(1 + exp(\mathbf{2}_{i,j} \cdot (\hat{c}_i - \mathcal{C}_{i,j}))) \tag{3}$$

With $\mathbf{2}_{i,j} = -\mathbf{1}_{i \neq j} + \mathbf{1}_{i=j}$, $\hat{c}_i$ the i-th coordinate of the network output and $\mathcal{C}$ the error table.

4.4 Mixed Supervision

To be tractable, training of attention-MIL architectures requires freezing the feature extractor weights. While SSL allows the feature extractor to build meaningful representations [Saillard et al., 2021, Dehaene et al., 2020], they are not specialized to the actual classification problems we try to solve. Several studies have shown that such SSL models benefit from fine-tuning specific to the downstream task [Chen et al., 2020]
We therefore added a training step to leverage the tile-level annotation and fine-tune the self-supervised model. However, as the final WSI classification task is not identical to the tile classification task, we suspect that fine-tuning solely on the tile classification task may over-specialize the feature extractor and thus sacrifice the generalizability of SSL (and for this reason ultimately also degrading the WSI classification performances). To avoid this, we developed a training process that optimizes the self-supervised and tile-classification objectives jointly.

Two different heads, plugged before the final classification layer, are used to compute both loss functions $\mathcal{L}_{SSL}$ and $\mathcal{L}_{SOSR}$ The final objective $\mathcal{L}$ is then:

$$\mathcal{L} = \beta \mathcal{L}_{SSL} + (1 - \beta) \mathcal{L}_{SOSR} \tag{4}$$

where β is a hyperparameter that has to be tuned. Here, we found $\beta = 0.3$ (see Supplementary).

5 Understanding the Feature Extractor with Activation Maximization

To further understand the features learned by the different pre-training policies (ImageNet, supervised, SSL and mixed), we used Activation Maximization (AM) to visualize extracted features and provide an explicit illustration of the specificity learned.
Methods to generate pseudo-images maximizing a feature activation have been introduced in [Erhan et al., 2009]. This technique consists in synthesizing the images that will maximize one feature activation. It is summarized as follow [Nguyen et al., 2019]:
If we consider a trained classifier with set of parameters θ that map an input image $x \in \mathbb{R}^{h \times w \times c}$, ($h$ and w are the height and width and c the number of channels) to a probability distribution over the classes, we can formulate the following optimization problem:

$$x^* = \arg\max_x (\sigma_i^l(\theta, x)) \tag{5}$$

where $\sigma_i^l(\theta, x)$ is the activation of the neuron i in a given layer l of the classifier. This formulation being a non-convex problem, local maximum can be found by gradient ascent, using the following

update step:

$$x_{t+1} = x_t + \epsilon \frac{\partial \sigma_i^l(\theta, x)}{\partial x_t} \tag{6}$$

The optimization process starts with a randomly initialized image. After a few steps, it generates an image which can help to understand what information is being captured by the feature. As we try to visualize meaningful representations of the features, some regularization steps are applied to the random noise input (random crop and rotations to generate more stable visualization, details can be found in Supplementary Materials). To generate filter visualization within the HE space, we transformed the RGB random image to HE input thanks to color deconvolution [Ruifrok and Johnston, 2001]. This preprocessing allowed to generate images with histology-like colors when converted back to the RGB space.

To select the most meaningful features for each class, we trained a Lasso classifier without bias to classify the extracted feature vectors into the four classes of the dataset for the four pre-training policies. The feature vectors for each tile were first normalized and divided element-wise by the vector of features' standard deviation across all the tiles. The L1 regularization factor λ was set to 0.01. Details about Lasso training can be found in Supplementary Materials. Contribution scores for each feature were therefore derived from the weights of the Lasso linear classifier: negative weights were removed and remaining positive weights were divided by their sum to obtain contribution scores $[0, 1]$. By filtering out the negative weights, the contribution score corresponds to the proportion of attribution among the features positively correlated to a class, and allows to select feature capturing semantic information related to the class, leaving out those containing information for other classes.

6 Experimental Setting

6.1 WSI Preprocessing

Preprocessing on a downsampled version of the WSIs was applied to select only tissue area and non-overlapping tiles of 224x224 pixels were extracted at a resolution of 1 mpp. (Details in Supplementary Materials)

6.2 Data Splits for Cross-Validation

To measure the performances of our models we performed 3-fold cross-validation for all our training settings. Because the annotated tiles used in our joint-optimization step were directly extracted from the slides themselves, we carefully split the tiles such that tiles in different folds were guaranteed to originate from different slides. The split divided the slides and tiles into a training set, a validation set and a test set.

All subsequent performance results are then reported as the average and standard deviation of the performance results on each of these 3 test folds.

6.3 Feature Extractor Pre-Training

The feature extractor is initialized with pre-trained weights obtained with three distinct supervision policies: fully supervised, self-supervised or a mix of supervision. These three policies rely on the fine-tuning of a DenseNet121 [Huang et al., 2016], pretrained on ImageNet. The fully-supervised archicture is fine-tuned solely on the tile classification task. The SSL architecture is derived from SimCLR framework and is trained on an unlabeled dataset of 1 million tiles extracted from the slides. Finally for the mixed-supervised architecture, a supervised branch is added to the previous SSL network and trained using the mixed objective function (see Fig. 1 and Eq. 4) on the fully supervised dataset. Technical details of these three training settings are available in the supplementary material.

6.4 Whole Slide Classification

After tiling the slides, the frozen feature extractor (DenseNet121) was applied to extract meaningful representations from the tiles. This feature extractor was initialized sequentially with the pre-trained weights mentioned above and generated as many sets of features. These bags of features were then used to train the Attention-MIL model with SOSR loss applied slide-wise. (Supplementary Materials).

6.5 Feature Visualization

To select the most relevant features, we trained an unbiased linear model on the feature vectors extracted from the annotated tiles. The feature vectors were standardised. The weights of the linear model were used to determine which features were the most impactful for each class. Feature visualizations were generated for the selected features and for each set of pre-trained weights. We extracted the tiles expressing the most of these features by selecting the feature vectors with the higher activation for the concerned feature. Implementation details are provided in Supplementary Materials.

7 Results

7.1 Self-Supervised Fine-Tuning

We saved the checkpoints of the self-supervised feature extraction model at each epoch of training, allowing us to investigate the amount of time needed to reach good WSI classification performances. We computed the embeddings of the whole dataset with each of the checkpoints and trained a WSI classifier from them. Figure 3 reports the performances of WSI classification models for each of these checkpoints. SSL training led to a higher Weighted Accuracy than using ImageNet weights after 3 epochs and resulted in a gain of +4.8% after 100 epochs. Interestingly, as little as 6 epochs

of training are enough to gain 4% of Weighted Accuracy: a significant boost in performance is possible with 50 GPU-hours of training. We then observe a small increase in performance until the 100th epochs.

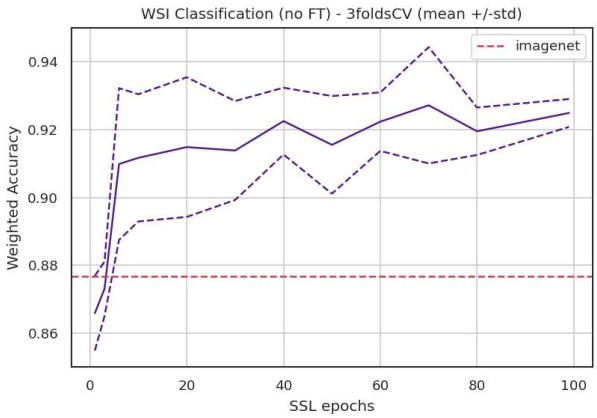

Figure 3: **Weighted Accuracy evolution** - Weighted Accuracy evolution on WS classification task with respect to the number of epochs of SSL training

7.2 Pre-Training Policy Comparison

To compare the weights obtained with the various supervision levels, we ran a 3-fold cross-validation on the WS classification task and summarized the results in table 3. The results indicate that the SSL pre-training substantially improves the WSI classification performance. In contrast, we see that initializing the feature extractor with fully-supervised weights gives an equivalent or poorer performance than any other initialization. SSL pre-training allows us to extract rich features that are generic, yet still relevant to the dataset (unlike ImageNet). On the other hand, fully supervised features are probably too specific and seem to not represent the full diversity of the image data. The joint-optimization process manages to balance out generic and specialized features without neutralizing them: mixing the supervision levels brings significant improvements (+2%) to the performance, leading to a Weighted Accuracy of 0.945.
We additionally compared the benefits introduced by the cost-sensitive loss (Eq. 3) with the cross-entropy loss. Our results show that with ImageNet weights the SOSR loss improves the Weighted Accuracy by 1% and the accuracy by 3%.
In conclusion, the combination of the SSL pre-trained model, its fully supervised fine-tuning, and the cost-sensitive loss leads to a notable improvement of 8 Weighted Accuracy points over the baseline MIL-imagenet model.

	Accuracy	SFP_metric
imagenet+ce	0.758 +/- 0.034	0.865 +/- 0.023
imagenet+sors	0.787 +/- 0.032	0.877 +/- 0.029
supervised+sors	0.772 +/- 0.055	0.874 +/- 0.027
ssl+sors	0.803 +/- 0.016	0.925 +/- 0.006
mixed+sors	**0.845 +/- 0.028**	**0.945 +/- 0.005**

Table 3: **Pre-training policies** - Performances summary

7.3 Number of Annotations vs Number of Epochs

We have seen that both SSL and supervised pre-training bring together an improvement in the WSI classification task. To further investigate the relationship between these two supervision regimes, we trained models with only some of the fully supervised annotations (15, 65, 100%) on top of intermediate SSL checkpoints. Results are reported in table 4.
It appears that without SSL pre-training (or with too few epochs of training), the supervised fine-tuning does not bring additional improvement in WSI classification. This is in line with the work of Chen et al. (Chen, Kornblith, Swersky, et al. 2020) that showed that an SSL model is up to 10x more label efficient than a supervised one.
However, for the 100-epoch checkpoint, we observe an improvement of 2 points of the Weighted Accuracy when using 100% of the tile annotations. Moreover, fine-tuning the models by mixed supervision with too few annotations (15%) leads to a slight drop in WSI classification performances. Finally, we see a diminution of the standard deviations across splits for the different pre-training policies, showing better stability for longer SSL training and more annotations.

We draw different conclusions from these observations:

- In this context, it is always better to pre-train the feature extractor with SSL rather than only invest in annotations.
- The supervised fine-tuning needs enough annotations to bring an improvement to the WSI classification task. We can note however that even when considering the 100% annotation settings, the supervised dataset (approx. 5000 images) is still rather small in comparison to traditional image datasets.
- A full SSL training is mandatory to leverage this small amount of supervised data.

7.4 Features Visualisations

We generated the pseudo-images of the most important features for each class and each pre-training policy and extracted the related tiles. The Figure 5 displays the most important features along with

	0 Annot.	∼1 Annot. / slide (1015 tiles)	∼4 Annot. / slide (3901 tiles)	∼6 Annot. / slide (5926 tiles)
ImageNet (no SSL)	0,877 +/- 0.029	0.872 +/- 0.024	0.872 +/- 0.023	0,874 +/-0.027
SSL-epoch10	0,912 +/- 0.019	0,907+/- 0.024	0,903 +/- 0.029	0,916 +/- 0.019
SSL-epoch50	0,915 +/- 0.014	0,913 +/- 0.024	0,916 +/- 0.014	0,914 +/- 0.022
SLL-epoch100	0,925 +/- 0.006	0,916 +/- 0.010	0,921 +/- 0.010	**0,945 +/- 0.005**

Table 4: **Relationship between self-supervision and full-supervision** - Study on the performance improvement on WS classification for different proportion of labelled data versus different training time of SSL

the tiles activating each feature the most for the class "Normal" (0). Although interpretation of such pseudo images must be treated carefully, we notice that the features obtained with SSL, supervised and mixed training are indubitably more specialized to histological data than those obtained with ImageNet. Some histological patterns, such as nuclei, squamous cells or basal layers are clearly identifiable in the generated images. The extracted tiles are strongly correlated with class-specific biomarkers. Feature **e** represents a normal squamous maturation, i.e. a layer of uniform and rounded basal cells, with slightly larger and bluer nuclei than mature cells. We can also observe several layers of mature cells with small nuclei and moderately abundant cytoplasm (pink halo around), equidistant from each other. Features **c** and **d** highlight clouds of small regular and rounded nuclei (benign cytological signs). Feature **g** and **h** are characteristic of squamous cells (polygonal shapes, stratified organization lying on a straight basal layer). Interestingly, features extracted with the supervised method (**g**, **h**) manage to sketch a normal epithelium with high resemblance, the features are more precise. On the other hand, features extracted with SSL (**c**, **d**) highlight true benign criteria but do not entirely summarize a normal epithelium (no basal maturation). The mixed model displays both, suggesting that mixed supervision highlights pathologically relevant patterns to a larger extent than the other regimes [WHO, 2020].
We can also note by looking at the real tiles that while features from ImageNet (**a**, **b**), SSL (**c**, **d**) and the supervised model (**g**, **h**) focus on the upper half of the cervix epithelium, it appears that features from the mixed supervision model (**e**, **f**) are focusing on the lower half which is known to be the relevant region for discrimination between class Normal (0) and Low Grade (1) (abnormal cells are constricted to the lower third of the epithelium).
In Figure 4 we can further identify class-related biomarkers for dysplasia and carcinoma grade. Tiles with visible koilocytes (cells with a white halo around the nucleus) have been extracted from the top features for Low Grade class. Koilocytes are symptomatic of infection by Human Papillomavirus and are a key element for this diagnosis (almost always responsible for precancerous lesions in the cervix, [WHO, 2020]). High Grade (2) generated image represents disorganised cells with a high nuclear-to-cytoplasmic ratio, marked variations in size and shape and loss of polarity. For the class "Carcinoma" (3), we observe irregular clusters of cohesive cells with very atypical nuclei, separated by a fibrous texture that can be identified as stroma reaction. All these criteria have been identified in [WHO, 2020] as key elements for diagnosis of dysplasia and invasive carcinoma.

In Figure 6, we observe that features extracted from ImageNet and SSL models are diverse, in particular, features extracted from SSL reflect rich tissue phenotypes which correlates to their generic capacities of image representations. On the other hand, features extracted with supervised and mixed methods are more redundant. We additionally observe in Figure 6 that feature visualisation from the mixed model picture realistic histopathological patterns specific to the class. Visualisation for other classes are available in Supplementary Materials.

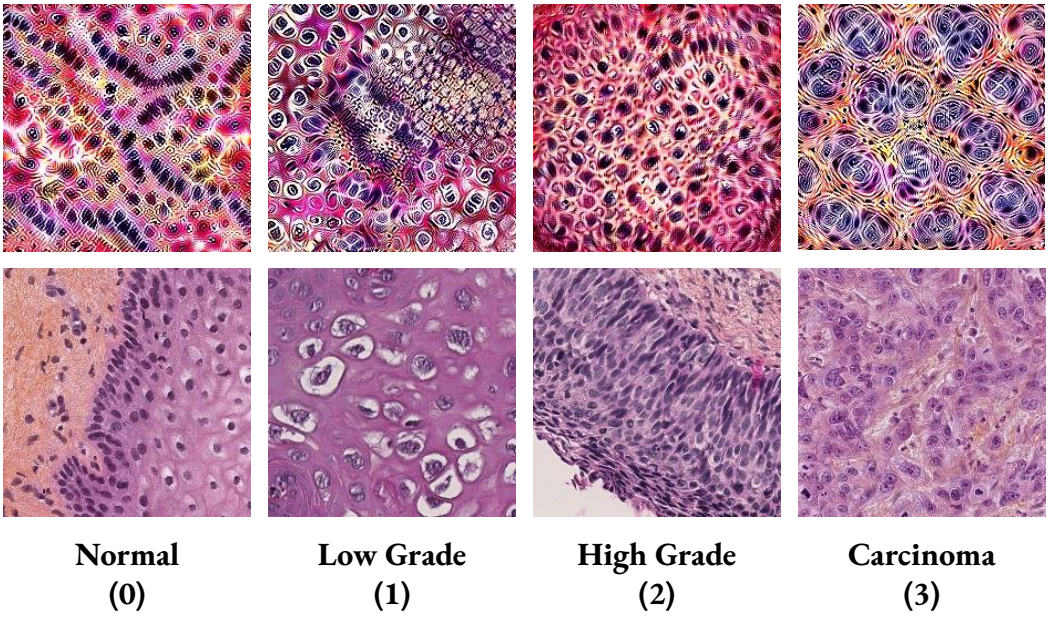

Figure 4: **Feature comparison per class** - The top row displays the top filter for the Mixed Supervised model for each class. The bottom row displays the tile expressing the feature the most

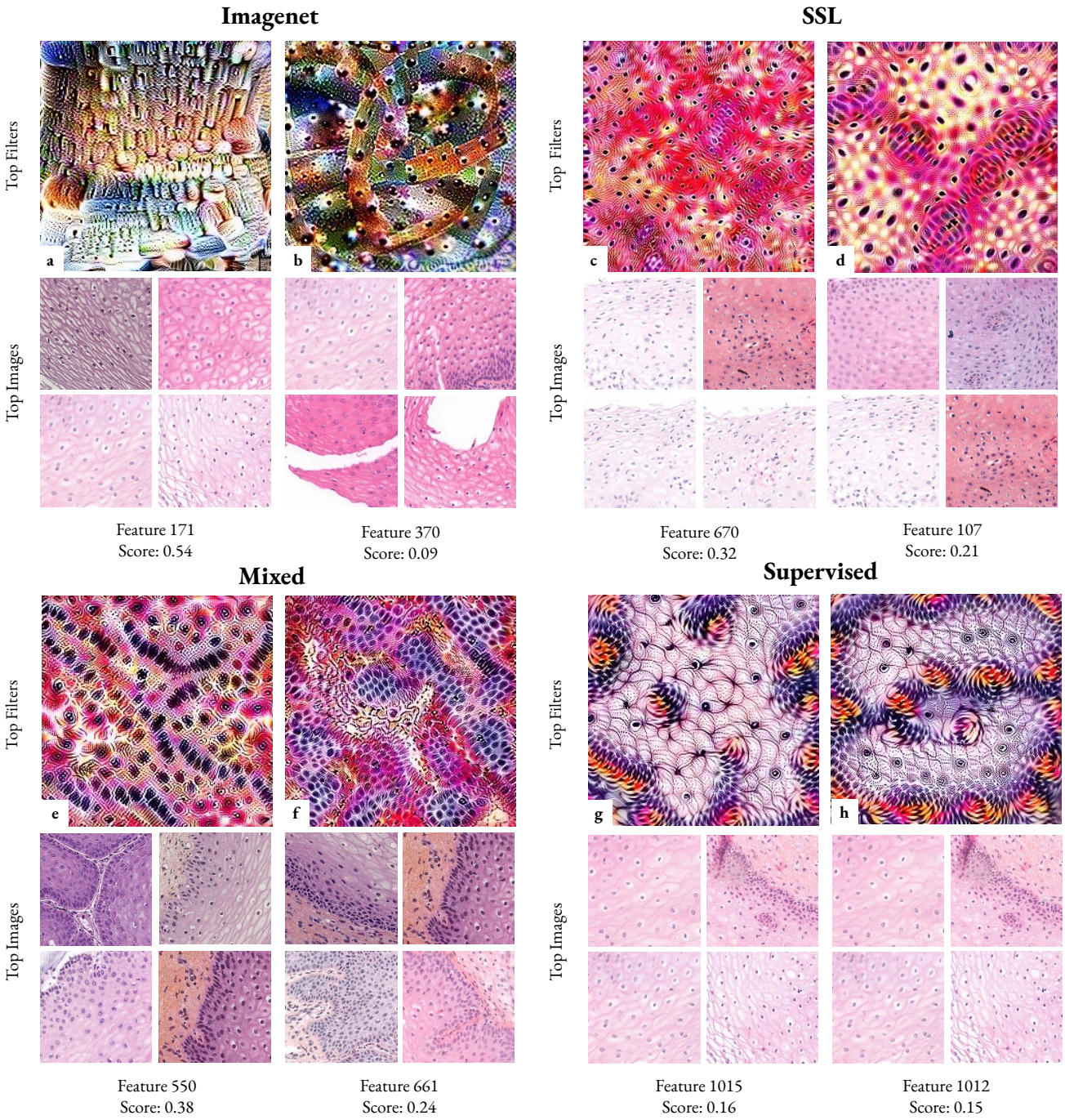

Figure 5: **Feature Visualization** - Top Features for class "Normal" (0) and associated tiles

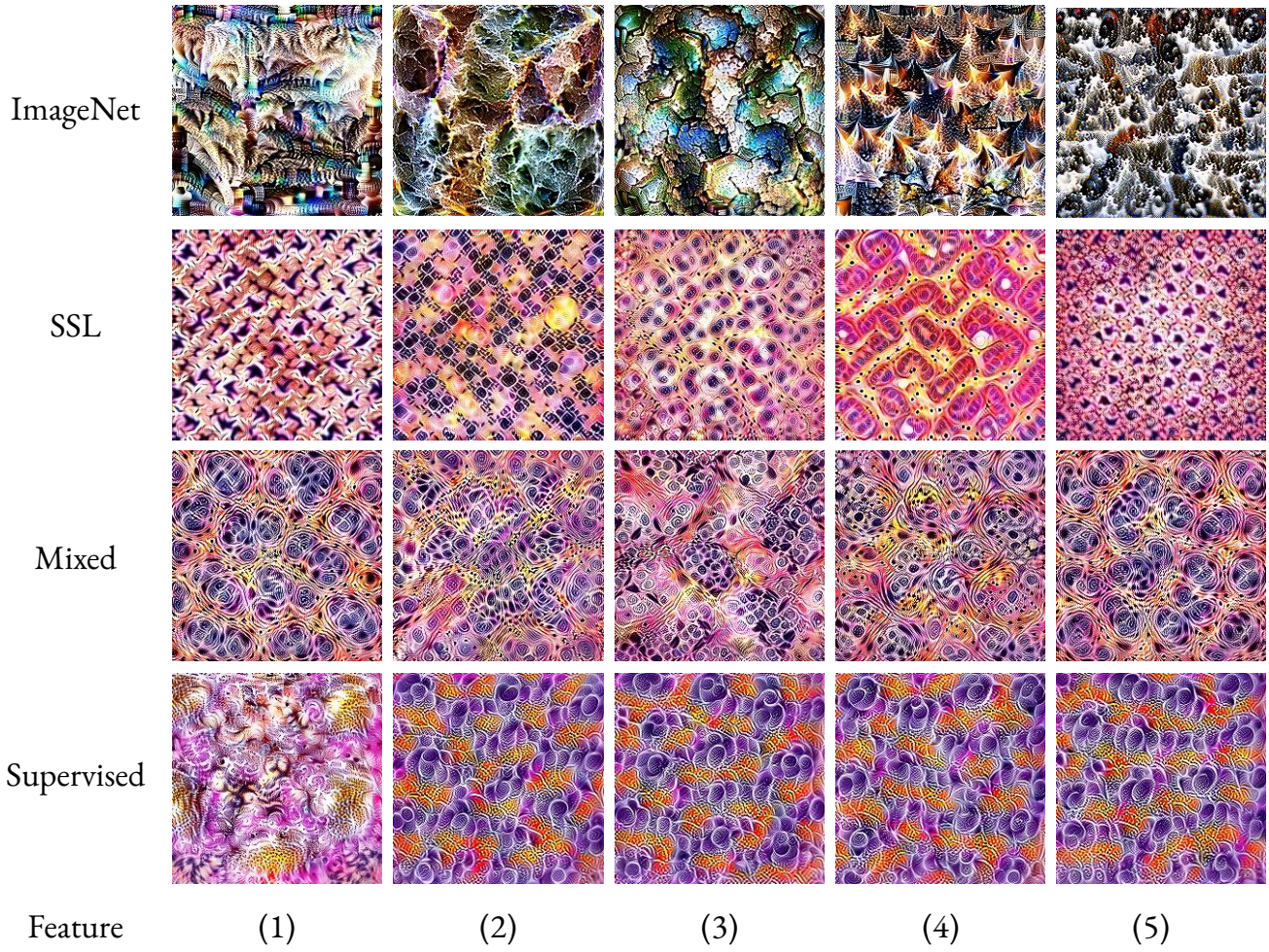

Figure 6: **Feature Diversity for the class "Carcinoma" (3) (top 5 features)** - Class "Normal" (0) and top 10 features in Supplementary Materials

8 Discussion

In pathology , expert annotations are usually hard to obtain. However, we are often in a situation where a small proportion of labeled annotation exists but not in sufficient quantities to support fully supervised techniques. Yet, even in small quantities, expert annotations carry meaningful information that one could use to enforce biological context to deep learning models and make sure that networks learn appropriate patterns. On the other hand, self-supervised methods have proven their efficacy to extract generic features in the histopathological domain and their usefulness for

downstream supervision tasks, even in the absence of massive ground truth data. Methods capable of reconciling self-supervision with strong supervision can therefore be useful and open the door to better performances.

In this paper, we presented a way to inject the fine-grained tile level information by fine-tuning the feature extractor with a joint optimization process. This process allowed to mix self-supervised learning features with tile classification ones and helped the downstream WSI classification task.

We applied our method to the TissueNet Challenge, a challenge for the automatic grading of cervix cancer, that provided annotations at the slide and tile level, thus representing an appropriate use case to validate our method of mixed supervision. We also propose in this study insights and guidelines for the training of a WSI classifier in the presence of tile annotations.

First, we showed that SSL is always beneficial to our downstream WSI classification tasks. Fine-tuning pre-trained weights with SSL for only 50 hours brings a 4% improvement over WSI classification weighted accuracy, and near to 5% when fine-tuning for longer (100 epochs).

Second, a small set of annotated tiles can bring benefit to the WSI classification task, up to 2% of weighted accuracy for a supervised dataset of around 5000 images.
Such a set of tiles can be obtained easily by asking the pathologist to select a few ROIs that guided his decision while labeling the WSIs, which can be achieved without a strong time commitment. However this boost in performance can be reached only if the feature extractor is pre-trained with SSL, and for sufficiently long: SSL unlocks the supervised fine-tuning benefits.

To further understand the differences between the range of supervision used to extract tile features, we conducted qualitative analysis on features visualizations by activation maximization and observed that features obtained from SSL, supervised or mixed trainings were more relevant for histological tasks and that class-discriminative patterns were indeed identified by the model. We also observed that supervised training on the tiles alone led to much less diversity in the features extracted by the model than the ones obtained with SSL.

The scope of this study contains by design three limitations. First, SSL models were trained by fine-tuning already pre-trained weights on imagenet. This may explain the rapid convergence and boost in performance observed; however it may also underestimate this boost if the SSL models were trained from scratch. We did not compare SSL trained from scratch and fine-tuned SSL, and left it to future work.

Second, all the conclusions reached are conditioned by the fact that we do not fine-tune the feature extractor network during the WSI classification training. Keeping these weights frozen, and even pre-computing the tiles representations brings a large computational benefit (both in memory

and speed of computations), but prevents the feature extractor from specializing during the WSI classification training.

Third, the tendency observed in table 4 of better performances correlated with larger numbers of annotations is modest and would require more annotations to validate it.

Finally, our method can be improved in several ways. First, SimCLR, was a pioneer method in self-supervised learning architecture and has proven to be efficient but it suffers from high performance drop when decreasing the batch size [Chen et al., 2020]. Other SSL models have been developed to alleviate this limitation. MoCo [He et al., 2020] actually propose a momentum mechanism allowing optimal performances even without large batch size and therefore, numerous available parallel GPUs. Other models like VICReg [Bardes et al., 2021] proposed techniques to maximize the variance between the features and therefore limit their redundancies. It will be interesting to benchmark these SSL variants with respect to their impact on WSI classification accuracy and feature interpretability.

To conclude, we present a method that provides an interesting alternative to using full supervision, pre-training on unrelated data sets or self-supervision. We convincingly show that the learned feature representations are both leading to higher performance and providing more intermediate features that are more adapted to the problem and point to relevant cell and tissue phenotypes. We expect that the mixed supervision will be adopted by the field and lead to better models.

Acknowledgments

The authors thank Etienne Decencière for the thoughful discussions that helped the project. ML was supported by a CIFRE PhD fellowship founded by KEEN EYE, Paris, France and ANRT (CIFRE 2019/1905). TL was supported by a Q-Life PhD fellowship (Q-life ANR-17-CONV-0005). Furthermore, this work was supported by the French government under management of Agence Nationale de la Recherche as part of the "Investissements d'avenir" program, reference ANR-19-P3IA-0001 (PRAIRIE 3IA Institute).

References

Adrien Bardes, Jean Ponce, and Yann LeCun. VICReg: Variance-Invariance-Covariance Regularization for Self-Supervised Learning. *arXiv:2105.04906 [cs]*, October 2021. URL `http://arxiv.org/abs/2105.04906`.

Gabriele Campanella, Matthew G. Hanna, Luke Geneslaw, Allen Miraflor, Vitor Werneck Krauss Silva, Klaus J. Busam, Edi Brogi, Victor E. Reuter, David S. Klimstra, and Thomas J.

Fuchs. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature Medicine*, 25(8):1301–1309, August 2019. ISSN 1546-170X. doi: 10.1038/s41591-019-0508-1. URL https://www.nature.com/articles/s41591-019-0508-1.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A Simple Framework for Contrastive Learning of Visual Representations. February 2020. URL https://arxiv.org/abs/2002.05709v3.

Yu-An Chung, Hsuan-Tien Lin, and Shao-Wen Yang. Cost-Aware Pre-Training for Multiclass Cost-Sensitive Deep Learning. *IJCAI*, 2016.

Nicolas Coudray, Paolo Santiago Ocampo, Theodore Sakellaropoulos, Navneet Narula, Matija Snuderl, David Fenyö, Andre L. Moreira, Narges Razavian, and Aristotelis Tsirigos. Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. *Nature Medicine*, 24(10):1559–1567, October 2018. doi: 10.1038/s41591-018-0177-5. URL https://www.nature.com/articles/s41591-018-0177-5.

Pierre Courtiol, Eric W. Tramel, Marc Sanselme, and Gilles Wainrib. Classification and Disease Localization in Histopathology Using Only Global Labels: A Weakly-Supervised Approach. *arXiv:1802.02212 [cs, stat]*, February 2020. URL http://arxiv.org/abs/1802.02212. arXiv: 1802.02212.

Olivier Dehaene, Axel Camara, Olivier Moindrot, Axel de Lavergne, and Pierre Courtiol. Self-Supervision Closes the Gap Between Weak and Strong Supervision in Histology. December 2020. URL https://arxiv.org/abs/2012.03583v1.

James A. Diao, Jason K. Wang, Wan Fung Chui, Victoria Mountain, Sai Chowdary Gullapally, Ramprakash Srinivasan, Richard N. Mitchell, Benjamin Glass, Sara Hoffman, Sudha K. Rao, Chirag Maheshwari, Abhik Lahiri, Aaditya Prakash, Ryan McLoughlin, Jennifer K. Kerner, Murray B. Resnick, Michael C. Montalto, Aditya Khosla, Ilan N. Wapinski, Andrew H. Beck, Hunter L. Elliott, and Amaro Taylor-Weiner. Human-interpretable image features derived from densely mapped cancer pathology slides predict diverse molecular phenotypes. *Nature Communications*, 12(1):1613, March 2021. ISSN 2041-1723. doi: 10.1038/s41467-021-21896-9. URL https://www.nature.com/articles/s41467-021-21896-9.

DrivenData. TissueNet: Detect Lesions in Cervical Biopsies. URL https://www.drivendata.org/competitions/67/competition-cervical-biopsy/page/254/.

Babak Ehteshami Bejnordi, Mitko Veta, Paul Johannes van Diest, Bram van Ginneken, Nico Karssemeijer, Geert Litjens, Jeroen A. W. M. van der Laak, and and the CAMELYON16 Consortium. Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer. *JAMA*, 318(22):2199–2210, December 2017. ISSN 0098-7484. doi: 10.1001/jama.2017.14585. URL https://doi.org/10.1001/jama.2017.14585.

Dumitru Erhan, Y. Bengio, Aaron Courville, and Pascal Vincent. Visualizing Higher-Layer Features of a Deep Network. *Technical Report, Univeristé de Montréal*, January 2009.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum Contrast for Unsupervised Visual Representation Learning. *arXiv:1911.05722 [cs]*, March 2020. URL http://arxiv.org/abs/1911.05722. arXiv: 1911.05722.

Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely Connected Convolutional Networks. Technical report, August 2016. URL https://ui.adsabs.harvard.edu/abs/2016arXiv160806993H.

Yi-Jie Huang, Weiping Liu, Xiuying Wang, Qu Fang, Renzhen Wang, Yi Wang, Huai Chen, Hao Chen, Deyu Meng, and Lisheng Wang. Rectifying Supporting Regions With Mixed and Active Supervision for Rib Fracture Recognition. *IEEE Transactions on Medical Imaging*, 39(12):3843–3854, December 2020. ISSN 1558-254X. doi: 10.1109/TMI.2020.3006138.

Maximilian Ilse, Jakub M. Tomczak, and Max Welling. Attention-based Deep Multiple Instance Learning. February 2018. URL https://arxiv.org/abs/1802.04712v4.

Jakob Nikolas Kather, Lara R. Heij, Heike I. Grabsch, Chiara Loeffler, Amelie Echle, Hannah Sophie Muti, Jeremias Krause, Jan M. Niehues, Kai A. J. Sommer, Peter Bankhead, Loes F. S. Kooreman, Jefree J. Schulte, Nicole A. Cipriani, Roman D. Buelow, Peter Boor, Nadina Ortiz-Brüchle, Andrew M. Hanby, Valerie Speirs, Sara Kochanny, Akash Patnaik, Andrew Srisuwananukorn, Hermann Brenner, Michael Hoffmeister, Piet A. van den Brandt, Dirk Jäger, Christian Trautwein, Alexander T. Pearson, and Tom Luedde. Pan-cancer image-based detection of clinically actionable genetic alterations. *Nature Cancer*, 1(8):789–799, August 2020. ISSN 2662-1347. doi: 10.1038/s43018-020-0087-6. URL https://www.nature.com/articles/s43018-020-0087-6.

Tristan Lazard, Guillaume Bataillon, Peter Naylor, Tatiana Popova, François-Clément Bidard, Dominique Stoppa-Lyonnet, Marc-Henri Stern, Etienne Decencière, Thomas Walter, and Anne Vincent Salomon. Deep Learning identifies new morphological patterns of Homologous Recombination Deficiency in luminal breast cancers from whole slide images. Technical report, September 2021. URL https://www.biorxiv.org/content/10.1101/2021.09.10.459734v1. Company: Cold Spring Harbor Laboratory Distributor: Cold Spring Harbor Laboratory Label: Cold Spring Harbor Laboratory Section: New Results Type: article.

Jiahui Li, Wen Chen, Xiaodi Huang, Shuang Yang, Zhiqiang Hu, Qi Duan, Dimitris N. Metaxas, Hongsheng Li, and Shaoting Zhang. Hybrid Supervision Learning for Pathology Whole Slide Image Classification. In Marleen de Bruijne, Philippe C. Cattin, Stéphane Cotin, Nicolas Padoy, Stefanie Speidel, Yefeng Zheng, and Caroline Essert, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, pages 309–318. Springer International Publishing, 2021.

Zhe Li, Chong Wang, Mei Han, Emily Xue, Wei Wei, Jia Li, and Fei-Fei Li. Thoracic Disease Identification and Localization with Limited Supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

Ming Y. Lu, Drew F. K. Williamson, Tiffany Y. Chen, Richard J. Chen, Matteo Barbieri, and Faisal Mahmood. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature Biomedical Engineering*, 5(6):555–570, June 2021. ISSN 2157-846X. doi: 10.1038/s41551-020-00682-w. URL https://www.nature.com/articles/s41551-020-00682-w.

Pawel Mlynarski, Hervé Delingette, Antonio Criminisi, and Nicholas Ayache. Deep learning with mixed supervision for brain tumor segmentation. *Journal of Medical Imaging*, 6(3):034002, August 2019. ISSN 2329-4302, 2329-4310. doi: 10.1117/1.JMI.6.3.034002. URL https://www.spiedigitallibrary.org/journals/journal-of-medical-imaging/volume-6/issue-3/034002/Deep-learning-with-mixed-supervision-for-brain-tumor-segmentation/10.1117/1.JMI.6.3.034002.full.

National Cancer Institute. Definition of cervical dysplasia - NCI Dictionary of Cancer Terms, February 2011. URL https://www.cancer.gov/publications/dictionaries/cancer-terms/def/cervical-dysplasia. Archive Location: nciglobal,ncienterprise.

Anh Nguyen, Jason Yosinski, and Jeff Clune. Understanding Neural Networks via Feature Visualization: A survey. *arXiv:1904.08939 [cs, stat]*, April 2019. URL http://arxiv.org/abs/1904.08939. arXiv: 1904.08939.

A. C. Ruifrok and D. A. Johnston. Quantification of histochemical staining by color deconvolution. *Analytical and Quantitative Cytology and Histology*, 23(4):291–299, August 2001.

Charlie Saillard, Flore Delecourt, Benoit Schmauch, Olivier Moindrot, Magali Svrcek, Armelle Bardier-Dupas, Jean Francois Emile, Mira Ayadi, Louis De Mestier, Pascal Hammel, Cindy Neuzillet, Jean-Baptiste Bachet, Juan Iovanna, Dusetti J Nelson, Valerie Paradis, Mikhail Zaslavskiy, Aurelie Kamoun, Pierre Courtiol, Remy Nicolle, and Jerome Cros. Identification of pancreatic adenocarcinoma molecular subtypes on histology slides using deep learning models. *Journal of Clinical Oncology*, 39(15_suppl):4141–4141, May 2021. ISSN 0732-183X. doi: 10.1200/JCO.2021.39.15_suppl.4141. URL https://ascopubs.org/doi/abs/10.1200/JCO.2021.39.15_suppl.4141.

Paul Tourniaire, Marius Ilie, Paul Hofman, Nicholas Ayache, and Herve Delingette. Attention-based Multiple Instance Learning with Mixed Supervision on the Camelyon16 Dataset. In *Proceedings of the MICCAI Workshop on Computational Pathology*, pages 216–226. PMLR, September 2021. URL https://proceedings.mlr.press/v156/tourniaire21a.html.

Han-Hsing Tu and Hsuan-Tien Lin. One-sided Support Vector Regression for Multiclass Cost-sensitive Classification. page 8, 2010.

Mitko Veta, Paul J. van Diest, Stefan M. Willems, Haibo Wang, Anant Madabhushi, Angel Cruz-Roa, Fabio Gonzalez, Anders B. L. Larsen, Jacob S. Vestergaard, Anders B. Dahl, Dan C. Cireşan, Jürgen Schmidhuber, Alessandro Giusti, Luca M. Gambardella, F. Boray Tek, Thomas Walter, Ching-Wei Wang, Satoshi Kondo, Bogdan J. Matuszewski, Frederic Precioso, Violet Snell, Josef Kittler, Teofilo E. de Campos, Adnan M. Khan, Nasir M. Rajpoot, Evdokia Arkoumani, Miangela M. Lacle, Max A. Viergever, and Josien P. W. Pluim. Assessment of algorithms for mitosis detection in breast cancer histopathology images. *Medical Image Analysis*, 20(1):237–248, February 2015. ISSN 1361-8423. doi: 10.1016/j.media.2014.11.010.

WHO. Colposcopy and treatment of cervical intraepithelial neoplasia: a beginners' manual, 2020. URL `https://screening.iarc.fr/colpochap.php?chap=2`.