MS-CLAM: Mixed Supervision for the classification and localization of tumors in Whole Slide Images

Paul Tourniaire, Marius Ilie, Paul Hofman, Nicholas Ayache, Hervé Delingette

MS-CLAM: Mixed Supervision for the classification and localization of tumors in Whole Slide Images

Paul Tourniaire[a,*], Marius Ilie[b,c,d], Paul Hofman[b,c,d], Nicholas Ayache[a,d], Hervé Delingette[a,d]

[a] *Université Côte d'Azur, Inria, Epione project-team, Sophia Antipolis, France*
[b] *Laboratory of Clinical and Experimental Pathology, Pasteur Hospital, Université Côte d'Azur Nice, France*
[c] *Hospital-Related Biobank BB-0033-00025*
[d] *FHU OncoAge*

ARTICLE INFO

Keywords: Digital Pathology, Mixed Supervision, Deep Learning, Camelyon16, DigestPath2019

ABSTRACT

Given the size of digitized Whole Slide Images (WSIs), it is generally laborious and time-consuming for pathologists to exhaustively delineate objects within them, especially with datasets containing hundreds of slides to annotate. Most of the time, only slide-level labels are available, giving rise to the development of weakly-supervised models. However, it is often difficult to obtain from such models accurate object localization, e.g., patches with tumor cells in a tumor detection task, as they are mainly designed for slide-level classification. Using the attention-based deep Multiple Instance Learning (MIL) model as our base weakly-supervised model, we propose to use mixed supervision – i.e., the use of both slide-level and patch-level labels – to improve both the classification and the localization performances of the original model, using only a limited amount of patch-level labeled slides. In addition, we propose an attention loss term to regularize the attention between key instances, and a paired batch method to create balanced batches for the model. First, we show that the changes made to the model already improve its performance and interpretability in the weakly-supervised setting. Furthermore, when using only between 12 and 62% of the total available patch-level annotations, we can reach performance close to fully-supervised models on the tumor classification datasets DigestPath2019 and Camelyon16.

1. Introduction

With the digitization of histological slides, deep learning algorithms have reached state-of-the-art performance on several tasks, e.g., cancer detection (Wang et al., 2016), tumor grading (Bulten et al., 2020) or mutation predictions and prognosis (Fu et al., 2020). Nonetheless, Whole Slide Images (WSIs) still represent an atypical challenge in medical image analysis, as they often reach sizes of billions of pixels that are beyond the capacity of any current deep learning framework. For that matter, they are usually split into patches (or tiles) of smaller dimensions (e.g., 256x256 pixels), which are in turn processed by the models. As patch-level labels are usually unknown, because they are too time-consuming to obtain from expert pathologists, WSI analysis often falls under the Multiple Instance Learning (MIL) framework (Dietterich et al., 1997), where the slide is seen as a bag of which the tiles are instances. MIL often comes with weak supervision, meaning that only the slide-level label is known. Under this particular setting, two natural problems arise:

- Given the embeddings (or the probabilities) obtained at the instance level by a deep learning model, how can we recover the bag label? This is the MIL classification task.
- Given the bag label, is it possible to detect which are the key instances?

Regarding the latter, in a binary classification problem where one has to classify slides as containing tumorous tissue or not, the point is to be able to locate the tumor within the image, which is what we will refer to as tumor localization. This is a secondary task compared to the MIL classification one, but it is of great importance in histological image analysis, as it allows medical experts to confirm that the model's key instance selection *matches the slide prediction*. Liu et al. (2012) refer to this as Key Instance Detection (KID), while putting forward that a good KID method allows for better bag classification.

*Corresponding author
e-mail: paul.tourniaire@inria.fr (Paul Tourniaire)

1.1. Weakly-supervised classification

Under weak supervision, where only slide-level labels are known, several methods have been proposed to solve the binary tumor classification problem: in Coudray et al. (2018), the authors simply follow the assumption that the tile labels are the same as the slide label, and proceed with this assumption to train a neural network to classify tiles, the outputs of which are averaged to recover the slide-level prediction. In Campanella et al. (2019), the authors retrieve the S most suspicious tiles within the slide to feed a recurrent neural network (RNN) which in turn gives the slide-level classification. In Coudray et al. (2018), the authors adapt the WELDON method by Durand et al. (2016) to MIL for binary classification of WSIs. However, these methods mainly focused on the bag-level classification. More recently, Lerousseau et al. (2021) proposed a refined weak supervision approach by making use of the tumor cell percentage associated with each slide, instead of the sole binary label. With a training set of more than 18,000 slides from multiple cancer sites, they showed that they could outperform a fully-supervised model that was trained with tile-level labels on a binary tumor classification problem, while also producing convincing segmentation masks. For most of the aforementioned methods however, very large datasets were leveraged to obtain such good results, which are not necessarily available for every histological classification problem.

Self-supervision methods such as contrastive learning (Chopra et al., 2005), the aim of which is to minimize the distance between similar samples within a latent space, have been used as a training method to improve the feature extractors in histopathology. Using the Momentum Contrast v2 (Chen et al., 2020b) self-supervised framework to train a ResNet-50 neural network (He et al., 2016) as their tile-level feature extractor, Dehaene et al. (2020), showed that they could reach slide-level classification scores close to the best fully-supervised method on the Camelyon16 challenge dataset (Bejnordi et al., 2017), but at the cost of intensive and time-consuming training, using many processing units.

1.2. Attention pooling

On the topic of pooling methods, Ilse et al. (2018) proposed a new pooling operator called attention pooling: instance-specific attention scores are learned during training to compute a weighted average of the instance feature vectors, which is then used to compute the bag-level prediction. The goal of this approach is to select the most relevant instances in a bag, in order to obtain a more accurate representation of the latter thanks to the weighted mean. Such a method also provides a way to interpret the decision result, by looking at the instances' attention scores during inference. There have since been several proposed modifications of the original architecture: Lu et al. (2021) presented CLAM, which stands for Clustering-constrained Attention Multiple Instance Learning. CLAM is a generalization of the attention-based MIL to multi-class classification, with data efficient performance.

Self-attention (Vaswani et al., 2017), which can be used to model the interactions between instances within the bags, was used by Rymarczyk et al. (2021) in addition to the classic attention-based MIL model (in which the instances are assumed to be independent), while exploring other MIL assumptions such as the presence-based or threshold-based assumptions. Unfortunately, Rymarczyk et al. (2021) do not linger on the consequences of self-attention on the resulting attention scores, as bag-level classification is the main focus of their study. Self-attention was also used by Li et al. (2021a), but this time as a distance measurement between the instance selected using max-pooling (denoted critical instance) and the other instances within the same bag, in a dual-stream model based on a self-supervised, multiscale feature extractor. Although their model integrates an instance classifier, it is mainly used for the critical instance selection, and serves no purpose in the localization of the tumor at inference time. Shi et al. (2020) on the other hand, establish several theorems regarding attention-based MIL, notably showing how the instances' attention scores influence the bag-level prediction. They propose another method to compute the attention scores, called loss-based attention, and show on several datasets that it yields higher bag-level classification scores, and also boosts instance recall. However, the method was only tested on small MIL datasets, with few instances per bags (e.g. tens of instances) compared to what is commonly found in histopathology datasets (e.g. thousands of instances), and the method does not consider in particular the case of negative bags in binary MIL classification.

1.3. Mixed Supervision

Shah et al. (2018) introduced mixed-supervision for image segmentation: strong supervision (i.e. pixel-level) and weak-supervision (bounding boxes, landmarks) were used together to improve segmentation while reducing the supervision cost. Mlynarski et al. (2019) defined *mixed supervision* in their work as the joint use of image-level and pixel-level labels within an image. Compared to using only a few fully-annotated images, they showed that using the latter along with additional weakly-labeled images to train the same model also improved the segmentation results.

As for computational pathology, a similar approach was used by Ciga and Martel (2021), where a two-headed patch-level ResNet-18 (He et al., 2016) was trained with both segmentation and classification labels, so as to reduce the segmentation labeling burden. In the case of WSI analysis, we define mixed supervision as the joint use of instance- and bag-level labels. This kind of approach has previously been associated to semi supervision (Marini et al., 2021), where instance or patch-level labels are denoted "strong annotations", and bag or slide-level labels "weak annotations". Other works based on this joint approach have been published, especially on the topic of prostate cancer grading (Arvaniti and Claassen, 2018; Bulten et al., 2020). However, all these works either focused on WSI or instance classification, but did not try to perform both tasks simultaneously, using the labels from the two distinct levels of supervision. More recently, Schmidt et al. (2022) proposed a model trained with both slide-level labels and a limited number of tile-level labels for tumor classification. The model is a tile-level classifier, that yields slide-level labels to generate

pseudo labels for weakly-augmented tiles. After obtaining the pseudo labels, the same tiles are strongly augmented and classified by the same model using their respective true or pseudo labels all together. A loss function composed of both supervised and unsupervised terms is used to train the network. Like most of the previously mentioned frameworks, this one falls more in the semi-supervised learning category than in the mixedly-supervised one. Indeed, the purpose of their method lies in the training of the feature extractor, and not in a model that works at the slide-level. In Lubrano di Scandalea et al. (2022), the authors devised a three-step approach, where a tile-level feature extractor is first trained in a self-supervised fashion using the SimCLR method (Chen et al., 2020a), then trained using both self-supervision and strong supervision on a small number of tiles. Once the feature extractor training is done, it is frozen, and used to extract features that will feed an attention-based deep MIL WSI classifier, in a similar way to Lu et al. (2021). This work, like the one by Dehaene et al. (2020), aims at improving the performance of a WSI classifier by using a feature extractor specifically trained on histological data: mixed supervision is only used during the fine-tuning of the feature extractor, and is not fully integrated in the WSI classification process.

Although these previous methods made use of both tile and slide labels, it was always in the form of a combination of (1) semi supervision at the tile level and then (2) weak supervision at the slide level. Therefore, the joint use of tile and slide labels has been limited in prior work with usually separate statistical models to exploit both types of labels and distinct training process to perform mixed supervision.

1.4. Contributions

In this work, we propose to generalize the use of mixed supervision to both tile- and slide-level classification tasks in a joint framework to train a model more suited for histological slide analysis, with higher performance and interpretability. For that matter, we rely on the CLAM architecture by Lu et al. (2021), as the model is able to operate at both slide and tile levels, allowing for tumor localization in addition to slide classification. Moreover, we make use of a limited amount of tile-labeled slides, in the hope of reducing the tedious work required from expert pathologists for the precise tumor delineation in histological datasets. Our contributions are listed as follows:

- To improve the performance of the tile-level classifier, it is trained on both true tile labels (when available) and pseudo-labels generated using the tiles' attention scores. This allows the classifier to leverage more training samples with accurate labels. To correct the potential class imbalance between tumorous and non-tumorous tiles, we also propose a paired batch method that uses both kinds of slides at the same time at each training step.

- To better target the key instances responsible for the bag label and obtain a model less focused on few specific instances, we design a new loss function based on the attention scores of the slide-level classifier. The loss also enforces a uniform spread of the attention on the relevant tiles in the slide, which improves the slide-level classification as well as the interpretability of the model during inference. We propose an exponential weighted sampling strategy, designed to simplify the training procedure in a single step, using both annotated and unannotated slides at the same time.

- We evaluate our method on two different histology datasets tasked with binary tumor classification and localization, and show that it indeed improves the consistency of the model between the tile- and the slide-level predictions, i.e., classification and localization. Throughout the rest of the paper, our model is coined MS-CLAM, for *Mixedly Supervised*-CLAM.

Compared to Tourniaire et al. (2021), we added loss functions that supervise the attention mechanism which directly impacts the slide-level classification, but also the interpretability of the model. Moreover, the training strategy has been simplified to fit in a single step thanks to the exponential weighted sampling strategy. This avoids to separate the slides between a subset where both slide-level and tile-level labels are available, and another set where only slide-level labels are known. Therefore, the method presented hereafter fully exploits mixed supervision in the context of MIL, for both bag classification and key instance detection.

2. Methods

2.1. CLAM

The CLAM framework (Lu et al., 2021) derives from the attention-based multiple instance framework introduced by Ilse et al. (2018), and adds what is referred to as an instance-level clustering task, where instances with highest and lowest attention scores are extracted from the slides, assigned positive and negative pseudo-labels (with respect to the target label), and then clustered by a label-specific instance clustering layer. As CLAM is designed to work on either binary or multiple class classification tasks, there can be multiple attention branches (CLAM MB), where each attention branch corresponds to a target label, or a single attention branch (CLAM SB). In the first case, each attention branch is followed by a slide-level binary classification layer, whereas in the latter, a single multi-class layer is used as the final output.

As multi-class classification is out of the scope of this article, we will focus on the single branch implementation of the algorithm, and show our contributions based on this model (Figure 1). Given we tackle the problem of tumor vs. non-tumor classification of WSIs, we refer to non-tumorous slides and tissue as "normal slides" and "normal tissue", regardless of the presence of artifacts in the tissue (tear, air bubble, etc...). In CLAM, the WSIs are first tiled, and each tile is converted into a 1024-dimensional vector by a modified frozen ResNet50 (He et al., 2016) (pre-trained on ImageNet). A slide is therefore represented as a feature matrix $\mathbf{z}$ containing N 1024-dimensional feature vectors. Each feature vector $\mathbf{z}_k$ is further reduced into a 512-dimensional vector $\mathbf{h}_k$ by a first layer $\mathbf{W}_1$. Then, attention scores are computed for each tile following:

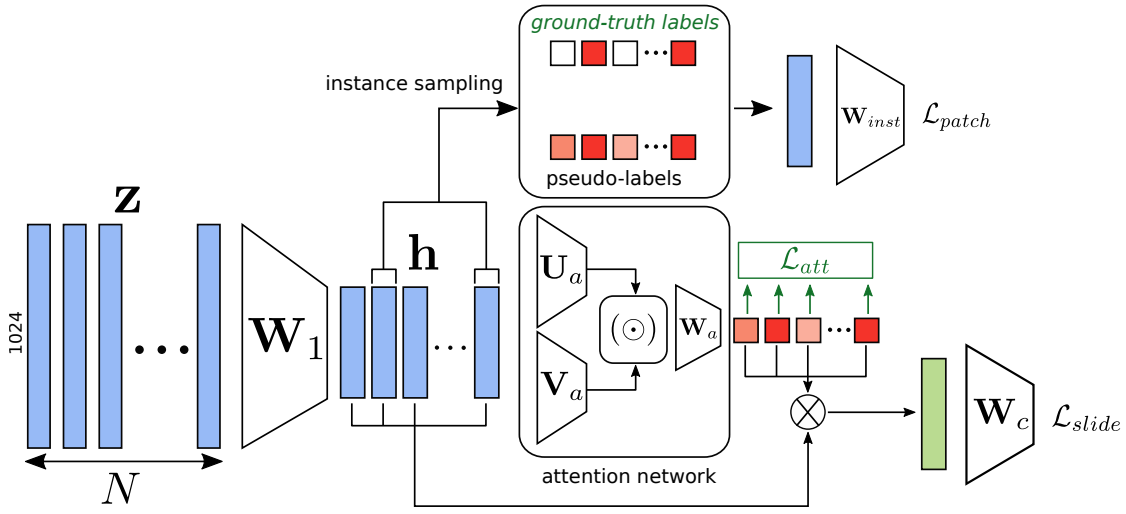

Fig. 1. Overview of the MS-CLAM model. Regarding the attention scores, a faint (resp. bright) color represents a low (resp. high) score. For the ground-truth colors, red means the instance is positive (with respect to the bag label), while no color means the instance is negative. The light-green rectangle represents the attention-weighted average of the feature vectors. Our contributions to the original CLAM architecture are printed in dark green.

$$a_k = \frac{\mathbf{W}_a(\tanh(\mathbf{V}_a\mathbf{h}_k^\top) \odot \mathrm{sigm}(\mathbf{U}_a\mathbf{h}_k^\top))}{\sum_{i=1}^{N} \exp\{\mathbf{W}_a(\tanh(\mathbf{V}_a\mathbf{h}_i^\top) \odot \mathrm{sigm}(\mathbf{U}_a\mathbf{h}_i^\top))\}} \quad (1)$$

where $\mathbf{U}_a$, $\mathbf{V}_a$, $\mathbf{W}_a$ are fully-connected layers. Given the softmax normalization, the attention scores sum to 1 for any number of tiles. The final classification layer $\mathbf{W}_c$ receives as an input the weighted sum of the slide instances $\mathrm{A}^\top\mathbf{h}$, where the weights are the attention scores. A cross-entropy loss $\mathcal{L}_{slide}$ is eventually computed on the output probability of the classifier.

For the instance-level clustering, a fully-connected layer $\mathbf{W}_{inst}$ performs a binary classification of the sampled instances. This classification is evaluated by a top-1 SVM loss (Lapin et al., 2016) $\mathcal{L}_{patch}$, as this loss allows for a higher noise tolerance over input data, which is desirable since the instances at this stage are only pseudo-labeled. In the original CLAM model, there are as many instance classifiers as there are classes.

2.2. *Instance-level classification supervision*

In CLAM, under the weak supervision setting, tile labels are unknown: based on the attention scores, pseudo labels are generated. To do so, the B tiles with the highest (resp. lowest) attention scores are labeled as positive (resp. negative) with respect to the slide-level label. For each of the latter, a binary tile classifier is trained using the sampled tiles and their corresponding pseudo-labels. This approach unfortunately has its share of flaws: first, the parameter B is fixed, meaning that tiles are invariably sampled within slides, regardless of the actual number of tiles representative of the slide-level class. Therefore, only small values of B can help avoid sampling the wrong tiles, but in turn limit the number of training samples. Another issue with this approach is that in every slide, regardless of its label, B tiles are labeled as negative evidences of the slide class. This relies on the assumption that a bag contains both positive and negative evidences of its class: however, based on the binary MIL assumptions, we know that all instances from a normal slide are normal. In normal slides, all tiles are evidences of the normal class, which we know directly from the slide-level label. Therefore, CLAM's assumption for pseudo-labeling is wrong for this particular case.

With MS-CLAM, we propose to solve some of these issues with the help of mixed supervision and the paired batch method. First, in the context of binary tumor classification, we propose to use a single instance (or tile) classifier, as two classifiers performing opposite tasks would be redundant. Second, for each slide with available tile-level labels, the instances are sampled and assigned their true label instead of the pseudo-generated one (see Figure 1, top). This not only allows us to train the tile-level classifier without potentially erroneous labels, it also helps sampling more tiles within the slides, since for all of them the label is accessible. To distinguish between the cases where tile labels are known or not, we use two different hyperparameters to sample the instances: B_+ when labels are available, and B_- when they are not ($B_+ > B_-$). For the case where B_+ might be greater than the actual number of tumorous tiles N_{tum} in the slide, we set $B_+ = N_{tum}$. Third, since in normal slides all tiles are normal (thanks to the MIL assumption), we use a different tile-sampling strategy, as the original method assigned wrong labels to the tiles with low attention scores. In our case, in normal slides, sampled tiles are only assigned the same label as the slide. Moreover, if we sample B tumorous tiles in tumorous slides then we only sample $B/2$ normal tiles in both tumorous and normal slides, to improve the balance between the two classes. A summary of the two tile labeling approaches is shown in Figure 2.

To train the original CLAM model, a single slide is sampled at each step, and a tile batch containing $2B$ instances is generated. With the modifications we made to the tile labeling, this approach is no longer recommended, because for normal slides, where only a single class is represented among the tiles, this would mean that the tile batch would correspond to a single label. Alternating between tumorous slides – with both labels represented at the tile level, and normal slides – with only a single label present in the tiles – could lead to unstable gradient

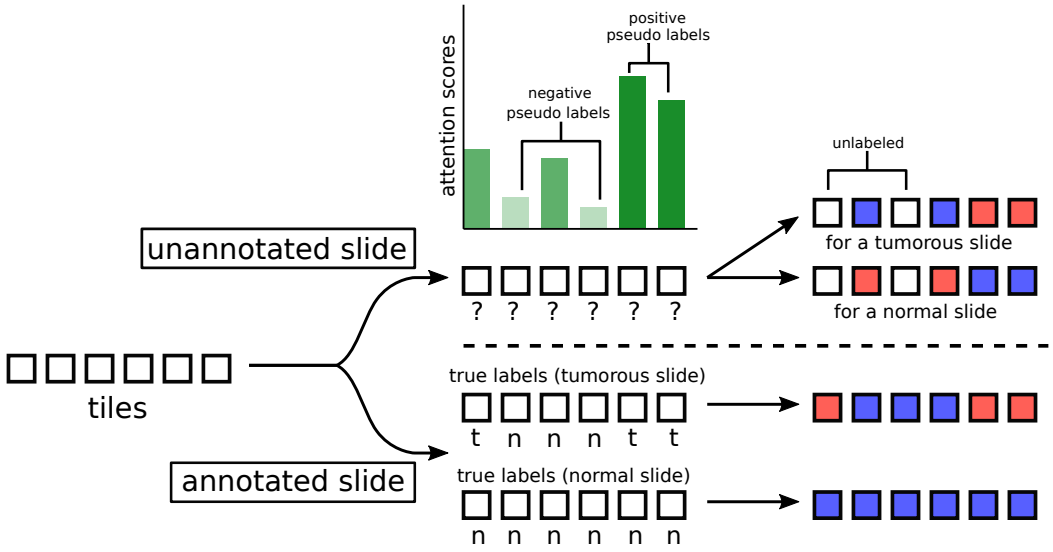

Fig. 2. The two methods for labeling tiles (represented as colored squares) in WSIs: the top part corresponds to the weakly-supervised case (already used in CLAM), where attention scores are used to generate pseudo-labels for the tiles. The bottom part on the other hand corresponds to the case where the tile labels are available (only for MS-CLAM). The number of sampled tiles in the weakly-supervised setting here is $B = 2$. Red (resp. blue) squares represent tumorous (resp. normal) tiles.

computation for the tile classifier. To circumvent this issue, we propose to simultaneously process one tumorous and one normal slide (which we refer to as *Double Sampling*), and build a tile batch using instances from both slides: we call this process the paired batch method. Figure 3 represents how the batch of tiles is produced.

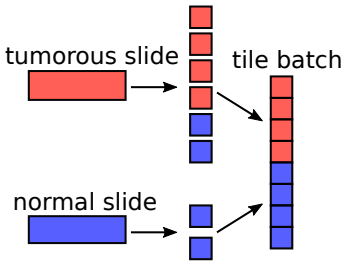

Fig. 3. The paired batch creation process. The tumorous slide provides B tumorous and $B/2$ normal tiles, while the normal slide provides $B/2$ normal tiles to make a $2B$-sized tile batch ($B = 4$ in the figure).

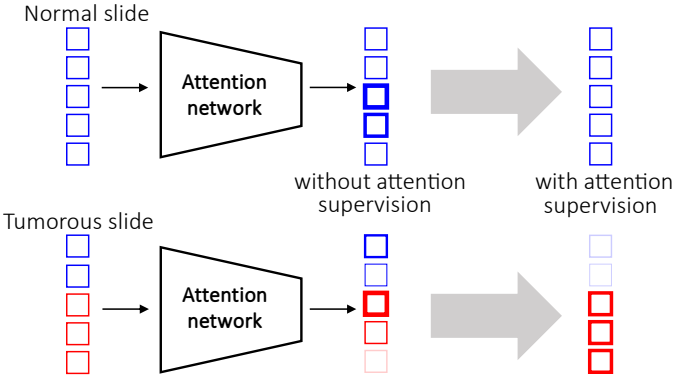

Fig. 4. The goal of the supervision of the attention scores. Instances are represented as colored squares. The red color (resp. blue) represents instances with tumor (resp. without). A thick, bright square means the instance was given a high attention score, whereas a thin, faint square means the instance was given a low attention score. In normal slides, attention scores should be even, so as to weight each instance equally. In tumorous slides, tumorous patches' attention scores should be higher than the non-tumorous ones, but equally weighted between them. The attention loss is designed to guide the attention on the most relevant patches.

2.3. Attention Loss

Until now, the mixed supervision was only designed for the instance-level classification, with a collateral impact on the slide-level classification. In CLAM, no constraint is applied on the attention scores, except that their sum must be equal to 1 (to be invariant to the bag size). In the weakly-supervised setting, we noticed that the attention scores associated to the patches were highly unbalanced, with only a few instances weighted much higher than the rest, whatever the slide label be. Although this effect is rather undetectable on small bag sizes, e.g., a few hundreds of instances, when facing bags with tens of thousands of instances, which is typical of WSIs, the attention tends to be focused unevenly on a few instances only, or even sometimes on a single one. Here, we propose a new loss term based on the attention scores to orient the attention spread towards the most important patches, but also to equalize the attention among them (Figure 1, bottom). Figure 4 describes the purpose of the supervision. Oddly enough, the imbalance between attention scores is noticeable in both normal and tumorous slides. Ideally, we wish the attention scores to be distributed differently depending on the slide label. For normal slides, we want the attention scores to be all equal, i.e., $\forall (i, j) \in \{1, ..., N\}^2, a_i = a_j = 1/N$. Seeing the attention scores as a probability distribution, this condition can be expressed through the means of the Shannon entropy (Shannon, 1948). The entropy reaches a maximum when all of the outcomes are equally likely, and the maximum value is $\log N$ given there are

N possible outcomes. For normal slides, the attention loss is therefore written as follows:

$$\mathcal{L}_{att} = \frac{1}{\log N} \sum_{k=1}^{N} a_k \log a_k \tag{2}$$

The entropy of the attention scores distribution should be maximum, meaning each instance is given equal importance in the weighted sum before the final classification layer. In other words, the attention pooling should mimic the mean pooling in the case of a normal slide. We take as a penalty term the negative entropy, normalized by $\log N$ to account for any number of instances within a WSI. This loss term does not require the availability of tile-level labels, since all tiles have the same label, which is the one of the slide. Such a regularization term on attention scores was also proposed by Lu et al. (2019) and Sharma et al. (2021), in the form of a KL-divergence with respect to a uniform distribution, which is equivalent to the entropy up to a constant. For tumorous slides on the other hand, the expression involves three terms, because we want to reach the three following objectives:

1. The attention scores of non-tumorous instances should be close to zero, as they have little to no impact on the slide label.
2. From the previous condition, we have that the tumorous instances' attention scores should be the only non-zero ones. Put differently, the sum of the attention scores of tumorous instances should be close to 1.
3. The entropy of the tumorous attention scores should be maximum, i.e., each tumorous instance should be weighted equally before the final classification. This is to ensure that all instances containing tumor contribute as equally as possible to the final prediction.

Therefore, the attention loss for tumorous slides is expressed as follows:

$$\mathcal{L}_{att} = \sum_{i=n_1}^{n_s} a_i + \frac{1}{\log m} \sum_{j=t_1}^{t_m} a_j \log a_j - \sum_{j=t_1}^{t_m} a_j \tag{3}$$

where m (resp. s) is the number of tumorous (resp. non-tumorous) instances, $t_1, ..., t_m$ are the indices of the tumorous attention scores, and $n_1, ..., n_s$ the indices of the non-tumorous ones. Contrary to the previous case, this loss term requires the knowledge of tile-level labels, hence why the mixed supervision is used during training. Table 1 summarizes the various losses computation depending on the slide's label.

2.4. Exponential Weighted Sampling

In Tourniaire et al. (2021), a two-step training procedure was designed to train the model where first only slides with tile-level labels were used, and then the entire training set in a second phase. This was done to ensure that the tile-level classifier was first trained on true labels, before being trained on pseudo-labels. In this paper, we introduce a sampling strategy devised to train the model in a single phase, where slides with tile-level labels are decreasingly more likely to be sampled during training, until all slides are sampled uniformly. Assuming we have $N_t = \hat{N}_t + \tilde{N}_t$ tumorous slides, where $\hat{N}_t$ is the number of tile-level labeled slides and $\tilde{N}_t$ the number of slides with unlabeled tiles: each slide i is assigned a sampling weight w_i, equal to a value $W > 1$ when the tile-level labels are known, or 1 otherwise. Then, these weights are converted into probabilities following $p_i = w_i / \sum_j w_j$. These probabilities serve as the parameters of a multinomial distribution used to sample the slides at each epoch. To progressively reduce the oversampling of annotated slides, their corresponding weights w_k are multiplied by a decay factor $\gamma < 1$ at the end of each epoch, until all slides' weights are equal to 1, resulting in equal sampling probabilities. Algorithm 1 summarizes the sampling strategy.

Algorithm 1: Tumorous slides' sampling strategy.

```
Data: Initial weight W, decay factor γ, number of
      training epochs E, the set of tumorous slides
      {S_1, ..., S_{N_t}}
for e ← {1, ..., E} do
    for i ← {1, ..., N_t} do
        if S_i has tile-level labels then
            w_i ← W
        else
            w_i ← 1
        end
        p_i ← w_i / Σ_i w_i
    end
    sample slides from Multinomial(p_1, ..., p_{N_t})
    W ← γW
end
```

2.5. MS-CLAM without tile-level labels

The absence of tile-level annotated slides (weak supervision only) can be seen as a particular case, which requires several adjustments to the method. Concerning the attention loss, we can only use Eq. 2, since Eq. 3 requires the knowledge of tile-level labels. Still, the paired batch method remains applicable, along with the other modifications we made to the model. Concerning the exponential weighted sampling, we fix $W = 1$ and $\gamma = 1$ so that all tumorous slides are sampled randomly (equal weights and no decay). This setting corresponds to the Double Sampling strategy we mention in section 2.2.

3. Materials

3.1. The Camelyon16 dataset

The Camelyon16 challenge (Bejnordi et al., 2017), was proposed to classify lymph node slides and detect metastatic regions within them, with a dataset of 399 WSIs from two different centers, partitioned between a test set (129 WSIs) and a training/validation set (270 WSIs). All metastatic slides have been precisely and exhaustively annotated by a panel of expert pathologists, except for 20 slides which were only partially annotated. The class distribution of the dataset is detailed in Table 2. This dataset is particularly challenging among histological datasets, as the metastasis size from one slide to the other

Table 1. Summary of the losses for each kind of slide label. The table also indicates how the losses are handled depending on the tile labels availability. The H function stands for the Shannon entropy, and A represents the vector of all attention scores. Similarly, A_t is the vector of tumorous tiles' attention scores, and A_n is the vector of normal tiles' attention scores. CE stands for cross-entropy.

Slide label	Inst. label availability	Inst. label	Att. loss	Inst. label nature	Inst. loss	Number of inst. in batch
Normal	yes	Normal	$-H(A)$	true label	CE	$B_+/2$ or $B_-/2$
Tumor	no	N/A	N/A	pseudo-label	CE	B_- (tum.) and $B_-/2$ (norm.)
	yes	Normal	$\|\|A_n\|\|_1$	true label	CE	$B_+/2$
		Tumor	$-H(A_t) - \|\|A_t\|\|_1$	true label		B_+

Table 2. Summary of the Camelyon16 dataset class distribution.

Slide class	Train set	Test set
Normal	159	80
Metastatic	111	49

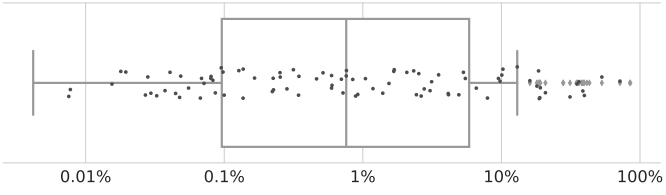

Fig. 5. A box plot showing the percentage of tumorous tiles (log-scaled) in tumorous slides in the training set of Camelyon16. The grey, diamond-shaped points represent the outliers, while the black, circular points correspond to the data points themselves.

greatly varies. From a MIL standpoint, this means that the number of positive instances per bag can differ significantly from one bag to the other, to the point where there can be only a few positive instances for tens of thousands of negative ones in one bag, while in the other there are nearly no negative instances. This variability is expressed in the form of a box plot in Figure 5 (notice the log scale on the horizontal axis). The mean percentage of tumorous tiles within tumorous slides is 6.3%, while the median percentage is only 0.76%. Therefore, we expect the attention loss to help the model coping with this variability among the positive bags. During the experiments, we split the training/validation set 5 times into a training (80%) and a validation (20%) set in a 5-fold cross-validation fashion, and report the average performance of the model on the competition test set

3.2. The DigestPath2019 dataset

The DigestPath2019 challenge (Li et al., 2019) was organized around two different gastric and colon histology datasets. In this work, we focused on the second dataset, for which the challenge task was to classify and segment tissue in colonoscopy images. It does not contain entire WSIs but regions selected within these colonoscopy slides. The resulting images have an average size of 5000x5000 pixels. As the competition test set is unavailable for download, we performed all of our experiments on the competition training set, which contains 660 images in total (from 324 patients, coming from 4 different centers), of which 250 (from 93 patients) display tumor regions. We perform a 5-fold cross-validation of the models on the competition training set (i.e., the available 660 images). For each fold, the training set is again divided between 80% training and 20% validation.

The challenge website mentions that some malignant glands were missed by pathologists, so the annotations are not exhaustive per se, but are considered as such during the experiments. As opposed to the ones in the Camelyon16 dataset, tumorous slides in DigestPath2019 display a much wider tumorous area with respect to the total tissue area: the mean percentage of tumorous tiles in tumorous images in this case is 31.8%, with a median value at 28.9%. The boxplot Figure 6 summarizes the ratio of tumorous tiles within tumorous images for DigestPath2019.

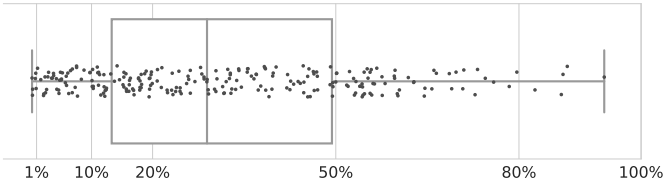

Fig. 6. A box plot showing the percentage of tumorous tiles in tumorous images in the training set of DigestPath2019.

3.3. Data pre-processing

For both datasets, we followed the usual procedure for histological image analysis: first, the tissue region is filtered in the images using a threshold on the saturation channel in the HSV color space. For some images in the DigestPath2019 dataset, blurry regions were filtered out using a blur detector (Golestaneh and Karam, 2017). Then, the images are split into squared tiles of dimensions 256x256 (for Camelyon16), 128x128 (for DigestPath2019). For DigestPath2019, the choice of a proper tile size is a more critical choice than for Camelyon16, since the original images have much smaller sizes (they are not original WSI files). Therefore, smaller tiles allow for a more accurate localization of the tumor within the images. However, since the feature extractor was initially trained with inputs of 224x224 pixels, using an input size too different from the initial dimensions would result in a performance decay. To assign a label to the tiles, we used a different approach for the two datasets since they both display very different tumorous area ratios as shown in Figures 5 and 6. For Camelyon16, we looked at the slide with the smallest tumorous region, centered

a tile around this region, and computed the percentage of tumor inside the tile to obtain a threshold: the value we obtained was 20%. On DigestPath2019, as tumor regions were quite wide and similar from one slide to the other, we stuck to a 50% threshold. We used the same Imagenet pre-trained, Resnet-50 backbone as Lu et al. (2021) to pre-extract the features from the tiles before training the attention layers.

As all images from each dataset were annotated, we randomly selected k% of the slides to be used with tile-level labels, with $k \in \{0, 6, 12, 25, 62, 100\}$, to evaluate the model's performance with an increasing percentage of available annotations.

3.4. Experimental setting

For all the experiments, we used the Adam optimizer (Kingma and Ba, 2014) during training, with the default values $\beta_1 = 0.9$ and $\beta_2 = 0.99$, with a learning rate of 2×10^{-4}. We took $B_+ = 1024$ for Camelyon16, $B_+ = 256$ for DigestPath2019, and $B_- = 8$ for the tile sampling parameters. For the exponential weighted sampling, we kept $W = 90$ and $\gamma = 0.9$. All models were trained for either 50 epochs (DigestPath2019) or 90 epochs (Camelyon16) on a single NVIDIA GeForce 1080 GTX Ti GPU. A 5-fold training on Camelyon16 takes approximately 5 hours, while on DigestPath2019 it takes only 2 hours 30 minutes.

4. Results

To evaluate the results of the experiments, we use several classification and localization metrics. Although the model does not directly produce tumor masks, we use the outputs of the tile classifier to compute binary tile masks of the slides, which are then compared to the reference tumor masks using localization metrics. For the slide-level classification, we mainly look at the accuracy, the F1-score, and the AUC (Area Under the ROC Curve, which was the reference metric for both datasets). For both datasets, to evaluate the quality of the tumor masks, we first compute the reference tile-accurate masks based on the tumor delineations done by experts and using the assumptions made in 3.3 as to which tiles are considered tumorous. This is done to allow for a fairer comparison between the predicted and the reference masks, since none of the models are pixel-accurate. To compute the masks using the tile-level predictions, we use a threshold of 0.5 on the output probability of the tile classifier. For CLAM SB or MB, as two different tile-level classifiers coexist, the one corresponding to the tumor class is used. On Camelyon16, all masks are computed at the 5^{th} magnification level, the same magnification level used during the challenge for localization evaluation. Predicted masks are evaluated using the Dice score on tumorous slides. On normal slides on the other hand, we simply compute the tile-level specificity of each model.

4.1. Baselines

Aside from CLAM, we also compare the performance of our model with several baselines:

- **Weakly-supervised baselines.** We compare our model with other weakly-supervised models such as TransMIL (Shao et al., 2021) and DS-MIL (Li et al., 2021b). For the latter, an important part of the method is the training of a feature extractor in a self-supervised fashion. According to the authors, it took 2 weeks on 6 GPUs to complete this part for Camelyon16 only; therefore, we decided to use the pre-computed features they provide on github for this dataset in particular. We also show the performance of DS-MIL when using the same feature extractor as for CLAM, MS-CLAM and TransMIL (i.e., the ImageNet pre-trained ResNet-50). However, since neither TransMIL nor DS-MIL have a dedicated tile-level classifier, we only compare the performances of these models on the slide-level classification task, but not on the tumor localization one.

- **Backbone fine-tuning.** We investigate the potential benefit of fine-tuning the feature extractor on the available tile labels before training CLAM and MS-CLAM. To this end, for each fold, we fine-tune the ImageNet pre-trained ResNet-50 backbone on tiles taken from the same slides used to supervise the training of MS-CLAM. The fine-tuning can be seen as supervised tile classification, after which we discard the final classification layer to recover the features. We use the Adam optimizer and cross-entropy loss, with a learning rate of 1×10^{-3}, divided by ten every time the loss plateaus for 15 epochs. The model is trained during 200 epochs for each fold, or until the loss stops decreasing during 20 epochs. We indicate *(FT)* in the tables next to models trained atop a fine-tuned backbone.

4.2. Slide-level classification

Table 3 shows the results of the image classification for DigestPath2019. Without the use of any tile-level annotation, the addition of the attention loss on normal slides along with the paired batch method lead to an improvement of all the classification metrics, in particular the accuracy (improved by 1.7% compared to CLAM SB, 3.5% compared to CLAM MB) and the F1-score (improved by 2% compared to CLAM SB, 4.2% compared to CLAM MB). With an annotation burden as low as 12% of the available slides in the training set, we also notice an improvement of all the classification metrics for the MS-CLAM model compared to the CLAM baseline. While the gain in AUC is rather marginal, there is a gradual improvement regarding the accuracy and the F1 score, meaning the model is less prone to classification errors.

Among the weakly-supervised baselines, TransMIL reaches the highest performance, on par with MS-CLAM using 6% of the annotated slides. However, the model does not provide tumor localizations, therefore the slide-level predictions suffer from a lack of interpretability. For DS-MIL, without access to a dedicated backbone for this dataset, the model obtains poor performance compared to all the other methods.

Finally, when training CLAM on top of a fine-tuned feature extractor, its performance improves as the number of annotated slides increases. Nonetheless, when using only 12%, it still does not reach the performance of MS-CLAM (without FT) with the same amount of annotations in terms of accuracy

Table 3. Classification metrics over a 5-fold CV of the DigestPath2019 training set (± standard error is indicated for each experiment and metric).

Model	% of annot. images	AUC (↑)	Acc. (↑)	F1-score (↑)
CLAM SB	0	0.973 ± 0.022	0.924 ± 0.039	0.901 ± 0.052
CLAM MB	0	0.953 ± 0.043	0.906 ± 0.048	0.879 ± 0.063
TransMIL	0	0.982 ± 0.015	0.932 ± 0.05	0.927 ± 0.053
DS-MIL (ImageNet)	0	0.615 ± 0.091	0.511 ± 0.138	0.212 ± 0.291
MS-CLAM	0	0.977 ± 0.014	0.941 ± 0.022	0.921 ± 0.026
MS-CLAM	6	0.982 ± 0.012	0.941 ± 0.026	0.922 ± 0.029
MS-CLAM	12	0.982 ± 0.013	**0.955 ± 0.034**	**0.940 ± 0.042**
MS-CLAM	25	0.980 ± 0.016	0.944 ± 0.038	0.927 ± 0.047
MS-CLAM	62	**0.984 ± 0.016**	0.947 ± 0.036	0.933 ± 0.041
MS-CLAM	100	0.981 ± 0.019	0.946 ± 0.035	0.931 ± 0.041
CLAM SB (FT)	12	0.987 ± 0.005	0.945 ± 0.014	0.928 ± 0.020
MS-CLAM (FT)	12	0.989 ± 0.009	0.953 ± 0.024	0.940 ± 0.027
CLAM SB (FT)	62	0.983 ± 0.013	0.953 ± 0.017	0.938 ± 0.022
MS-CLAM (FT)	62	**0.991 ± 0.008**	0.955 ± 0.029	0.941 ± 0.037
CLAM SB (FT)	100	0.990 ± 0.006	**0.964 ± 0.023**	**0.954 ± 0.025**
MS-CLAM (FT)	100	**0.991 ± 0.007**	0.950 ± 0.027	0.936 ± 0.033

and F1-score. Notwithstanding the case with 100% annotated slides, MS-CLAM (FT) still manages to improve the performance of CLAM (FT).

Classification results for Camelyon16 are detailed in Table 4. For the MS-CLAM models, there is again an improvement of all the classification metrics compared to CLAM when using 12% of annotated slides or more, which corresponds to only 11 tumorous slides in the Camelyon16 training set. We see similar effects to the ones observed on DigestPath2019: accuracy and F1-score largely profit from the added supervision (F1-score improved by 5% with only 6% annotated slides).

Just like for DigestPath2019, TransMIL is comparable to MS-CLAM trained with 6% annotated slides (with a 2% overhead in AUC), yet still without localization information. For DS-MIL, this time, we have access to a specifically trained backbone. The self-supervised version of the model reaches the highest AUC, but when looking at the accuracy and the F1-score, it falls behind TransMIL or MS-CLAM with 6% of annotated slides. Similar to what we observed on DigestPath2019, DS-MIL achieves poor results when using the ImageNet pretrained backbone.

The results obtained by the FT models on Camelyon16 are similar to the ones obtained on DigestPath2019, although this time MS-CLAM (FT) is still superior to its CLAM (FT) counterpart when using 100% of the annotated samples. Interestingly enough, for the 12% and the 62% settings, CLAM (FT) and MS-CLAM (without FT) reach similar accuracy and F1-score, although CLAM (FT) obtains a higher AUC. Nonetheless, the computational burden is much lower for MS-CLAM, as no backbone fine-tuning is required.

4.3. Localization of tumor regions

Localization results (i.e., the tumor masks derived from the tile classifier) for DigestPath2019 (resp. Camelyon16) are detailed in Table 5 (resp. Table 6). For both datasets, with the exception of MS-CLAM with 0 and 6% annotations on DigestPath2019, we obtain both a higher mean Dice score on tumorous images, and a higher mean specificity for normal images, close to 1, indicating very few false positives. In both cases, the baseline models lack specificity and fail to point at the tumor region accurately, triggering many false positives in the case of normal images. Yet, for DigestPath2019 in particular, the tumor region in tumorous slides spans a relatively wide area (see Figure 6), sometimes covering nearly all of the tissue. Therefore, the Dice score for CLAM is fairly high (0.520) . Conversely, MS-CLAM models tend to have a much higher specificity, at the cost of a lower sensitivity, which gradually improves with the percentage of tile-level labeled slides. This, combined with the previous observation on the ratio between tumor and healthy tissue in the case of DigestPath2019, explains why the Dice score of MS-CLAM with 0 and 6% annotations is lower for this dataset in particular compared to CLAM SB. It is in the images that contain smaller tumor regions regarding the total tissue area that the difference is perceptible. An example of the latter case is visible Figure 7, where the CLAM model has classified all tissue as tumorous, whereas it is in fact limited to sub-parts of the image. With MS-CLAM on the other hand, the tumorous tissue is correctly located within all the annotated regions. Furthermore, this image is a likely example of incomplete annotations, as several small unmarked regions in the image, in particular at the bottom right, are likely to be tumorous. Nonetheless, these regions are still correctly picked by the models, as shown on the two right-most masks in Figure 7. We can also notice that adding more supervision in MS-CLAM improves the recall of tumorous instances, which is expected since the model trains on more tumorous tile samples. In terms of localization, although fine-tuning the feature extractor improves the performance of CLAM in terms of both Dice score and specificity, the latter still remains far below what is achieved with MS-CLAM, regardless of the backbone used. When using more annotated samples (62% onward), the performance of the model improves with fine-tuning.

On Camelyon16, contrary to what was observed on DigestPath2019, all of the MS-CLAM models outperform CLAM re-

Table 4. Classification metrics on the Camelyon16 test set. All metrics are averaged on a 5-fold cross validation split of the training set (± a standard error reported).

Model	% of annot. images	AUC (↑)	Acc. (↑)	F1-score (↑)
CLAM SB	0	0.883 ± 0.033	0.863 ± 0.027	0.797 ± 0.049
CLAM MB	0	0.907 ± 0.010	0.870 ± 0.028	0.806 ± 0.049
TransMIL	0	0.910 ± 0.021	0.874 ± 0.030	0.857 ± 0.039
DS-MIL (SS)	0	**0.966 ± 0.021**	0.879 ± 0.066	0.849 ± 0.065
DS-MIL (ImageNet)	0	0.467 ± 0.185	0.478 ± 0.130	0.331 ± 0.302
MS-CLAM	0	0.884 ± 0.020	0.888 ± 0.026	0.830 ± 0.044
MS-CLAM	6	0.889 ± 0.017	0.898 ± 0.026	0.859 ± 0.022
MS-CLAM	12	0.908 ± 0.013	0.899 ± 0.028	0.861 ± 0.031
MS-CLAM	25	0.911 ± 0.016	0.902 ± 0.028	0.867 ± 0.035
MS-CLAM	62	0.932 ± 0.008	0.938 ± 0.009	0.913 ± 0.013
MS-CLAM	100	0.939 ± 0.008	**0.938 ± 0.012**	**0.916 ± 0.017**
CLAM SB (FT)	12	0.932 ± 0.029	0.898 ± 0.030	0.860 ± 0.039
MS-CLAM (FT)	12	0.921 ± 0.036	0.910 ± 0.018	0.877 ± 0.029
CLAM SB (FT)	62	0.967 ± 0.011	0.936 ± 0.025	0.915 ± 0.033
MS-CLAM (FT)	62	0.970 ± 0.012	0.935 ± 0.015	0.916 ± 0.018
CLAM SB (FT)	100	0.980 ± 0.008	0.946 ± 0.016	0.928 ± 0.020
MS-CLAM (FT)	100	**0.982 ± 0.013**	**0.950 ± 0.018**	**0.935 ± 0.022**

Table 5. Localization metrics on DigestPath2019. All metrics are averaged on a 5-fold cross validation split of the training set (± a standard error reported).

Model	% of annot. slides	Dice score (tum)	Specificity (norm)
CLAM SB	0	0.520 ± 0.075	0.525 ± 0.075
CLAM MB	0	0.443 ± 0.087	0.443 ± 0.101
MS-CLAM	0	0.310 ± 0.051	**0.998 ± 0.001**
MS-CLAM	6	0.460 ± 0.073	0.998 ± 0.002
MS-CLAM	12	0.530 ± 0.063	0.997 ± 0.003
MS-CLAM	25	0.595 ± 0.061	0.993 ± 0.005
MS-CLAM	62	**0.677 ± 0.026**	0.978 ± 0.009
MS-CLAM	100	0.676 ± 0.027	0.960 ± 0.016
CLAM SB (FT)	12	0.598 ± 0.069	0.695 ± 0.233
MS-CLAM (FT)	12	0.453 ± 0.060	**0.997 ± 0.001**
CLAM SB (FT)	62	0.582 ± 0.123	0.590 ± 0.306
MS-CLAM (FT)	62	**0.715 ± 0.027**	0.976 ± 0.011
CLAM SB (FT)	100	0.596 ± 0.094	0.609 ± 0.289
MS-CLAM (FT)	100	0.714 ± 0.014	0.967 ± 0.014

Table 6. Localization metrics on the Camelyon16 test set. All metrics are averaged on a 5-fold cross validation split of the training set (± a standard error reported).

Model	% of annot. slides	Dice score (tum)	Specificity (norm)
CLAM SB	0	0.212 ± 0.005	0.740 ± 0.034
CLAM MB	0	0.223 ± 0.031	0.755 ± 0.029
MS-CLAM	0	0.331 ± 0.015	**1.000 ± 0.000**
MS-CLAM	6	0.425 ± 0.052	**1.000 ± 0.000**
MS-CLAM	12	0.473 ± 0.023	**1.000 ± 0.000**
MS-CLAM	25	0.503 ± 0.039	0.999 ± 0.001
MS-CLAM	62	**0.513 ± 0.029**	0.996 ± 0.002
MS-CLAM	100	0.475 ± 0.023	0.991 ± 0.003
CLAM SB (FT)	12	0.210 ± 0.024	0.691 ± 0.125
MS-CLAM (FT)	12	0.425 ± 0.043	**1.000 ± 0.000**
CLAM SB (FT)	62	0.287 ± 0.031	0.872 ± 0.041
MS-CLAM (FT)	62	**0.533 ± 0.033**	0.998 ± 0.001
CLAM SB (FT)	100	0.270 ± 0.032	0.840 ± 0.082
MS-CLAM (FT)	100	0.442 ± 0.048	0.984 ± 0.008

gardless of the percentage of annotated slides used, in terms of both Dice score and specificity. Again, part of the explanation lies in the ratio between tumor and healthy tissue in tumorous slides: with many false positives, CLAM models tend to overestimate far more the tumorous regions than in the previous dataset. One noticeable result in Table 6 is the slight decrease of specificity and Dice score for the MS-CLAM models with 100% of annotated slides. This is because this model has a much higher recall at the cost of more false positives, and is therefore penalized by the relatively small tumorous regions in Camelyon16. However, it offers a much higher tile-level AUC and Average Precision (AP) than its counterpart with few tile-level labels (MS-CLAM with 100% of annotated slides reaches a mean AUC of 0.950 and a mean AP of 0.763, against an AUC of 0.736 and an AP of 0.429 for MS-CLAM with 6%). In the same way, although it seems like MS-CLAM with 62% of annotated slides performs better than the one with 100%, these two models have in fact very close performance. With 62%, the mean tile-level AUC is 0.948, but the mean recall is 0.551 (against 0.605 with 100%). Given the small number of tumorous tiles per slide in Camelyon16, false positives are more hurtful to the Dice score than false negatives. Figure 8 shows an example of a tumorous slide from Camelyon16, where the mask computed using the weakly-supervised model lacks specificity, which is higher for the MS-CLAM models, while recall increases with the annotation percentage. When dealing with minute tumorous regions, which is the case for the slides in Camelyon16, a high specificity is essential to accurately pick the tumor region: with many false positives, the inspection of the slide becomes tedious, whereas a high specificity guarantees that the user can rapidly check the regions raised suspicious by the model. The same observations made on DigestPath2019 regarding the effect of backbone fine-tuning on the localization performance can be made on Camelyon16, although this time MS-CLAM (FT) systematically obtains higher Dice score and specificity compared to CLAM (FT). When com-

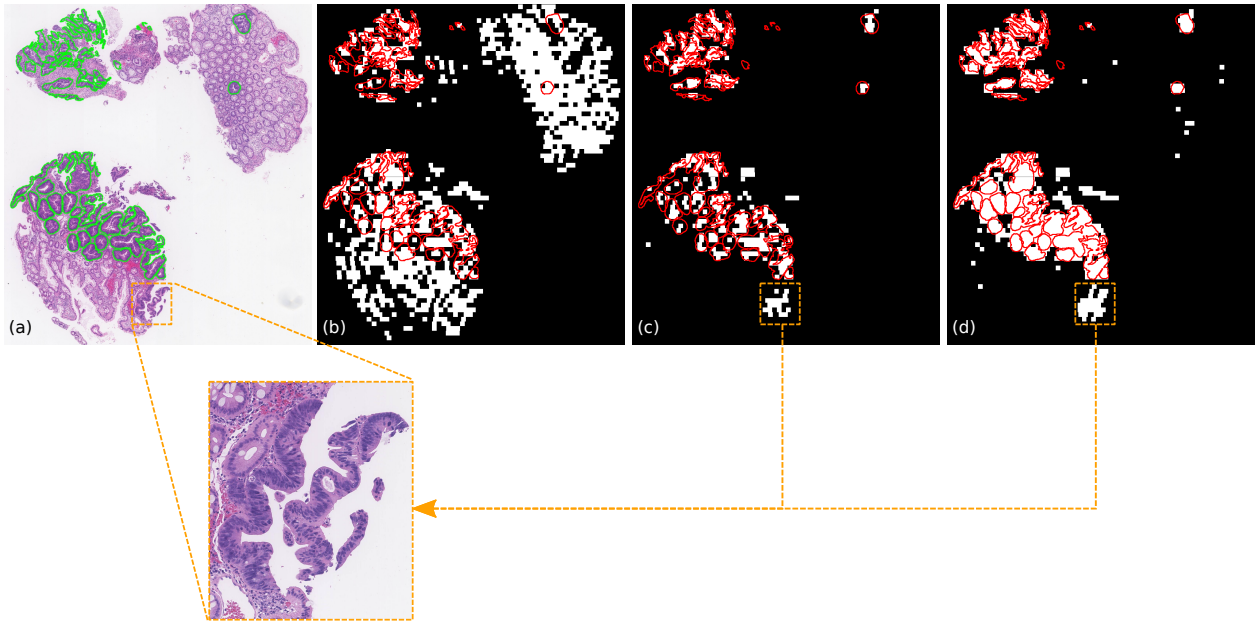

Fig. 7. Examples of tumor masks obtained on a tumorous image from the DigestPath2019 dataset. (a) The slide region with the tumorous tissue delineated in green on the left image, and in red on the four binary masks. (b)-(d) The tile-level masks computed by the models' tile-level classifier with various amounts of supervision. (b) CLAM SB (0%). (c) MS-CLAM (6%). (d) MS-CLAM (62%). The orange square contains a tissue region, likely tumorous, that is not delineated in the ground truth annotations.

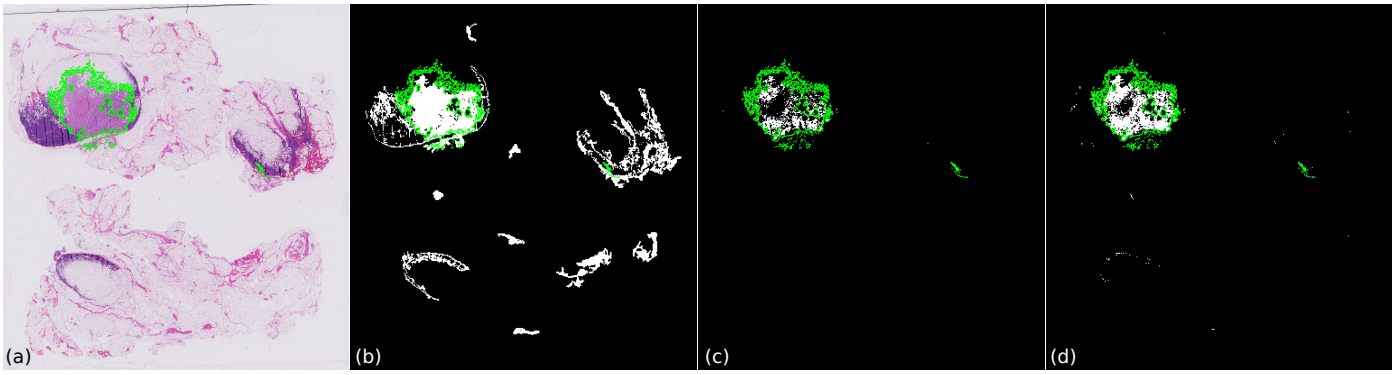

Fig. 8. Slide #26 from the test set of Camelyon16, along with the tile-level tumor mask computed by each model using the tile-level classifier. (a) The slide thumbnail (metastasis delineated in green). (b)-(d) The tile-level masks computed by the models with various amounts of supervision. (b) CLAM SB (0%). (c) MS-CLAM (6%). (d) MS-CLAM (62%).

paring MS-CLAM with and without fine-tuning, it is only in the case of 62% annotated samples that we obtain a performance gain, while all the other amounts of annotation are negatively affected by the fine-tuning procedure. Fine-tuning alone seems insufficient to increase the localization performance of the model, whereas our framework does bring significant improvements.

5. Ablation studies

In this section, we show the contributions of each main module of our model to the global performance, both in terms of WSI classification and localization. We also highlight the impact of the attention loss on the attention scores and the visualized attention maps. All ablation experiments are performed on the Camelyon16 dataset.

5.1. Attention loss

Slide-level classification. Table 7 shows the impact of the attention loss on the slide classification task. Our attention loss improves the classification results from 25% annotated slides onward, although the performance is only slightly superior (< 0.5% difference) without the loss for 12% annotated slides. Regarding the impact on the attention scores: for normal slides, an attention loss $\mathcal{L}_{att}$ (norm) close to -1 means that the attention scores are nearly all equal. In this case, the model was able to give equal importance to each tile instead of focusing on just a few. For tumorous slides on the other hand, $\mathcal{L}_{att}$ (tum) has two main effects: first, the attention scores of tumorous tiles are higher than the ones of normal tiles; second, the attention scores of tumorous tiles have similar values, which means the tiles contribute equally to the final attention-based mean of the features. These observations regarding the attention scores are

Table 7. Impact of the attention loss on the slide-level classification performance. All metrics are averaged on a 5-fold cross validation split of the training set (± a standard error reported).

Model	% of annot. images	AUC (↑)	Acc. (↑)	F1-score (↑)	$\mathcal{L}_{att}$ (norm) (↓)	$\mathcal{L}_{att}$ (tum) (↓)
CLAM SB	0	0.883 ± 0.033	0.863 ± 0.027	0.797 ± 0.049	-0.664 ± 0.219	-0.441 ± 0.100
MS-CLAM (no att. loss)	12	0.910 ± 0.018	0.902 ± 0.019	0.865 ± 0.027	-0.537 ± 0.086	-0.780 ± 0.081
MS-CLAM	12	0.908 ± 0.013	0.899 ± 0.028	0.861 ± 0.031	-0.963 ± 0.042	-0.954 ± 0.182
MS-CLAM (no att. loss)	25	0.901 ± 0.031	0.895 ± 0.020	0.848 ± 0.036	-0.545 ± 0.087	-0.762 ± 0.145
MS-CLAM	25	0.911 ± 0.016	0.902 ± 0.028	0.867 ± 0.035	-0.966 ± 0.023	-1.170 ± 0.105
MS-CLAM (no att. loss)	62	0.914 ± 0.018	0.905 ± 0.013	0.866 ± 0.016	-0.497 ± 0.164	-0.777 ± 0.066
MS-CLAM	62	0.932 ± 0.008	0.938 ± 0.009	0.913 ± 0.013	-0.946 ± 0.017	-1.271 ± 0.107
MS-CLAM (no att. loss)	100	0.919 ± 0.006	0.907 ± 0.019	0.874 ± 0.023	-0.397 ± 0.147	-0.831 ± 0.065
MS-CLAM	100	0.939 ± 0.008	0.938 ± 0.012	0.916 ± 0.017	-0.915 ± 0.032	-1.318 ± 0.085

visible in the coarse attention maps represented in Figures 9 and 10, where each tile is colored according to its score. In the first one, the tumorous region is much better described by the attention scores when using the attention loss. In the second, although not perfectly uniform, the coarse attention map shows a more widespread attention distribution on the slide.

Localization. Table 8 shows the impact of the attention loss on the tumor localization task. Both the Dice score and the specificity are strongly affected: when using 62% of annotated slides, the presence of the attention loss increases the Dice score by 16%, while also preserving a very high specificity of 0.996. In general, the model's tendency to make more false positives (decrease in specificity) as the percentage of annotated slides increases is better contained thanks to the attention loss, as the specificity does not fall below 0.991 (against 0.927 without it).

Overall, the attention loss improves the slide-level classification performance from 25% annotated slides onward, and yields remarkable improvements for tumor localization for every annotation ratio.

Table 8. Impact of the attention loss on localization. All metrics are averaged on a 5-fold cross validation split of the training set (± a standard error reported).

Model	% of annot. slides	Dice score (tum)	Specificity (norm)
CLAM SB	0	0.212 ± 0.005	0.740 ± 0.034
MS-CLAM (no att. loss)	12	0.437 ± 0.015	0.989 ± 0.003
MS-CLAM	12	0.473 ± 0.023	1.000 ± 0.000
MS-CLAM (no att. loss)	25	0.423 ± 0.034	0.981 ± 0.008
MS-CLAM	25	0.503 ± 0.039	0.999 ± 0.001
MS-CLAM (no att. loss)	62	0.351 ± 0.031	0.946 ± 0.035
MS-CLAM	62	0.513 ± 0.029	0.996 ± 0.002
MS-CLAM (no att. loss)	100	0.323 ± 0.019	0.927 ± 0.021
MS-CLAM	100	0.475 ± 0.023	0.991 ± 0.003

5.2. Exponential Weighted Sampling

In this section, we compare the exponential weighted sampling strategy with 2 other sampling strategies:

- RandSamp: The same slide sampling strategy as CLAM, i.e. sample randomly a single slide at each iteration.
- DoubleSamp: We use the Double Sampling strategy introduced in section 2.5 regardless of the amount of annotated slides used. In this setting, a normal slide and a tumorous slide are both sampled at each step without any specific weight for the annotated samples.
- EWSamp: The Exponential Weighted Sampling strategy: a normal slide and a tumorous slide are sampled simultaneously like in DoubleSamp, but annotated tumorous slides are more likely to be sampled than non-annotated ones, as detailed in Algorithm 1.

Slide-level classification. Table 9 shows the impact of the sampling strategy on the slide classification task. When the amount of annotated slides is > 25%, the EWSamp strategy yields the best results in terms of F1-score and accuracy, while being comparable to the DoubleSamp strategy in terms of AUC. It is only when there are a few annotated slides (e.g., 12%) that the RandSamp strategy reaches higher accuracy and F1-score, albeit with a slightly lower AUC than the other two.

Localization. Table 10 shows the impact of the sampling strategy on the tumor localization task. This time, the RandSamp strategy is far behind the other two: the tile-level paired batch method we present in section 2.2 greatly improves the localization performance of the model. When the number of annotated slides is small (12%), the EWSamp strategy is advantageous compared to the DoubleSamp one, yielding a higher Dice score. When the amount of annotated slides increases, the exponential weighting of the annotated samples is likely to become less important, since the probability that an annotated slide is sampled at each step is higher.

Like the attention loss, the two sampling strategies we proposed bring much higher localization performance compared to the standard sampling method. When only a few annotated slides are available, the exponential weighted sampling is preferable.

6. Discussion

With MS-CLAM, we showed the benefits of using a few slides with tile-level labels in addition to the slide-level ones on both DigestPath2019 and Camelyon16. On the latter, using only 62% of the tile-level labeled slides would have been enough to reach the 1st position on the challenge leaderboard

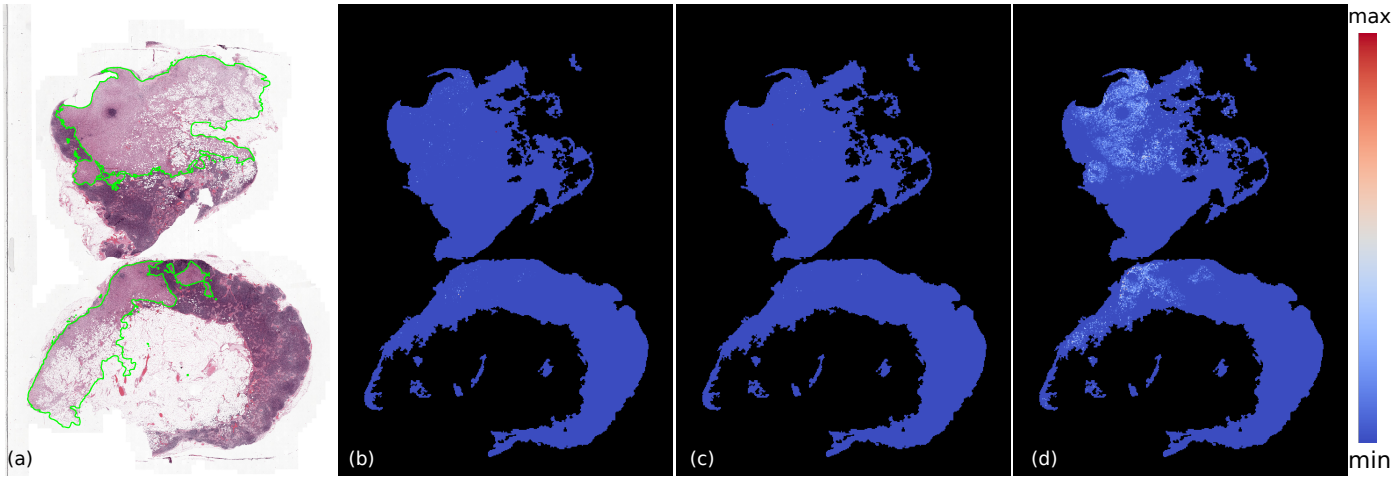

Fig. 9. Slide #90 (tumorous) from the test set of Camelyon16, along with the tile-level coarse attention map computed by each model using the attention scores. The color scale on the right indicates the mapping between colors and attention scores. The former have been rescaled following $a'_k = (a_k - \min(a))/(\max(a) - \min(a))$. (a) The slide thumbnail (metastasis delineated in green). (b)-(d) The coarse attention maps computed by the models with various amounts of supervision. (b) CLAM SB (0%). (c) MS-CLAM (12%, no attention loss). (d) MS-CLAM (12%)

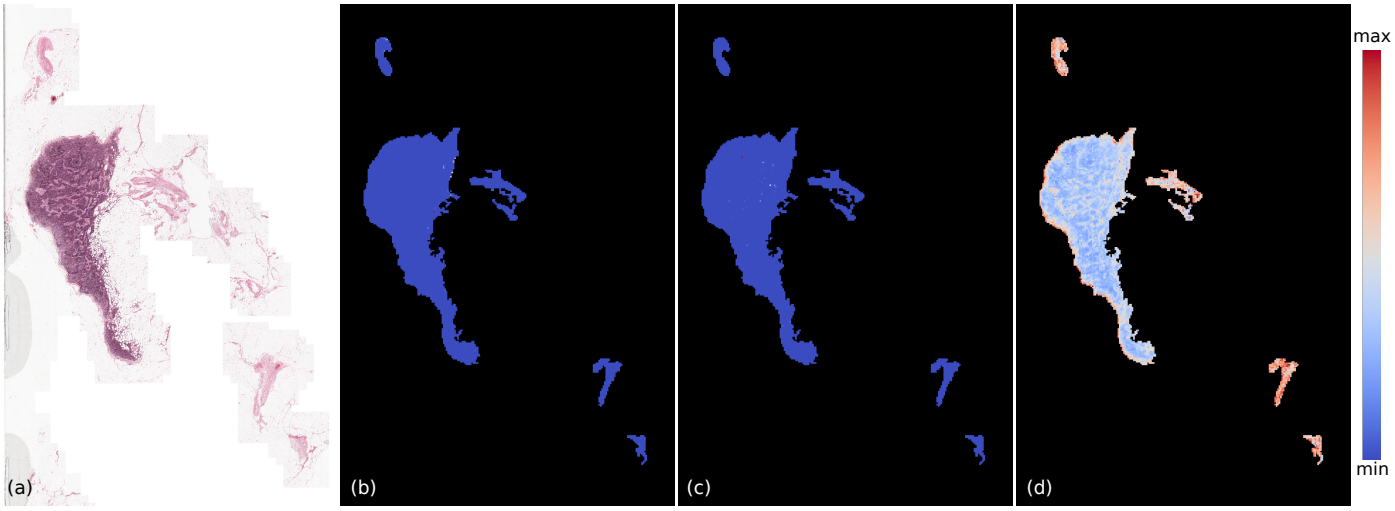

Fig. 10. Slide #119 (normal) from the test set of Camelyon16, along with the tile-level coarse attention map computed by each model using the attention scores. The attention scores have been rescaled and matched to colors following the same procedure as in Figure 9. (a) The slide thumbnail. (b)-(d) The coarse attention maps computed by the models with various amounts of supervision. (b) CLAM SB (0%). (c) MS-CLAM (12%, no attention loss). (d) MS-CLAM (12%).

Table 9. Impact of the exponential weighted sampling on the slide-level classification performance. (± a standard error reported).

Model	% of annot. images	AUC (↑)	Acc. (↑)	F1-score (↑)
CLAM SB	0	0.883 ± 0.033	0.863 ± 0.027	0.797 ± 0.049
MS-CLAM (RandSamp)	12	0.898 ± 0.013	0.916 ± 0.003	0.877 ± 0.007
MS-CLAM (DoubleSamp)	12	0.904 ± 0.006	0.902 ± 0.007	0.866 ± 0.009
MS-CLAM (EWSamp)	12	0.908 ± 0.013	0.899 ± 0.028	0.861 ± 0.031
MS-CLAM (RandSamp)	25	0.902 ± 0.020	0.916 ± 0.019	0.878 ± 0.028
MS-CLAM (DoubleSamp)	25	0.906 ± 0.013	0.899 ± 0.015	0.861 ± 0.019
MS-CLAM (EWSamp)	25	0.911 ± 0.016	0.902 ± 0.028	0.867 ± 0.035
MS-CLAM (RandSamp)	62	0.926 ± 0.010	0.930 ± 0.016	0.900 ± 0.025
MS-CLAM (DoubleSamp)	62	0.933 ± 0.007	0.929 ± 0.003	0.901 ± 0.005
MS-CLAM (EWSamp)	62	0.932 ± 0.008	0.938 ± 0.009	0.913 ± 0.013
MS-CLAM (RandSamp)	100	0.937 ± 0.008	0.933 ± 0.015	0.904 ± 0.025
MS-CLAM (DoubleSamp)	100	0.943 ± 0.004	0.936 ± 0.013	0.913 ± 0.019
MS-CLAM (EWSamp)	100	0.939 ± 0.008	0.938 ± 0.012	0.916 ± 0.017

Table 10. Evaluation of the impact of the slide exponential weighted sampling strategy. All metrics are averaged on a 5-fold cross validation split of the training set (± a standard error reported).

Model	% of annot. slides	Dice score (tum)	Specificity (norm)
CLAM SB	0	0.212 ± 0.005	0.740 ± 0.034
MS-CLAM (RandSamp)	12	0.318 ± 0.027	1.000 ± 0.000
MS-CLAM (DoubleSamp)	12	0.456 ± 0.038	1.000 ± 0.000
MS-CLAM (EWSamp)	12	0.473 ± 0.023	1.000 ± 0.000
MS-CLAM (RandSamp)	25	0.337 ± 0.020	1.000 ± 0.000
MS-CLAM (DoubleSamp)	25	0.507 ± 0.033	1.000 ± 0.000
MS-CLAM (EWSamp)	25	0.503 ± 0.039	0.999 ± 0.001
MS-CLAM (RandSamp)	62	0.351 ± 0.025	1.000 ± 0.000
MS-CLAM (DoubleSamp)	62	0.510 ± 0.028	0.996 ± 0.001
MS-CLAM (EWSamp)	62	0.513 ± 0.029	0.996 ± 0.002
MS-CLAM (RandSamp)	100	0.329 ± 0.016	1.000 ± 0.000
MS-CLAM (DoubleSamp)	100	0.482 ± 0.022	0.993 ± 0.003
MS-CLAM (EWSamp)	100	0.475 ± 0.023	0.991 ± 0.003

based on the AUC results (5th position on the final leaderboard). On DigestPath2019 on the other hand, using 6% of the tile-level labeled slides would have reached the 5th rank in terms of AUC on the second task of the challenge (Da et al., 2022). For both of the challenges, the best results were obtained with the help of deep neural networks trained from scratch on the challenge data, with additional post-processing steps, and sometimes using an ensemble of various heavy architectures (ensemble of networks) to reach the highest possible score. Furthermore, each challenge had its unique best methods, while here we presented a model that reaches near top performance without any post-processing steps, on both datasets. On the DigestPath2019 data, the AUC improvement with respect to the CLAM baseline is rather modest, but MS-CLAM clearly reduces the number of classification errors compared to CLAM (see accuracy and F1-score in Table 3). What is more, MS-CLAM trains in a matter of hours, and reaches state-of-the-art performance with an out-of-domain pre-trained feature extractor, proving the efficiency of such a model. With both higher classification scores, and lower attention losses, the MS-CLAM models provide better performance thanks to a higher key instance recall, that the attention loss promotes. Coupled with paired batch sampling, it allows MS-CLAM to outperform CLAM even without tile-level labels, using weak supervision only. Mixed supervision also allows to sample more tiles within annotated slides, and provides ground-truth labels instead of pseudo-labels for these samples in particular. In turn, MS-CLAM models achieve higher localization performance than their weakly-supervised counterparts.

To profit even further from the annotations, we evaluated the impact of fine-tuning the feature extractor in CLAM and MS-CLAM. Although fine-tuning on its own was sufficient to reach higher results in terms of slide-level classification, its effect on localization was far from what we could achieve with MS-CLAM alone. Furthermore, fine-tuning is a long and costly process (16 hours per fold over 2 GPUs for Camelyon16 when using 62% of the annotated slides). Given this dataset contains only 270 samples, fine-tuning could be very expensive to scale to bigger datasets. MS-CLAM on the other hand needs nearly no additional time compared to CLAM, and is far superior in terms of localization, while even being competitive with fine-tuning in terms of slide-level classification.

There are still several limitations with this implementation of mixed supervision for attention-based MIL, the most critical one being that the tumor region must be exhaustively located by annotations within the annotated set. Missing tumorous regions could induce erroneous tile labels and hamper the tile-level classification. However, this need is limited to only a few slides as shown for both datasets (in Camelyon16, only 11 slides suffice for a performance improvement). Furthemore, we showed that the model was still robust to partial annotations, as some DigestPath2019 tumorous slides exhibit unannotated tumor tissue which was correctly classified (Figure 7). Moreover, although the ground-truth segmentation masks for the Camelyon16 challenge are particularly meticulous, coarser segmentation masks could suffice for our models as a tiling approach is used. This however, brings us to the second limitation of the model: the tile-level localization suffers from inaccuracy, due to square tiles only approximately fitting the tissue parts. It is therefore impossible to obtain more subtle tumor localization using tiles only, although they still offer a good first approximation of the tumor location. In the case of DigestPath2019, where tiles are often overlapping with benign tissue or background, predicted tumorous tiles tend to overestimate the tumor region. The labeling of tiles is also imperfect for the very same reasons. It would be interesting to supervise the attention of the model with finer tile-level labels, accounting for instance for the ratio of tumor within the tile, instead of its mere presence obtained after a hard threshold. Finally, the model presented here was only designed for binary classification, where tile- and slide-level labels coincide: it could be extended to multi-label classification, with different labels at the tile and the slide levels. A good example for this is the Gleason grading of prostate cancer, where tile-level Gleason patterns are insufficient to qualify the entire slide, without the knowledge of the area spanned by the patterns, which is typically accessible via the slide-level label. The tile-to-slide cooperation offered by this kind of model, along with mixed supervision could potentially be of great interest in this scenario.

7. Conclusion

In this paper, we presented a new loss function, coined attention loss, that leverages partially available tile-level labels to constrain the attention distribution in CLAM, an attention-based, weakly-supervised MIL model. Using mixed supervision to exploit both slide- and tile-level labels, we were able to improve the performances of the model for the classification of both entities. With greater coherence between classification and localization, these newly trained models offer better interpretability and fewer false positives among the suspicious regions, furthering their usability in a clinical setting. The framework was built atop an already cost-efficient architecture (Lu et al., 2021), that required few slide-level labels, and limited computational resource, and extended in a similar fashion this effectiveness to the mixed supervision setting, narrowing the amount of required labels to improve upon the baseline. Although for the moment limited to binary classification, with local and global label coherence, we aim to extend the application of mixed supervision to multi-class classification, with

fewer constraints on the relations between the labels at different scales.

Acknowledgments

The authors are grateful to the OPAL infrastructure from Université Côte d'Azur for providing resources and support. This work has been supported by the French government, through the 3IA Côte d'Azur Investments in the Future project managed by the National Research Agency (ANR) with the reference number ANR-19-P3IA-0002.

The authors would like to thank Hind Dadoun for her suggestions on the paired batch method.

References

Arvaniti, E., Claassen, M., 2018. Coupling weak and strong supervision for classification of prostate cancer histopathology images. arXiv preprint arXiv:1811.07013 .

Bejnordi, B.E., Veta, M., Van Diest, P.J., Van Ginneken, B., Karssemeijer, N., Litjens, G., Van Der Laak, J.A., Hermsen, M., Manson, Q.F., Balkenhol, M., et al., 2017. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. Jama 318, 2199–2210.

Bulten, W., Pinckaers, H., van Boven, H., Vink, R., de Bel, T., van Ginneken, B., van der Laak, J., Hulsbergen-van de Kaa, C., Litjens, G., 2020. Automated deep-learning system for gleason grading of prostate cancer using biopsies: a diagnostic study. The Lancet Oncology 21, 233–241.

Campanella, G., Hanna, M.G., Geneslaw, L., Miraflor, A., Werneck Krauss Silva, V., Busam, K.J., Brogi, E., Reuter, V.E., Klimstra, D.S., Fuchs, T.J., 2019. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nature medicine 25, 1301–1309.

Chen, T., Kornblith, S., Norouzi, M., Hinton, G., 2020a. A simple framework for contrastive learning of visual representations, in: III, H.D., Singh, A. (Eds.), Proceedings of the 37th International Conference on Machine Learning, PMLR. pp. 1597–1607. URL: https://proceedings.mlr.press/v119/chen20j.html.

Chen, X., Fan, H., Girshick, R., He, K., 2020b. Improved baselines with momentum contrastive learning. arXiv preprint arXiv:2003.04297 .

Chopra, S., Hadsell, R., LeCun, Y., 2005. Learning a similarity metric discriminatively, with application to face verification, in: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), pp. 539–546 vol. 1. doi:10.1109/CVPR.2005.202.

Ciga, O., Martel, A.L., 2021. Learning to segment images with classification labels. Medical Image Analysis 68, 101912. URL: https://www.sciencedirect.com/science/article/pii/S1361841520302760, doi:https://doi.org/10.1016/j.media.2020.101912.

Coudray, N., Ocampo, P.S., Sakellaropoulos, T., Narula, N., Snuderl, M., Fenyö, D., Moreira, A.L., Razavian, N., Tsirigos, A., 2018. Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. Nature medicine 24, 1559–1567.

Da, Q., Huang, X., Li, Z., Zuo, Y., Zhang, C., Liu, J., Chen, W., Li, J., Xu, D., Hu, Z., et al., 2022. Digestpath: a benchmark dataset with challenge review for the pathological detection and segmentation of digestive-system. Medical Image Analysis , 102485.

Dehaene, O., Camara, A., Moindrot, O., de Lavergne, A., Courtiol, P., 2020. Self-supervision closes the gap between weak and strong supervision in histology. arXiv preprint arXiv:2012.03583 .

Dietterich, T.G., Lathrop, R.H., Lozano-Pérez, T., 1997. Solving the multiple instance problem with axis-parallel rectangles. Artificial Intelligence 89, 31–71. URL: https://www.sciencedirect.com/science/article/pii/S0004370296000343, doi:https://doi.org/10.1016/S0004-3702(96)00034-3.

Durand, T., Thome, N., Cord, M., 2016. Weldon: Weakly supervised learning of deep convolutional neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Fu, Y., Jung, A.W., Torne, R.V., Gonzalez, S., Vöhringer, H., Shmatko, A., Yates, L.R., Jimenez-Linan, M., Moore, L., Gerstung, M., 2020. Pan-cancer computational histopathology reveals mutations, tumor composition and prognosis. Nature Cancer 1, 800–810.

Golestaneh, S.A., Karam, L.J., 2017. Spatially-varying blur detection based on multiscale fused and sorted transform coefficients of gradient magnitudes., in: CVPR, pp. 596–605.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Ilse, M., Tomczak, J., Welling, M., 2018. Attention-based deep multiple instance learning, in: International conference on machine learning, PMLR. pp. 2127–2136.

Kingma, D.P., Ba, J., 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 .

Lapin, M., Hein, M., Schiele, B., 2016. Loss functions for top-k error: Analysis and insights, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Lerousseau, M., Classe, M., Battistella, E., Estienne, T., Henry, T., Leroy, A., Sun, R., Vakalopoulou, M., Scoazec, J.Y., Deutsch, E., et al., 2021. Weakly supervised pan-cancer segmentation tool, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 248–256.

Li, B., Li, Y., Eliceiri, K.W., 2021a. Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14318–14328.

Li, B., Li, Y., Eliceiri, K.W., 2021b. Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 14318–14328.

Li, J., Yang, S., Huang, X., Da, Q., Yang, X., Hu, Z., Duan, Q., Wang, C., Li, H., 2019. Signet ring cell detection with a semi-supervised learning framework, in: International Conference on Information Processing in Medical Imaging, Springer. pp. 842–854.

Liu, G., Wu, J., Zhou, Z.H., 2012. Key instance detection in multi-instance learning, in: Hoi, S.C.H., Buntine, W. (Eds.), Proceedings of the Asian Conference on Machine Learning, PMLR, Singapore Management University, Singapore. pp. 253–268. URL: https://proceedings.mlr.press/v25/liu12b.html.

Lu, M.Y., Chen, R.J., Wang, J., Dillon, D., Mahmood, F., 2019. Semi-supervised histology classification using deep multiple instance learning and contrastive predictive coding. arXiv preprint arXiv:1910.10825 .

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

Marini, N., Otálora, S., Müller, H., Atzori, M., 2021. Semi-supervised training of deep convolutional neural networks with heterogeneous data and few local annotations: An experiment on prostate histopathology image classification. Medical Image Analysis 73, 102165. URL: https://www.sciencedirect.com/science/article/pii/S1361841521002115, doi:https://doi.org/10.1016/j.media.2021.102165.

Mlynarski, P., Delingette, H., Criminisi, A., Ayache, N., 2019. Deep learning with mixed supervision for brain tumor segmentation. Journal of Medical Imaging 6, 034002.

Rymarczyk, D., Borowa, A., Tabor, J., Zielinski, B., 2021. Kernel self-attention for weakly-supervised image classification using deep multiple instance learning, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pp. 1721–1730.

Lubrano di Scandalea, M., Lazard, T., Balezo, G., Bellahsen-Harrar, Y., Badoual, C., Berlemont, S., Walter, T., 2022. Automatic grading of cervical biopsies by combining full and self-supervision. URL: https://hal.archives-ouvertes.fr/hal-03533712, doi:10.1101/2022.01.14.476330. working paper or preprint.

Schmidt, A., Silva-Rodríguez, J., Molina, R., Naranjo, V., 2022. Efficient cancer classification by coupling semi supervised and multiple instance learning. IEEE Access 10, 9763–9773. doi:10.1109/ACCESS.2022.3143345.

Shah, M.P., Merchant, S.N., Awate, S.P., 2018. Ms-net: Mixed-supervision fully-convolutional networks for full-resolution segmentation, in: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (Eds.), Medical Image Computing and Computer Assisted Intervention –

MICCAI 2018, Springer International Publishing, Cham. pp. 379–387.

Shannon, C.E., 1948. A mathematical theory of communication. The Bell system technical journal 27, 379–423.

Shao, Z., Bian, H., Chen, Y., Wang, Y., Zhang, J., Ji, X., zhang, y., 2021. Transmil: Transformer based correlated multiple instance learning for whole slide image classification, in: Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., Vaughan, J.W. (Eds.), Advances in Neural Information Processing Systems, Curran Associates, Inc.. pp. 2136–2147. URL: https://proceedings.neurips.cc/paper/2021/file/10c272d06794d3e5785d5e7c5356e9ff-Paper.pdf.

Sharma, Y., Shrivastava, A., Ehsan, L., Moskaluk, C.A., Syed, S., Brown, D., 2021. Cluster-to-conquer: A framework for end-to-end multi-instance learning for whole slide image classification, in: Medical Imaging with Deep Learning, PMLR. pp. 682–698.

Shi, X., Xing, F., Xie, Y., Zhang, Z., Cui, L., Yang, L., 2020. Loss-based attention for deep multiple instance learning, in: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 5742–5749.

Tourniaire, P., Ilie, M., Hofman, P., Ayache, N., Delingette, H., 2021. Attention-based multiple instance learning with mixed supervision on the camelyon16 dataset, in: Atzori, M., Burlutskiy, N., Ciompi, F., Li, Z., Minhas, F., Müller, H., Peng, T., Rajpoot, N., Torben-Nielsen, B., van der Laak, J., Veta, M., Yuan, Y., Zlobec, I. (Eds.), Proceedings of the MICCAI Workshop on Computational Pathology, PMLR. pp. 216–226. URL: https://proceedings.mlr.press/v156/tourniaire21a.html.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. Advances in neural information processing systems 30.

Wang, D., Khosla, A., Gargeya, R., Irshad, H., Beck, A.H., 2016. Deep learning for identifying metastatic breast cancer. arXiv preprint arXiv:1606.05718 .