Processing Megapixel Images with Deep Attention-Sampling Models Supplementary Material

Angelos Katharopoulos [1 2] **François Fleuret** [1 2]

A. Introduction

This supplementary material is organised as follows: In § B and § C we provide the detailed derivation of the gradients for our *attention sampling*. Subsequently, in § D we mention additional related work that might be of interest to the readers. In § E and § F we present experiments that analyse the effect of our entropy regularizer and the number of patches sampled on the learned attention distribution. In § G, we visualize the attention distribution of our method to show it focuses computation on the informative parts of the high resolution images. Finally, in § H we provide details with respect to the architectures trained for our experiments.

B. Sampling with replacement

In this section, we detail the derivation of equation 11 in our main submission. In order to be able to use a neural network as our attention distribution we need to derive the gradient of the loss with respect to the parameters of the attention function $a(\cdot; \Theta)$ through the sampling of the set of indices Q. Namely, we need to compute

$$\frac{\partial \frac{1}{N} \sum_{q \in Q} f(x; \Theta)_q}{\partial \theta} \tag{1}$$

for all $\theta \in \Theta$ including the ones that affect $a(\cdot)$.

By exploiting the Monte Carlo approximation and the multiply by one trick, we get

$$\begin{aligned}
&\frac{\partial}{\partial \theta} \frac{1}{N} \sum_{q \in Q} f(x; \Theta)_q && (2)\\
&\approx \frac{\partial}{\partial \theta} \sum_{i=1}^{K} a(x; \Theta)_i f(x; \Theta)_i && (3)\\
&= \sum_{i=1}^{K} \frac{\partial}{\partial \theta} \left[a(x; \Theta)_i f(x; \Theta)_i \right] && (4)\\
&= \sum_{i=1}^{K} \frac{a(x; \Theta)_i}{a(x; \Theta)_i} \frac{\partial}{\partial \theta} \left[a(x; \Theta)_i f(x; \Theta)_i \right] && (5)\\
&= \mathbb{E}_{I \sim a(x; \Theta)} \left[\frac{\frac{\partial}{\partial \theta} \left[a(x; \Theta)_I f(x; \Theta)_I \right]}{a(x; \Theta)_I} \right]. && (6)
\end{aligned}$$

C. Sampling without replacement

In this section, we derive the gradients of the attention distribution with respect to the feature network and attention network parameters. We define

- $f_i = f(x; \Theta)_i$ for $i \in \{1, 2, \dots, K\}$ to be the K features
- $a_i = a(x; \Theta)_i$ for $i \in \{1, 2, \dots, K\}$ to be the probability of the i-th feature from the attention distribution a
- $w_i = \sum_{j \neq i} a_j$

We consider sampling without replacement to be sampling an index i from a and then sampling from the distribution $p_i(j)$ defined for $j \in \{1, 2, \dots, i-1, i+1, \dots, K\}$ as follows,

$$p_i(j) = \frac{a_j}{w_i}. \tag{7}$$

Given samples i, j sampled from a and p_i, we can make an unbiased estimator for $\mathbb{E}_{I \sim a}[f_I]$ as follows,

$$\begin{aligned}
a_i f_i + w_i f_j &\simeq && (8)\\
\mathbb{E}_{I \sim a}[\mathbb{E}_{J \sim p_I}[a_I f_I + w_I f_J]] &= && (9)\\
\mathbb{E}_{I \sim a}[a_I f_I + \mathbb{E}_{J \sim p_I}[w_I f_J]] &= && (10)\\
\mathbb{E}_{I \sim a}\left[a_I f_I + \sum_{j \neq I} a_j f_j \right] &= && (11)\\
\mathbb{E}_{I \sim a}\left[\sum_{j=1}^{K} a_j f_j \right] &= && (12)\\
\mathbb{E}_{I \sim a}[f_I]\,. & && (13)
\end{aligned}$$

Using the same i, j sampled from a and p_i accordingly, we can estimate the gradient as follows,

$$\frac{\partial}{\partial\theta}\mathbb{E}_{I\sim a}[f_I] = \tag{14}$$

$$\frac{\partial}{\partial\theta}\mathbb{E}_{I\sim a}[\mathbb{E}_{J\sim p_I}[a_I f_I + w_I f_J]] = \tag{15}$$

$$\frac{\partial}{\partial\theta}\sum_{i=1}^{K}\sum_{j\neq i} a_i p_i(j)\left(a_i f_i + w_i f_j\right) = \tag{16}$$

$$\sum_{i=1}^{K}\sum_{j\neq i}\frac{\partial}{\partial\theta} a_i p_i(j)\left(a_i f_i + w_i f_j\right) = \tag{17}$$

$$\sum_{i=1}^{K}\sum_{j\neq i}\frac{a_i p_i(j)}{a_i p_i(j)}\frac{\partial}{\partial\theta} a_i p_i(j)\left(a_i f_i + w_I f_j\right) = \tag{18}$$

$$\mathbb{E}_{I\sim a}\left[\mathbb{E}_{J\sim p_I}\left[\frac{\frac{\partial}{\partial\theta} a_I p_I(J)\left(a_I f_I + w_I f_J\right)}{a_I p_I(J)}\right]\right] \simeq \tag{19}$$

$$\frac{\frac{\partial}{\partial\theta} a_i p_i(j)\left(a_i f_i + w_i f_j\right)}{a_i p_i(j)} = \tag{20}$$

$$\frac{p_i(j)\left(a_i f_i + w_i f_j\right)\frac{\partial}{\partial\theta} a_i}{a_i p_i(j)} + \frac{a_i \frac{\partial}{\partial\theta} p_i(j)\left(a_i f_i + w_i f_j\right)}{a_i p_i(j)} = \tag{21}$$

$$\left(a_i f_i + w_i f_j\right)\frac{\frac{\partial}{\partial\theta} a_i}{a_i} + \frac{\frac{\partial}{\partial\theta} p_i(j)\left(a_i f_i + w_i f_j\right)}{p_i(j)} = \tag{22}$$

$$\left(a_i f_i + w_i f_j\right)\frac{\partial}{\partial\theta}\log(a_i) + \frac{\frac{\partial}{\partial\theta} p_i(j)\left(a_i f_i + w_i f_j\right)}{p_i(j)}. \tag{23}$$

When we extend the above computations for sampling more than two samples, the logarithm in equation 23 allows us to avoid the numerical errors that arise from the cumulative product at equation 20.

D. Extra related work

For completeness, in this section we discuss parts of the literature that are tangentially related to our work.

Combalia & Vilaplana (2018) consider the problem of high-resolution image classification from the Multiple Instance Learning perspective. The authors propose a two-step procedure; initially random patches are sampled and classified. Subsequently, more patches are sampled around the patches that resulted in confident predictions. The most confident prediction is returned. Due to the lack of the attention mechanism, this model relies in identifying the region of interest via the initial random patches. However, in the second pass the prediction is finetuned if informative patches are likely to be spatially close with each other.

Maggiori et al. (2017) propose a neural network architecture for the pixelwise classification of high resolution images. The authors consider features at several resolutions and train a pixel-by-pixel fully connected network to combine the features into the final classification. The aforementioned approach could be used with our *attention sampling* to approach pixelwise classification tasks such as semantic segmentation.

E. Ablation study on the entropy regularizer

To characterize the effect of the entropy regularizer on our *attention sampling*, we train with the same experimental setup as for the histopathology images of § 4.3 but varying the entropy regularizer $\lambda \in \{0, 0.01, 0.1, 1\}$. The results are depicted in Figure 1. Using no entropy regularizer results in a very selective attention distribution in the first 60 epochs of training. On the other hand, a high value for λ, the entropy regularizer weight, drives the sampling distribution towards uniform.

In our experiments we observed that values close to 0.01 (e.g. 0.005 or 0.05) had no observable difference in terms of the final attention distribution.

F. Ablation study on the number of patches

According to our theory, the number of patches should not affect the learned attention distribution. Namely, the expectation of the gradients and the predictions should be the same and the only difference is in the variance.

In Figure 2, we visualize, in a similar fashion to E, the attention distributions learned when sampling various numbers of patches per image for training. Although the distributions are different in the beginning of training after approximately 100 epochs they converge to a very similar attention distribution.

G. Qualitative results of the learned attention distribution

In this section, we provide additional visualizations of the learned attention distribution using both *attention sampling* and *Deep MIL* on our two real world datasets, namely the Histopathology images § G.1 and the Speed limits § G.2.

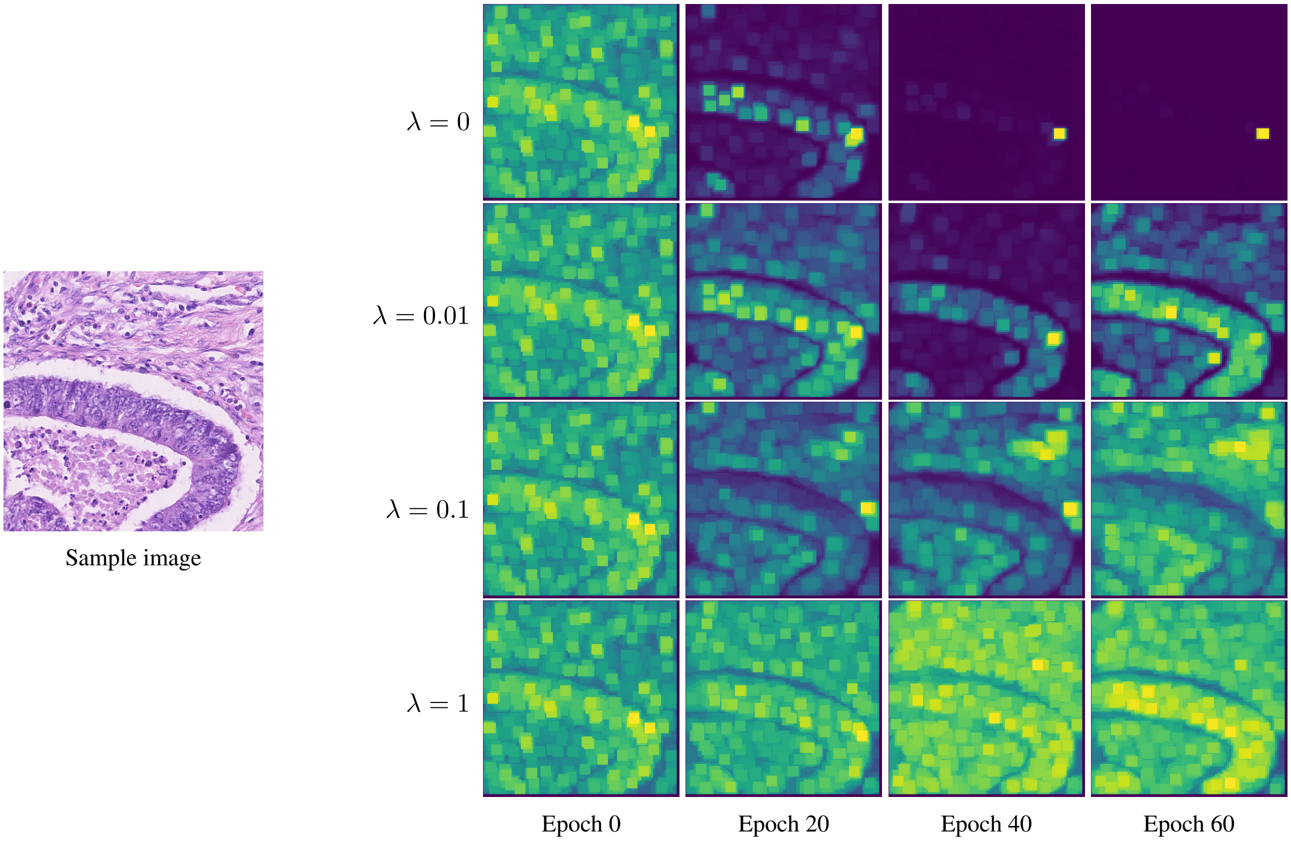


Figure 1. We visualize the effects of the entropy regularizer on the sampling distribution computeed from a test image of the *colon cancer* dataset in the first 60 epochs of training. We observe that no entropy regularizer results in our attention becoming very selective early during training which might hinder the exploration of the sampling space.

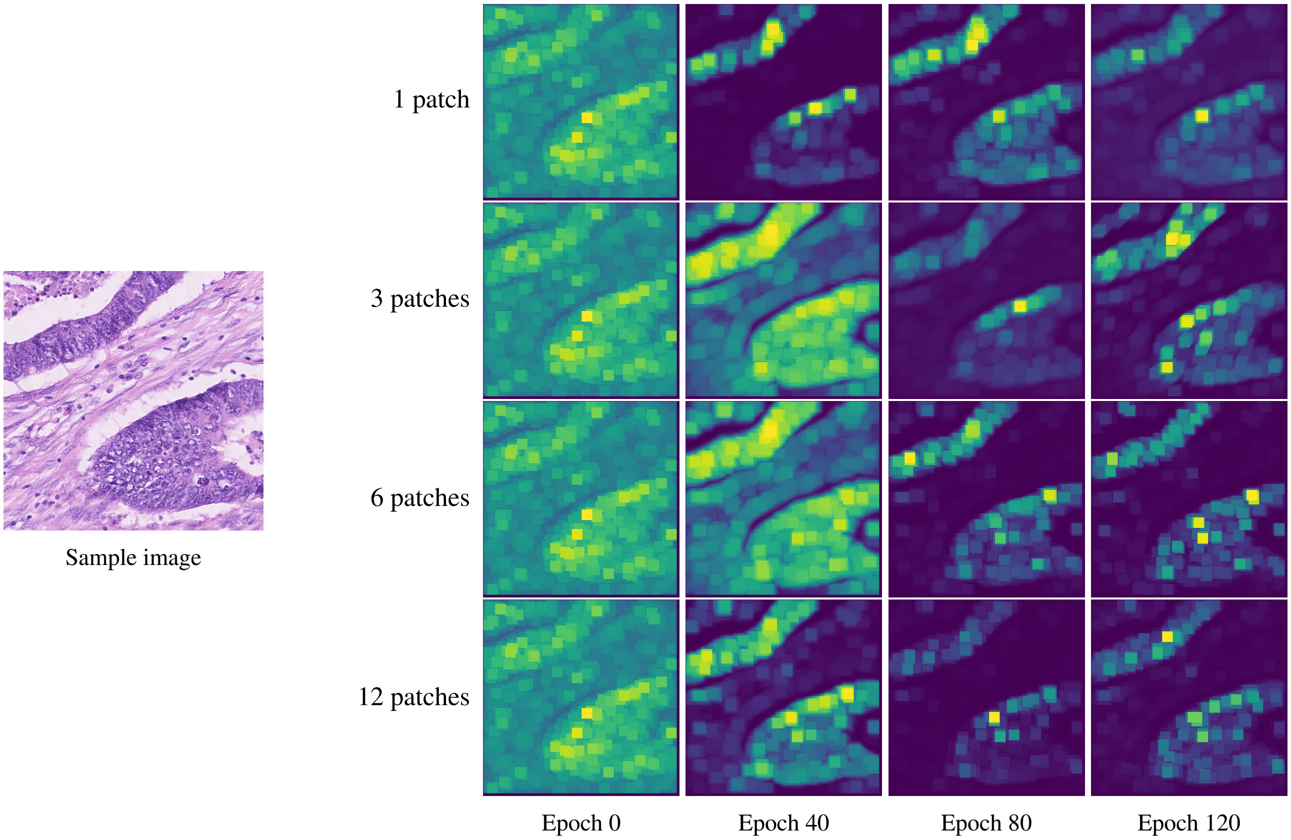


Figure 2. Visualization of the attention distribution when training with varying number of patches. All the distributions converge to approximately the same after ∼100 epochs.

G.1. Histopathology images

In Figure 3 we visualize the learned attention distribution of *attention sampling* and we compare it to *Deep MIL* and the ground truth positions of epithelial cells in an subset of the test set.

We observe that the learned attention distribution is very similar to the one learned by *Deep MIL* even though our model processes a fraction of the image at any iteration. In addition, it is interesting to note that the two methods produce distributions that agree even on mistakenly tagged patches, one such case is depicted in figures 11 and 12 where both methods the top right part of the image to contain useful patches.

G.2. Speed limits

Figure 4 compares the attention distributions of *Deep MIL* and *attention sampling* on the Speed Limits dataset (§ 4.4 in the main paper). This dataset is hard because it presents large variations in scale and orientation of the regions of interest, namely the speed limit signs. However, we observe that both methods locate effectively the signs even when there exist more than one in the image. Note that for some of the images, such as 6 and 15, the sign is not readable from the low resolution image.

H. Network Architecture Details

In this section, we detail the network architectures used throughout our experimental evaluation. The ultimate detail is always code, thus we encourage the reader to refer to the github repository `https://github.com/idiap/attention-sampling`.

H.1. Megapixel MNIST

We summarize the details of the architectures used for the current experiment. For ATS, we use a three layer convolutional network with 8 channels followed by a ReLU activation as the attention network and a convolutional network inspired from LeNet-1 (LeCun et al., 1995) with 32 channels and a global max-pooling as a last layer as the feature network. We also use an entropy regularizer with weight 0.01. The CNN baseline is a ResNet-16 that starts with 32 channels for convolutions and doubles them after every two residual blocks.

We train all the networks with the Adam (Kingma & Ba, 2014) optimizer with a fixed learning rate of 10^{-3} for 500 epochs.

H.2. Histopathology images

We summarize the details of the architecture used for the experiment on the H&E stained images. For ATS, we use a three layer convolutional network with 8 channels followed by ReLU non linearities as the attention network with an entropy regularizer weight 0.01. The feature network of is the same as the one proposed by (Ilse et al., 2018). Regarding, the CNN baseline, we use a ResNet (He et al., 2016) with 8 convolutional layers and 32 channels instead.

We train all the networks for 30,000 gradient updates with the Adam optimizer with learning rate 10^{-3}.

H.3. Speed Limits

We detail the network architectures used for the current experiment. For *attention sampling*, we use an attention network that consists of four convolutions followed by ReLU non-linearities starting with 8 channels and doubling them after each layer. Furthermore, we add a max pooling layer with pool size 8 at the end to reduce the sampling space and use an entropy regularizer weight of 0.05. The feature network of both our model and *Deep MIL* is a ResNet with 8 layers and 32 channels. The CNN baseline is a ResNet-16 that starts with 32 channels for convolutions and doubles them after every two residual blocks.

Again, we we use the Adam (Kingma & Ba, 2014) optimizer with a fixed learning rate of 10^{-3} for 300,000 iterations.

References

Combalia, M. and Vilaplana, V. Monte-carlo sampling applied to multiple instance learning for whole slide image classification. 2018.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Ilse, M., Tomczak, J., and Welling, M. Attention-based deep multiple instance learning. In Dy, J. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2127–2136, Stockholmsmssan, Stockholm Sweden, 10–15 Jul 2018. PMLR. URL `http://proceedings.mlr.press/v80/ilse18a.html`.

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

LeCun, Y., Jackel, L., Bottou, L., Brunot, A., Cortes, C., Denker, J., Drucker, H., Guyon, I., Muller, U., Sackinger, E., et al. Comparison of learning algorithms for handwritten digit recognition. In *International conference on*

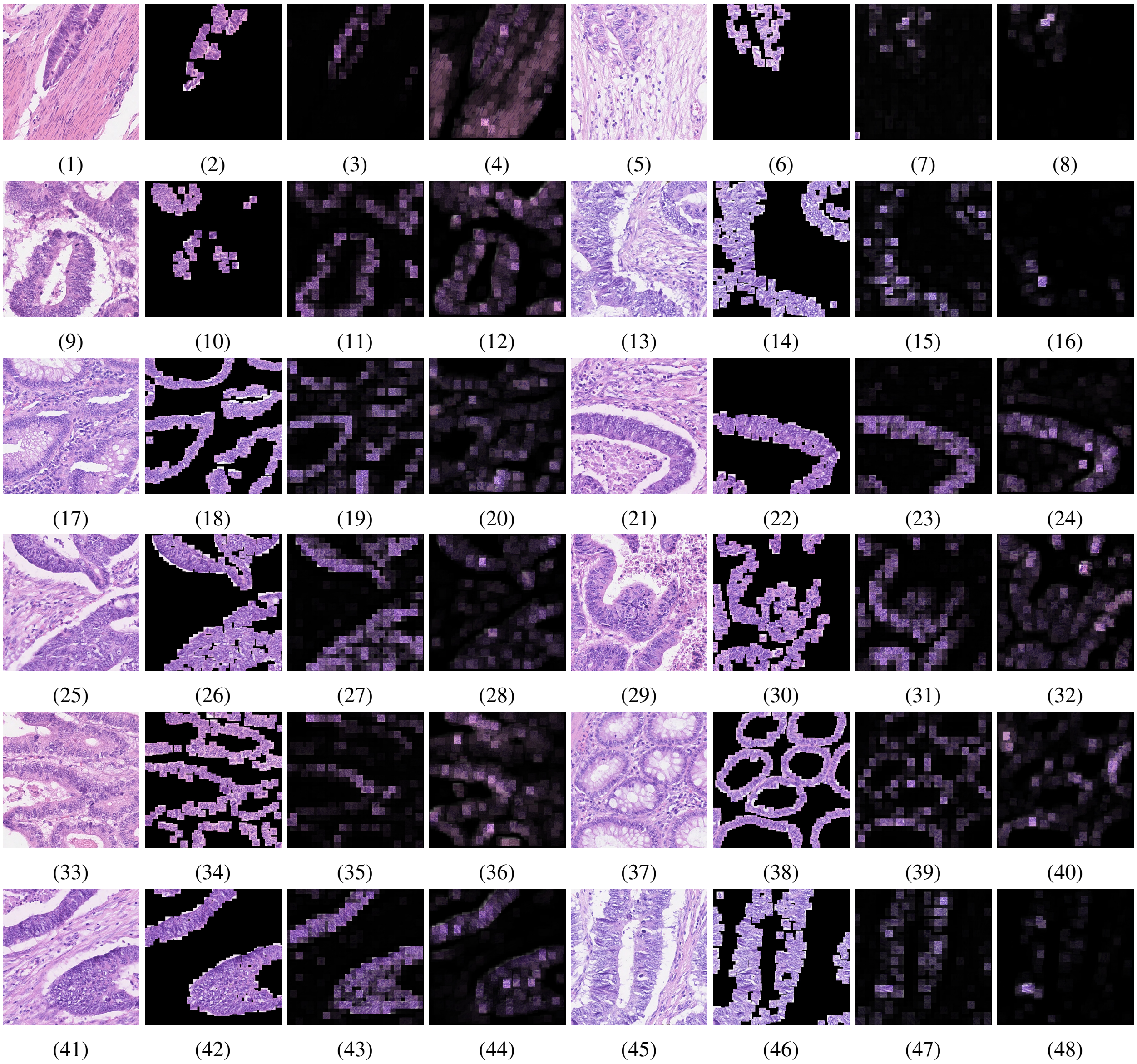
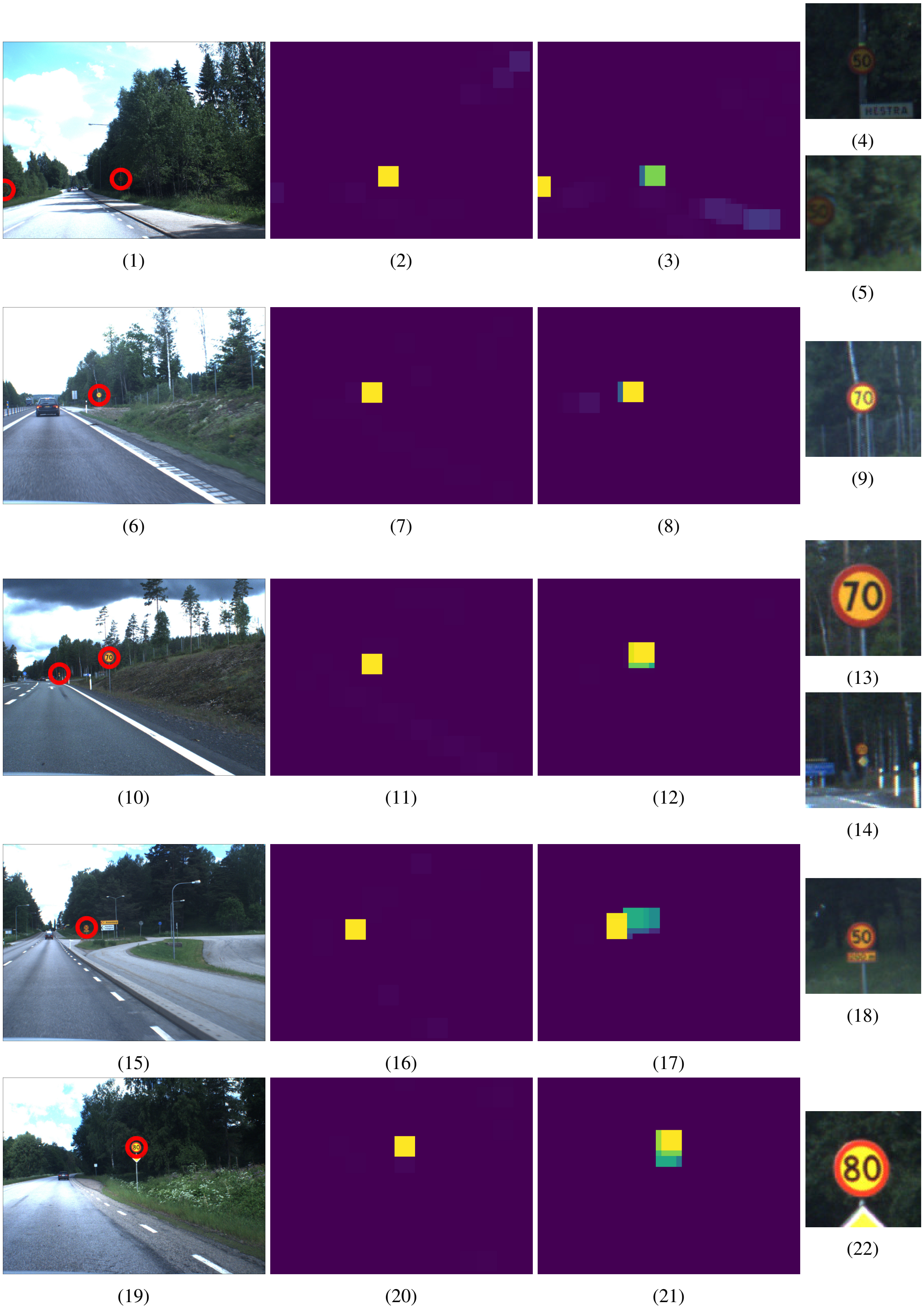


Figure 3. We visualize in groups of 4, the H&E stained image, the ground truth positions of epithelial cells, the attention distribution of *Deep MIL* and the attention distribution of *attention sampling*. We observe that indeed our method learns to identify regions of interest without per patch annotations in a similar fashion to *Deep MIL*.



(1) (2) (3) (4) (5)

(6) (7) (8) (9)

(10) (11) (12) (13) (14)

(15) (16) (17) (18)

(19) (20) (21) (22)

Figure 4. Visualization of the positions of the speed limit signs in test images of the dataset as well as the two attention distributions of *Deep MIL* (left) and *attention sampling* (right) and the patches extracted from the high resolution image at the positions of the signs. Both methods identify effectively the speed limit in the high resolution image.

artificial neural networks, volume 60, pp. 53–60. Perth, Australia, 1995.

Maggiori, E., Tarabalka, Y., Charpiat, G., and Alliez, P. High-resolution image classification with convolutional networks. In *2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pp. 5157–5160. IEEE, 2017.