Sparse-shot Learning with Exclusive Cross-Entropy for Extremely Many Localisations

Andreas Panteli[1,2], Jonas Teuwen[1,2,3], Hugo Horlings[1] and Efstratios Gavves[2,4]
[1]Netherlands Cancer Institute, [2]University of Amsterdam,
[3]Radboud University Medical Center, [4]Ellogon.AI
{a.panteli, j.teuwen, h.horlings}@nki.nl, egavves@uva.nl

Abstract

Object localisation, in the context of regular images, often depicts objects like people or cars. In these images, there is typically a relatively small number of objects per class, which usually is manageable to annotate. However, outside the setting of regular images, we are often confronted with a different situation. In computational pathology, digitised tissue sections are extremely large images, whose dimensions quickly exceed 250'000 × 250'000 pixels, where relevant objects, such as tumour cells or lymphocytes can quickly number in the millions. Annotating them all is practically impossible and annotating sparsely a few, out of many more, is the only possibility. Unfortunately, learning from sparse annotations, or sparse-shot learning, *clashes with standard supervised learning because what is not annotated is treated as a negative. However, assigning negative labels to what are true positives leads to confusion in the gradients and biased learning. To this end, we present* exclusive cross-entropy, *which slows down the biased learning by examining the second-order loss derivatives in order to drop the loss terms corresponding to likely biased terms. Experiments on nine datasets and two different localisation tasks, detection with YOLLO and segmentation with Unet, show that we obtain considerable improvements compared to cross-entropy or focal loss, while often reaching the best possible performance for the model with only 10-40% of annotations.*

1. Introduction

With the advent of deep learning and big datasets, object localisation, be it bounding box detection [1, 2, 3, 4, 5, 6], semantic segmentation [7, 8, 9], or instance segmentation [10, 11, 12, 13], has progressed with leaps and bounds ever since deformable part models [14, 15] and selective search [16, 17]. The basic assumption for all above localisation methods is that all relevant objects in the image are annotated. This is a reasonable assumption for regular images like in PASCAL VOC 2007 [18] or MSCOCO [19], containing on average 512 × 512 -or sometimes up to 1'000 × 1'000- images with no more than a dozen objects per class per image. Outside the realm of regular images, however, we are often confronted with a different situation: digitised tissue sections are typically very large images, of file size around 1-10 GB, whose dimensions can quickly exceed 250'000 × 250'000 px, where relevant objects, such as tumour cells or lymphocytes can quickly number in the millions. Annotating them all, even relying on regions-of-interest, can be hard and in practice only sparse annotations are feasible. In this paper we focus on learning from sparse annotations, coined *sparse-shot learning*, especially when the objective is localising an extreme numbers of objects.

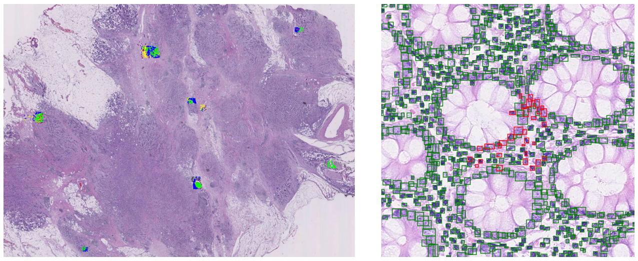

Figure 1: Left: A digitised tissue section containing millions of cells. An image typically corresponds to only a small part of smaller coloured-in regions in the whole slide. Only a handful of annotations are available, and after great effort (about 6'000 annotations in our data). Right: with red the non-exhaustively annotated objects in a small 1'000 by 1'000 region of a tissue slide image, roughly 30% of the total number of objects in green (right).

Learning from sparse annotations clashes with supervised learning, especially in the context of object localisation. In the absence of any other knowledge, the typical assumption is to assign a negative label to all locations in the image that are not annotated as (true) positives. This is a suboptimal choice on two grounds. For one, it is very likely that the annotator could not annotate all relevant objects or that simply they missed many of them. When blindly assuming as negative all unannotated areas, for example in digitised tissue sections, the unannotated objects often amount to more than 90% of the total number of objects [20, 21, 22]. Secondly and more importantly though, assigning a negative label to what is in reality a true positive leads to conflicting gradients [23], that in turn guide the model to poor convergence and generalisation. Sparse-shot learning describes a setting, where standard supervised methods are ill-suited, both from a practical and methodological perspective [24].

Learning given missing, or few, annotations has been explored in the past, albeit in different scenarios than sparse-shot learning. In weakly supervised learning [25, 26] an image-level label is provided, without localisation. The model is then asked to jointly infer likely object locations, as well as learn an accurate classification model. However, when there exist no image-level label, as is the case in many object detection datasets, this type of weak learning cannot infer any localisation labels. Some weakly supervised learning approaches include creating weak pseudo-labels for objects based on confident predictions [27, 28]. Sparse-shot learning is similar in that it assumes all unannotated areas to be potentially negatives, so in a way they correspond to weak negative labels. A key difference is that sparse-shot learning focuses on rejecting specific subsets of these weak negative labels that are likely to add bias and, it does not create new positive labels for object detection.

Focusing on whole images rather than locations in images, in semi-supervised learning [29] the goal is to learn from both annotated and unannotated images. Unannotated images are hence leveraged to learn better and more general image-level classifiers. Similarly, few-shot learning [30] utilises a small number of exhaustively annotated images. Sparse-shot learning, on the other hand, describes a succinctly different setting often encountered in practice: learning localisation models from large images, where only a minute portion of the relevant locations are annotated during training. In this work, our contributions are as follows:

1. We introduce the problem of sparse-shot setting, which is predominant in several imaging scenarios in medical imaging, where acquiring high-quality exhaustive annotations is quite often downright impossible.

2. We provide an analysis showing that the likely culprit leading to poor optimisations with sparse-shot learning is the high speed of learning attributed to biased annotations, and not the biased annotations themselves. To this end, we introduce a novel learning objective coined *exclusive cross-entropy (ECE)* that incorporates a simple cut-off threshold to discard samples contributing large second-order derivatives to the loss, which are the ones speeding up biased learning.

3. Via extensive experimentation on nine datasets and two state-of-the-art architectures, YOLLO [4] and Unet [7], we show that the exclusive cross-entropy generalises in both detection and segmentation. Interestingly, the learned models trained in data, where only 10-40% of the annotations are provided, often reach the same performance as the same models trained with exhaustive annotations, especially in segmentation tasks.

2. Related work

Learning with weak supervision has been a popular area of research. In the work of [26], a weakly supervised learning method is proposed to mediate the effect of partially annotated localisations. They rely on a hybrid dataset containing both image- and instance-level labels, thus rendering the method applicable on data with only instance-level labels. Recently, [23] propose to use the similarity between classes (organs) to merge them together and train on a simpler, more general task. Unlike our work, they rely on multiple classes that are similar while ignoring the background that is by definition dissimilar.

In the work of [31], noisy label learning was explored with loss regularisation focusing during the early stages of training. Similar to weak supervision, transfer learning is employed for the cases of only low information loss. In our sparse-shot learning, however, we have the extreme case of as little as 10% of annotated objects. This results in noisy label learning, creating noisy pseudo-labels which accumulate biased gradient updates. In addition, early learning regularisation [31] penalises the loss function based on the weak labels it iteratively creates, which can lead to a continuous cycle of wrong predictions caused by increasingly more mistakes.

Learning given outliers and imbalanced data has also been explored. In the work of [32] the Huber loss for dense object detection is used to address outlier samples. The Huber loss aims to put smaller weights on outlier cases of hard examples that generate larger errors. Non-exhaustive annotations, however, present themselves with a different challenge since the missing annotations are plentiful, and they are not outliers; down weighing them leads to discarding potentially important data during learning. Recently, focal loss [2] has also offered a significant step towards arbitrating the effect of unforeseen data imbalance, by using the model predictions to weigh more infrequent classes. However, with non-exhaustive annotations the model predictions are inevitably biased due to the incorrect assignment of pseudo-labels to the unannotated data points. Thus, focal loss is sensitive in the absence of exhaustive annotations.

3. Sparse-shot learning

We first introduce the problem setting of *sparse-shot learning*. We then discuss existing methods from the literature and present *exclusive cross-entropy (ECE)*.

3.1. Problem setting

Let $I = (I_m)_{m=1}^{M}$ be a dataset of M images, where each image $I_m = \{x_i, y_i\}, i \in [1, N]$ has a maximum of N relevant objects, and each object, x_i, in the image is assigned one class $y_i \in \{1, \ldots, C\}$ for C number of classes. To reduce notation clutter whenever the subscript m can be inferred by the context, we drop it. x_i can be a pixel, or a bounding box related to an object in an image. For clarity of exposition, we focus first on the binary case, $y_i \in \{0, 1\}$.

In the standard fully supervised setting, a popular choice is the cross-entropy loss

$$\mathcal{L} = -\sum_{x_i \in I} \log p(y_i|x_i). \tag{1}$$

For compactness, we will denote the positive predictions $p_i = p(y_i|x_i)$ and the negative ones by $1-p_i = 1-p(y_i|x_i)$.

In sparse-shot learning, we *do not have all* relevant labels at training time; that is, we do not have exhaustive knowledge of $y_i : \forall x_i \in I$. Instead, we have the annotations y_i for a few locations only, $x_i \in \mathcal{F}$, where $\mathcal{F} \subset I$ is our *foreground* knowledge. The rest of the unannotated image, $\overline{\mathcal{F}} = I - \mathcal{F}$, contains both irrelevant background (set B) $\overline{\mathcal{F}}_B$ for which $y_i = 0$, $\forall x_i \in \overline{\mathcal{F}}_B$, as well as locations $\overline{\mathcal{F}}_U$ that belong to one of the relevant classes, $y_i = 1, .., C$. Expanding equation (1) to incorporate these subsets, we have

$$\begin{aligned} \mathcal{L} = &-\sum_{x_i \in \mathcal{F}} \log p_i \\ &-\sum_{x_i \in \overline{\mathcal{F}}} \Big[y_i \log p_i + (1-y_i) \log \big(1-p_i\big) \Big] \\ = &\ \mathcal{L}_{\mathcal{F}} + \mathcal{L}_{\overline{\mathcal{F}}} \end{aligned} \tag{2}$$

In the absence of any knowledge of annotations in $\overline{\mathcal{F}}$, there exist two following options from the literature to compute the loss in equation (2).

Unannotated regions as background. Following the paradigm of standard object localisation [33, 34, 35, 2], all that is not included in the set of annotations is set to be background. That is, $\overline{\mathcal{F}} \equiv B$. This approach has the drawback that it includes true positive samples in the set of true negative samples, causing bias which adds to the loss as

$$\text{bias} = -\sum_{x_i \in \bar{\mathcal{F}}_U} \log \big(1 - p_i\big) \tag{3}$$

As a result, when optimising the parameters of the neural network, the model gets confused as it is asked to differentiate between samples that are virtually identical in appearance with opposite labels. This pushes the model parameters to poor local minima and, thus, conflicting predictions.

Weak supervision. The predominant paradigm, in similar setups where annotations are partly missing from an image, is weakly supervised learning. Many variants of weakly supervised learning have been explored [36, 26, 37, 38] in this context. The general idea amongst them is that the model f is trained for R rounds. The model from a previous round t is used to predict the labels of unknown samples, $y_i = \operatorname{argmax} p(y_i|x_i; \theta_t)$, often referred to as *pseudo-labels*. The pseudo-labels are then used together with the true labels to minimise cross-entropy in equation (2) and obtain the updated model parameters, θ_{t+1}. However, these new pseudo-labels introduce bias caused by wrong, false positive, assignments of objects, originally associated with the background set $\overline{\mathcal{F}}$. These mistakes add a bias term to the loss described as

$$\text{bias} = -\sum_{r \in R} \Big[\sum_{x_i \in \bar{\mathcal{F}}_{U_{r,t}}} \log \big(1 - p_i\big) + \sum_{x_i \in \bar{\mathcal{F}}_{r,t}} \log \big(1 - p_i\big) \Big] \tag{4}$$

where $\mathcal{F}_{U_{r,t}} \subseteq \mathcal{F}_U$, $\mathcal{F}_{r,t}$ corresponds to the weakly annotated labels by model t at round r. In that respect, weak supervision might eventually do more harm than good, because it biases the final classifier not only on one label side ($y_i = 0$) but all.

3.2. Motivation for exclusive cross-entropy

In the absence of exhaustive ground truth knowledge in the background, any learning algorithm will inevitably introduce bias to the model parameters. Ideally, for sparse-shot learning, we want an algorithm that takes advantage of the background without disproportionately biasing the model parameters *either towards pseudo-positive or pseudo-negative* labels.

To this end, rather than fixating on how to optimally infer the missing annotations $y_i : \forall x_i \in \overline{\mathcal{F}}$, we focus on the learning dynamics of the classifier and how we can optimally influence these dynamics in the sparse-shot learning setting. The objective is to discover background samples - positive or negative- that are likely to add significant bias to the learning, and skip them. Specifically, in the absence of any knowledge of annotations in the background, we tentatively consider all samples in the background as negative samples such that we at least do not add bias to the positive samples in the training set; as noted in equation (3).

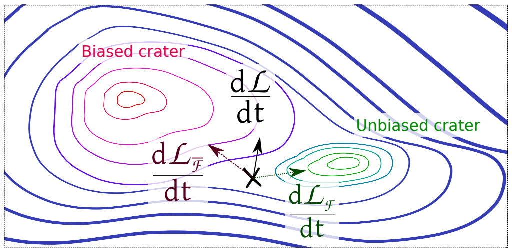
Biased crater
$\frac{d\mathcal{L}}{dt}$
Unbiased crater
$\frac{d\mathcal{L}_{\overline{\mathcal{F}}}}{dt}$
$\frac{d\mathcal{L}_{\mathcal{F}}}{dt}$

Figure 2: When annotations are missing, unannotated data cause learning models to near biased craters due to incorrect information. One way to improve, is to eliminate all instances containing bias, but that would mean eliminating all background examples in non-exhaustively annotated sets. Instead, we propose to simply slow down the pace of learning, captured by $\frac{d^2\mathcal{L}}{dt^2}$, coming from the unannotated data while maintaining the pace of learning from certain annotated data. That way, the model will move faster towards the unbiased crater and eventually the desired solution in spite of the unavoidable bias.

3.2.1 Exclusive cross-entropy for sparse-shot learning

Although in the beginning of the training, any prediction will likely be highly inaccurate, a model is still prone to returning highly confident predictions for both samples in $\mathcal{F}$ and $\overline{\mathcal{F}}$. This is known as overconfidence in neural network predictions when the softmax and sigmoid activation functions are used for classification [39, 40, 41]. This effect is due to the nature of cross-entropy in equation (1), which attains the lowest score when the model predictions $\log p$ are the highest (either $p = 1$ or $1 - p = 1$ for positive and negative samples respectively). It is particularly problematic in the case of missing annotations, as the model will be encouraged to make overconfident predictions for samples in the training, whose annotations are not given but inferred; thus often wrong.

To motivate how to exploit the learning dynamics to break out of this paradox, we illustrate in figure 2 a hypothetical optimisation landscape in gradient descent. Figure 2 highlights the scenario where the bias and the dynamics of learning may adversely affect the final solution. For the purpose of the explanation and without loss of generality, in this example we assume we have one *unbiased minimum* centred in an unbiased crater, which we would obtain if we had perfect knowledge of all relevant annotations in the background. Next to our unbiased crater, there exist multiple biased ones caused by the addition of biased annotations. In reality, neural networks exhibit multiple equivalent minima, however, this does not affect the motivation. Our hypothesis is that when learning from incorrect annotations, models converge to the minima in biased craters; *i.e.* their performance would *not* be as good as models trained on all correct annotations.

Ideally, we want the model to enter the unbiased crater, as in that case it will almost certainly converge to the optimal parameters with standard gradient descent. Unfortunately, the biased gradients will inevitably push the model towards one of the biased craters. One way to limit this, is to make sure the model learns from the unannotated background samples at a slower speed than it does from the certain foreground ones. As learning is captured by the first derivative of the loss with respect to time, $\frac{d\mathcal{L}}{dt}$ (derivatives with respect to parameters correspond to optimal model steps), the speed of learning is captured by the second derivative with respect to time, $\frac{d^2\mathcal{L}}{dt^2}$. In other words, we want the second derivative of the background loss to be small, or even zero, compared to the second derivative of the foreground loss, *i.e.*,

$$\frac{d^2\mathcal{L}_{\overline{\mathcal{F}}}}{dt^2} \ll \frac{d^2\mathcal{L}_{\mathcal{F}}}{dt^2} \tag{5}$$

If equation (5) holds, that indicates that the model learns faster from the positive samples, compared to negative ones, thus increasing the chances of reaching the unbiased crater before getting trapped in a biased one. Moving the detailed computations to the supplementary material, the derivative equation can be expressed as

$$\frac{d^2\mathcal{L}_{\overline{\mathcal{F}}}}{dt^2} \propto p^m(1-p)^n, \tag{6}$$

with polynomial roots $p = 0$ and $1 - p = 0$. To make sure that the second order derivative is zero or almost zero, we shall exclude training samples, in the unannotated areas $\overline{\mathcal{F}}$, which have high confidence predictions. Since all unannotated samples are assigned a weak negative label, we introduce an exclusivity threshold term ρ to the cross-entropy loss in equation (2), only for the unannotated areas $\overline{\mathcal{F}}$.

$$\mathcal{L} = -\sum_{x_i \in \mathcal{F}} \log p_i - \sum_{x_i \in \overline{\mathcal{F}}} \delta(p_i < \rho^\beta) \log\big(1 - p_i\big) \tag{7}$$

where β is an annealing hyper-parameter and $\delta(\cdot)$ is the Kronecker delta function. As learning progresses and the model improves, predictions will become successively more confidently accurate and, hence, the threshold requirement can be relaxed over time by a less strict β. Note that equation (7) can support multiple classes by modifying the log probability $\log\big(1-p_i\big)$ accordingly. We refer to the loss in equation (2) as *exclusive cross-entropy* (ECE).

3.2.2 Intuitive motivation and discussion

Using exclusive cross-entropy enforces that the model should not be over-confident when it is too early for any model to be accurate. In contrast to cross-entropy, exclusive cross-entropy attempts to ignore risky high-confidence predictions and does not encourage the model to assign high scores to as many samples as possible. High-confidence predictions, without sufficient training, run the risk of being false positives/negatives and will wrongfully push the model in the wrong direction. Low-confidence unannotated data corresponding to false negatives (*i.e.* unlabelled objects), on the other hand, will have small gradient magnitudes due to their low score, but their direction will, as learning progresses, hopefully tend to be in the right, approximate, direction. In order to avoid converging too quickly to spurious local minima, the goal is to slow down the learning speed from high-risk unannotated data and reach an unbiased crater first.

Specifically, in the beginning of standard training, the classifier is *de facto* imprecise. Any confidence in predictions are, thus, likely to be misplaced, certainly so for training samples that miss a manual annotation. Given that our background training samples are all considered as tentative negatives ($y_i = 0, \forall x_i \in \overline{\mathcal{F}}$), let us consider the case of a high confidence positive prediction, $p(y_i = 1|x_i) > \rho$, for a real true positive object. The first possible reasoning, is that the model is already capable of recognising objects correctly as positive predictions, $y_i = 1$. This means that the model is already accurate and there is no reason it should receive an update via back-propagation. The second possibility, is that the pseudo-negative annotation is wrong. Back-propagating would update the model towards an incorrect direction. Therefore, not only there is no big need to update the model, but we could be adding bias due to incorrect pseudo-annotations. Given that we do not really know the true label, it is, therefore, better to exclude the contribution of this training sample to the gradient at this round. A similar argument can be constructed for high confidence negative predictions, $p(y_i = 0|x_i) > \rho$.

It is important to note that the annealing and exclusivity threshold employed, are not equivalent as changing the learning rate nor ignoring unannotated objects altogether. Exclusive cross-entropy is similar to a dynamic switchable learning rate; where the rate is dynamically set to zero if training examples are unannotated.

Computational cost. As the exclusive cross-entropy is computed using the already calculated $p(y_i|x_i)$, the computational cost is virtually identical to standard cross-entropy. No retraining, compared to weakly supervised learning, or other expensive processes are required.

Annealing ρ. Our primary objective when satisfying equation (5) is that the model reaches the unbiased crater first. Once in the unbiased crater, the model will eventually reach the desired minimum. By annealing threshold ρ by parameter β we ensure that learning is influenced less by biased loss terms at the early stages and takes more samples into account at the later stages. In experiments, we find that the learning algorithm is robust with respect to ρ and β; so we use the same ρ and β for all our datasets and obtain consistently good performance.

Class imbalance. In object localisation, class imbalance can have strong effects on learning [3, 4]. Especially in large images such as the tissue sections, the amount of irrelevant or background object instances dwarf in comparison to the few positive annotations provided by the annotator. To account for the severe class imbalance, we can complement the exclusive cross-entropy with the focal loss re-weighting scheme, $u(p_i) = -\alpha(1-p_i)^\gamma \log(p_i)$, as originally proposed by [2]. In this case, the *focal* exclusive cross-entropy is computed as

$$\mathcal{L} = -\sum_{x_i \in \mathcal{F}} \log p_i - \sum_{x_i \in \overline{\mathcal{F}}} \delta(p_i < \rho^\beta) u\big(1 - p_i\big) \qquad (8)$$

4. Experiments

4.1. Experimental Setup

Data. We evaluate on the following nine datasets: CoNSeP [13], CPM15 [36], CPM17 [36], CRCHisto [6], Kumar [42], MoNuSeg [42], WBC-NuClick [43], TNBC [44], and our own tumour-infiltrating lymphocyte (TIL) localisation benchmark containing 16 Hematoxylin and Eosin (H&E) stained digital biopsies of whole slide images (WSIs). The largest dataset is TIL with 440’734 images and 45’127 cell annotations, including the 6’631 lymphocytes. The second-largest dataset is WBC-NuClick with 1’463 images, while the second most annotated dataset is the CRCHisto dataset with 29’748 cells. We provide all details and visual examples in the supplementary material.

Evaluation. All datasets, except for TIL, contain images that are only small portions of the digitised tissue sections, such that they can be exhaustively annotated. We create non-exhaustive annotation set variants with 10%, ..., 90% of the annotations (100% is the full set). To make sure the different variants are comparable, we include all annotations in every smaller variant in the variants above (the 80% variant annotations are also in the 90% variant and so on).

We evaluate segmentation using DICE and object detection using F1 score. The TIL dataset contains only a very small portion of all cells, thus we cannot use precision-related metrics, as unknown true positives would be counted as true negatives. Instead, given that in the TIL dataset we have annotations for other cell types that are similar to lymphocytes and are the most likely false positives, we propose the *exclusive recall* computed as $\text{Rec}_{\text{exc}}(y) = \text{Rec}(y) \cdot (1 - \text{Rec}(\neq y))$. While still not accounting for the missed true positives, exclusive recall down weighs the score when predictions correspond to wrong cell types and can quantitatively score *relative* performance between methods.

Architectures. The exclusive cross-entropy is agnostic to the specific segmentation or detection model and architecture. We experiment with two state-of-the-art methods: YOLLO [4] for object detection and Unet [7] for segmentation using standard open-source implementation. We train the models from scratch per non-exhaustive set (10 sets per dataset) and with no pre-training. For hyper-parameter tuning we rely only on the TNBC dataset for segmentation, and reuse the same parameters in all other datasets, experiments and tasks (both for segmentation and detection). We use *no task-specific or dataset-specific parameters* for the exclusive cross-entropy. In all 161 experiments with exclusive cross-entropy with YOLLO and Unet on all nine datasets use the same hyperparameter values, to demonstrate generality and robustness. We include in the supplementary material all the model and training parameters.

4.2. Ablation study

Cross-entropy variants and weakly supervised learning. We report results with exclusive cross-entropy as well as weakly-supervised learning on the 30% and 60% non-exhaustive variants of the TNBC dataset. We start with training using standard cross-entropy. Then, we use the trained model to update the labels in the respective non-exhaustively annotated training sets. If the prediction for an unannotated sample is $p_i > \tau$, then the sample becomes a pseudo-positive, we re-train and repeat the process. Noisy label learning with early learning regularisation [31], performs similar to the standard weakly-supervised learning, as shown further in the supplementary material.

We present results in table 1. We observe that weak supervised learning does not increase the performance of the standard cross-entropy training. A possible reason -in contrast to the regular uses of weak supervision [45]- is that objects in medical images are easy to confuse. Weak supervision works better when the expected confusion is not high. For standard cross-entropy, focal reweighing is not beneficial, likely due to overly down-weighing the actual true positives. Adding focal reweighing, to the unannotated $\overline{\mathcal{F}}$ group, in the exclusive cross-entropy is beneficial and hence, we use focal reweighing in all subsequent experiments with exclusive cross-entropy.

Exclusivity threshold and annealing schedules. Next, we ablate different exclusivity thresholds and annealing schedules. We present two experiments with fixed ρ at 0.5 and 0.75 ($\rho = 1$ is standard cross-entropy). We also present two experiments with linear and sigmoid scheduling in annealing ρ. We gather results in table 2. We observe consistently good performance no matter the type of threshold and scheduling, with sigmoid scheduling doing best. In the following experiments, we will use the sigmoid schedule.

Table 1: Exclusive cross-entropy *vs.* weak supervision for 30% and 60% annotations on TNBC for the detection task.

	τ	F1@30%	F1@60%
Cross-entropy		0.65	0.7
+weak supervision	0.75	0.64	0.68
+weak supervision	0.50	0.62	0.64
+focal loss		0.36	0.49
Exclusive cross-entropy		0.70	0.75
+focal loss		**0.74**	**0.80**

Table 2: Annealing scheduling study on the 30% and 60% sets of TNBC for the detection task.

	F1@30%	F1@60%
Fixed $\rho = 0.75$	0.68	0.70
Fixed $\rho = 0.50$	0.66	0.71
Linear $\rho = 0.75 \cdot \rho_t, \rho_t : 0 \to 1$	0.71	0.73
Sigmoid $\rho = \sigma(\rho_{Linear})$	**0.74**	**0.80**

Table 3: Quantitative results on the TIL localisation dataset scored by the exclusive recall metric.

	Cross-entropy	Focal loss	Huber loss	ECE
Rec_{exc} (↑)	0.85	0.81	0.69	**0.88**

4.3. Sparse-shot segmentation

We present results for segmentation in figure 4 using standard cross-entropy (CE) (assuming what is not annotated is a negative sample), focal loss (FL) [2], Huber loss [32], and exclusive cross-entropy (ECE). The exclusive cross-entropy attains top performance in most datasets and settings. Importantly, the exclusive cross-entropy reaches its near maximum performance consistently with only 40% of the annotations, no matter the dataset. Compared to standard cross-entropy, exclusive cross-entropy improves up to 85%, especially with sparser annotations (*e.g.*, 10% or 20% variants) and harder datasets (datasets that CE scores less than 0.5 in DICE score with 10% of annotations). A surprising result is that focal loss achieves relative better performance in some more sparsely annotated sets, but its performance drops with more exhaustively annotated sets; and is significantly worse than the other methods. A possible cause for this finding is that the focal loss was originally designed for class object loss imbalance during detection with exhaustive annotations [2]. Therefore, the focal loss cross-entropy component, applied to all terms of the loss function, down weighs both $\mathcal{F}$ and $\overline{\mathcal{F}}$ groups equally.

4.4. Sparse-shot detection

We present results for box detection in figure 5 with the same hyper-parameters as in segmentation. In MoNuSeg,

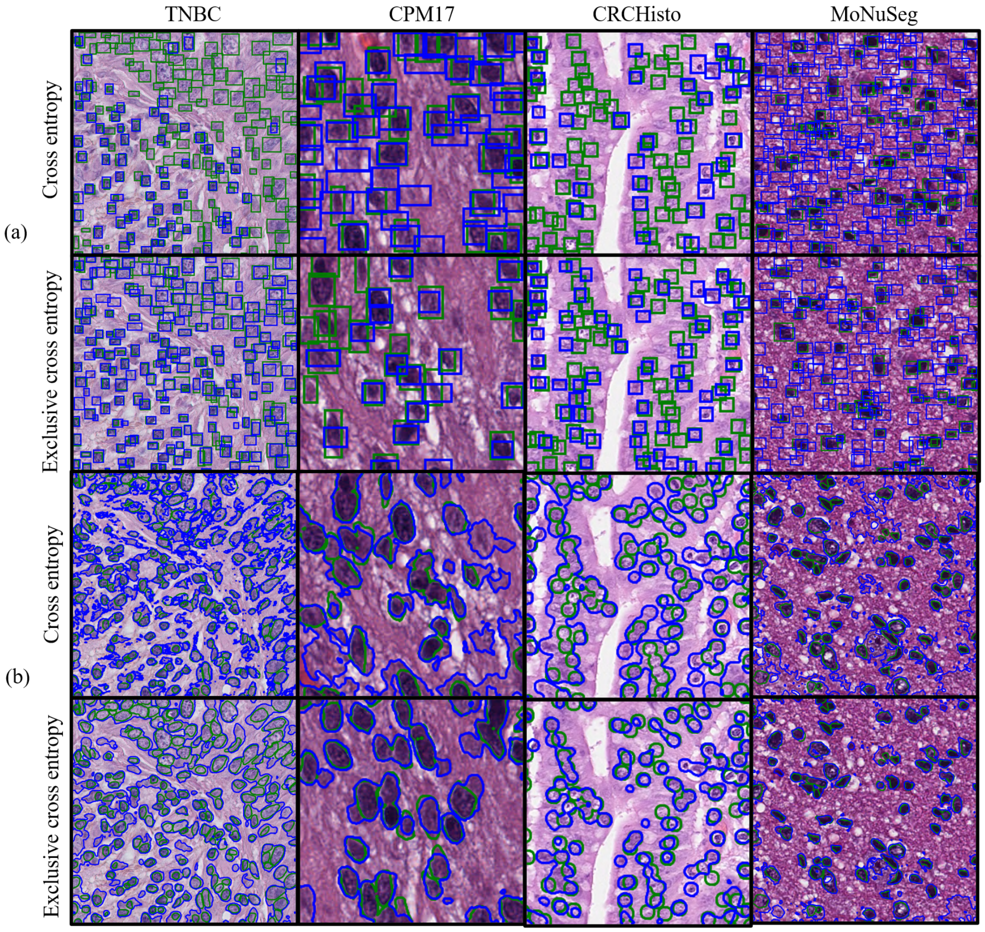

Figure 3: Qualitative results of **ground truth** and **predictions** for datasets TNBC, CPM17, CRCHisto, and MoNuSeg on the 30% non-exhaustive annotation variants for detection (a) and segmentation (b).

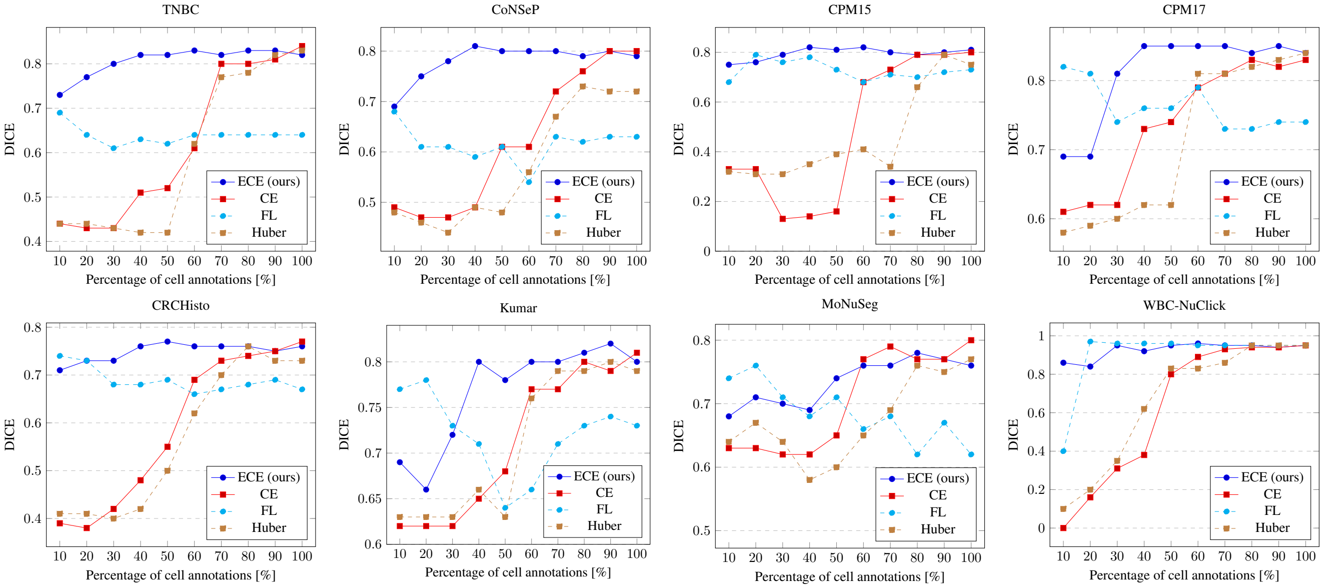

Figure 4: Segmentation results on the non-exhaustive sets of the datasets

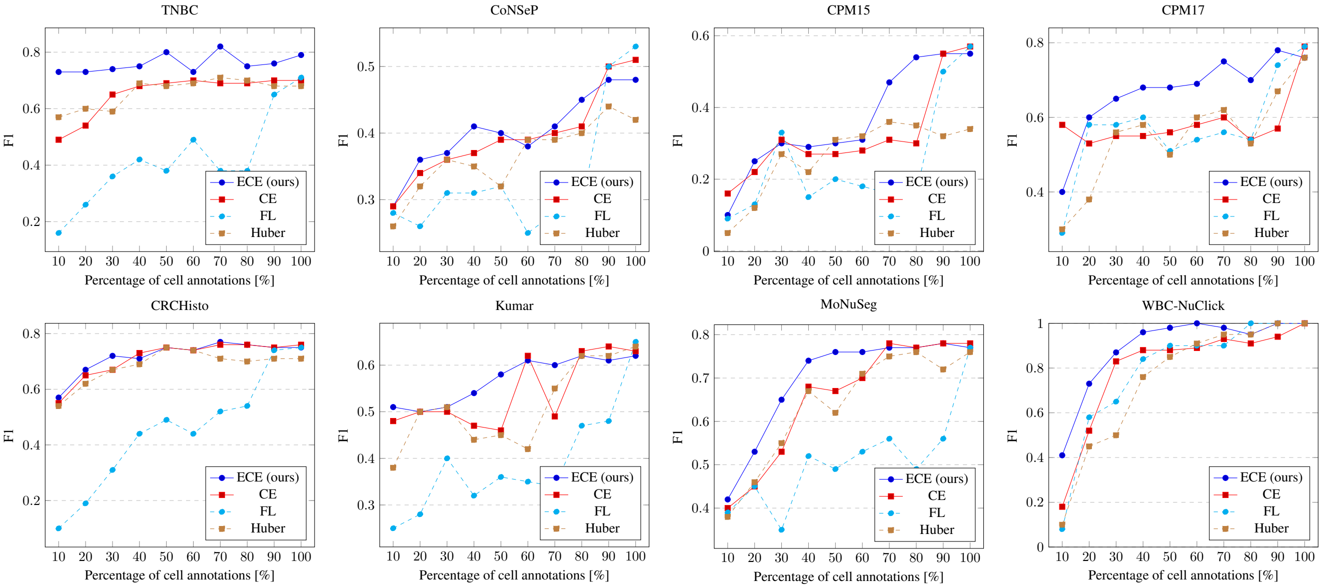

Figure 5: Detection results on the non-exhaustive sets of the datasets

WBC-Nuclick and TNBC exclusive cross-entropy loss outperforms the standard cross-entropy and the focal loss consistently by about 10% in the 10-50% variants. In CRCHisto, CoNSeP, Kumar and CPM15, exclusive cross-entropy still maintains top performance, but for different non-exhaustive variants it is matched by different methods, showing more robustness in the final predictions. A possible reason for the smaller increase of performance compared to segmentation is that segmentation is more challenging than detection. This can be due to the fact that the number of output objects, pixels, in segmentation is larger than the number of objects, cells, in detection. Hence, the number of unannotated objects is relatively lower in the detection task. Last, focal loss appears to have trouble with balancing between the foreground and background due to its uniform weighing strategy.

We, furthermore, present results in terms of exclusive recall on the TIL dataset in table 3. Exclusive cross-entropy performs best, locating correctly the most true positive lymphocytes, while not confusing them with other visually similar cell types like tumour cells or fibroblasts. Upon visual inspection, the difference between the methods is even greater but not quantitatively reflected due to the large number of missing annotations; as discussed in the supplementary material.

4.5. Qualitative results

We show in figure 3 qualitative results for cross-entropy and exclusive cross-entropy. Cross-entropy tends to either under-predict, mostly in detection, or over-predict in segmentation. Exclusive cross-entropy correctly detects most objects while avoiding erroneous background predictions.

5. Conclusion

In this work, we focus on the problem of *sparse-shot learning*, especially in the context of localising extremely many objects. Sparse-shot learning is particularly important for certain types of images, like digitised tissue sections in computational pathology, easily exceeding resolutions of 250’000 × 250’000 pixels and millions of cells to be localised. We show that standard cross-entropy assuming all background as negative labels leads to biased learning and poor optimisation, likely due to the contributions represented by large second-order derivatives in the loss. By ignoring these terms, we present *exclusive cross-entropy*. Extensive experiments on nine datasets and two localisation tasks, detection with YOLLO and segmentation with Unet, show that we obtain considerable improvements compared to cross-entropy or focal loss, while often reaching the best possible accuracy for the model with only 10-40% of annotations present.

Acknowledgements

The collaboration project is co-funded by the PPP Allowance made available by Health Holland[1]. Top Sector Life Sciences & Health, to stimulate public-private partnerships.

[1] https://www.health-holland.com

References

[1] David Tellez, Maschenka Balkenhol, Irene Otte-Höller, Rob van de Loo, Rob Vogels, Peter Bult, Carla Wauters, Willem Vreuls, Suzanne Mol, Nico Karssemeijer, et al. Whole-slide mitosis detection in h&e breast histology using phh3 as a reference to train distilled stain-invariant convolutional networks. *IEEE transactions on medical imaging*, 37(9):2126–2136, 2018.

[2] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[3] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

[4] Mart van Rijthoven, Zaneta Swiderska-Chadaj, Katja Seeliger, Jeroen van der Laak, and Francesco Ciompi. You only look on lymphocytes once. 2018.

[5] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. *arXiv preprint arXiv:2005.12872*, 2020.

[6] Korsuk Sirinukunwattana, Shan E Ahmed Raza, Yee-Wah Tsang, David RJ Snead, Ian A Cree, and Nasir M Rajpoot. Locality sensitive deep learning for detection and classification of nuclei in routine colon cancer histology images. *IEEE transactions on medical imaging*, 35(5):1196–1206, 2016.

[7] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 234–241, 2015.

[8] Thorsten Falk, Dominic Mai, Robert Bensch, Özgün Çiçek, Ahmed Abdulkadir, Yassine Marrakchi, Anton Böhm, Jan Deubner, Zoe Jäckel, Katharina Seiwald, et al. U-Net: deep learning for cell counting, detection, and morphometry. *Nature methods*, 16(1):67–70, 2019.

[9] Xiaomeng Li, Hao Chen, Xiaojuan Qi, Qi Dou, Chi-Wing Fu, and Pheng-Ann Heng. H-denseunet: hybrid densely connected unet for liver and tumor segmentation from ct volumes. *IEEE transactions on medical imaging*, 37(12):2663–2674, 2018.

[10] Amirreza Mahbod, Gerald Schaefer, Isabella Ellinger, Rupert Ecker, Örjan Smedby, and Chunliang Wang. A two-stage u-net algorithm for segmentation of nuclei in h&e-stained tissues. In *European Congress on Digital Pathology*, pages 75–82. Springer, 2019.

[11] Simon Graham, Hao Chen, Jevgenij Gamper, Qi Dou, Pheng-Ann Heng, David Snead, Yee Wah Tsang, and Nasir Rajpoot. Mild-net: Minimal information loss dilated network for gland instance segmentation in colon histology images. *Medical image analysis*, 52:199–211, 2019.

[12] Fidel A Guerrero-Pena, Pedro D Marrero Fernandez, Tsang Ing Ren, Mary Yui, Ellen Rothenberg, and Alexandre Cunha. Multiclass weighted loss for instance segmentation of cluttered cells. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 2451–2455. IEEE, 2018.

[13] Simon Graham, Quoc Dang Vu, Shan E Ahmed Raza, Ayesha Azam, Yee Wah Tsang, Jin Tae Kwak, and Nasir Rajpoot. Hover-net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images. *Medical Image Analysis*, 58:101563, 2019.

[14] Ross Girshick, Forrest Iandola, Trevor Darrell, and Jitendra Malik. Deformable part models are convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 437–446, 2015.

[15] Pedro F Felzenszwalb, Ross B Girshick, and David McAllester. Cascade object detection with deformable part models. In *2010 IEEE Computer society conference on computer vision and pattern recognition*, pages 2241–2248. IEEE, 2010.

[16] Jasper RR Uijlings, Koen EA Van De Sande, Theo Gevers, and Arnold WM Smeulders. Selective search for object recognition. *International journal of computer vision*, 104(2):154–171, 2013.

[17] Koen EA Van de Sande, Jasper RR Uijlings, Theo Gevers, and Arnold WM Smeulders. Segmentation as selective search for object recognition. In *2011 International Conference on Computer Vision*, pages 1879–1886. IEEE, 2011.

[18] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *International journal of computer vision*, 111(1):98–136, 2015.

[19] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[20] Shona Hendry, Roberto Salgado, Thomas Gevaert, Prudence A Russell, Tom John, Bibhusal Thapa, Michael Christie, Koen Van De Vijver, M Valeria Estrada, Paula I Gonzalez-Ericsson, et al. Assessing tumor infiltrating lymphocytes in solid tumors: a practical review for pathologists and proposal for a standardized method from the international immuno-oncology biomarkers working group: part 2: Tils in melanoma, gastrointestinal tract carcinomas, non-small cell lung carcinoma and mesothelioma, endometrial and ovarian carcinomas, squamous cell carcinoma of the head and neck, genitourinary carcinomas, and primary brain tumors. *Advances in anatomic pathology*, 24(6):311, 2017.

[21] Shona Hendry, Roberto Salgado, Thomas Gevaert, Prudence A Russell, Tom John, Bibhusal Thapa, Michael Christie, Koen Van De Vijver, M Valeria Estrada, Paula I Gonzalez-Ericsson, et al. Assessing tumor infiltrating lymphocytes in solid tumors: a practical review for pathologists and proposal for a standardized method from the international immuno-oncology biomarkers working group: part

2: Tils in melanoma, gastrointestinal tract carcinomas, non-small cell lung carcinoma and mesothelioma, endometrial and ovarian carcinomas, squamous cell carcinoma of the head and neck, genitourinary carcinomas, and primary brain tumors. *Advances in anatomic pathology*, 24(6):311, 2017.

[22] Shona Hendry, Roberto Salgado, Thomas Gevaert, Prudence A Russell, Tom John, Bibhusal Thapa, Michael Christie, Koen Van De Vijver, M Valeria Estrada, Paula I Gonzalez-Ericsson, et al. Assessing tumor infiltrating lymphocytes in solid tumors: A practical review for pathologists and proposal for a standardized method from the international immuno-oncology biomarkers working group: Part 1: Assessing the host immune response, tils in invasive breast carcinoma and ductal carcinoma in situ, metastatic tumor deposits and areas for further research. *Advances in anatomic pathology*, 24(5):235, 2017.

[23] Gonglei Shi, Li Xiao, Yang Chen, and S Kevin Zhou. Marginal loss and exclusion loss for partially supervised multi-organ segmentation. *Medical Image Analysis*, page 101979, 2021.

[24] Joel Saltz, Rajarsi Gupta, Le Hou, Tahsin Kurc, Pankaj Singh, Vu Nguyen, Dimitris Samaras, Kenneth R Shroyer, Tianhao Zhao, Rebecca Batiste, et al. Spatial organization and molecular correlation of tumor-infiltrating lymphocytes using deep learning on pathology images. *Cell reports*, 23(1):181–193, 2018.

[25] Nicolas Toussaint, Bishesh Khanal, Matthew Sinclair, Alberto Gomez, Emily Skelton, Jacqueline Matthew, and Julia A Schnabel. Weakly supervised localisation for fetal ultrasound images. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pages 192–200. Springer, 2018.

[26] Mengmeng Xu, Yancheng Bai, Bernard Ghanem, Boxiao Liu, Yan Gao, Nan Guo, Xiaochun Ye, Fang Wan, Haihang You, Dongrui Fan, et al. Missing labels in object detection. In *CVPR Workshops*, 2019.

[27] Ishan Misra, Abhinav Shrivastava, and Martial Hebert. Watch and learn: Semi-supervised learning for object detectors from video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3593–3602, 2015.

[28] Kihyuk Sohn, Zizhao Zhang, Chun-Liang Li, Han Zhang, Chen-Yu Lee, and Tomas Pfister. A simple semi-supervised learning framework for object detection. *arXiv preprint arXiv:2005.04757*, 2020.

[29] Fuyong Xing and Lin Yang. Robust nucleus/cell detection and segmentation in digital pathology and microscopy images: a comprehensive review. *IEEE reviews in biomedical engineering*, 9:234–263, 2016.

[30] Sung Whan Yoon, Jun Seo, and Jaekyun Moon. Tapnet: Neural network augmented with task-adaptive projection for few-shot learning. In *International Conference on Machine Learning*, pages 7115–7123. PMLR, 2019.

[31] Sheng Liu, Jonathan Niles-Weed, Narges Razavian, and Carlos Fernandez-Granda. Early-learning regularization prevents memorization of noisy labels. *arXiv:2007.00151*, 2020.

[32] Deepak Gupta, Barenya Bikash Hazarika, and Mohanadhas Berlin. Robust regularized extreme learning machine with asymmetric huber loss function. *Neural Computing and Applications*, 32(16):12971–12998, 2020.

[33] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.

[34] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[35] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[36] Quoc Dang Vu, Simon Graham, Tahsin Kurc, Minh Nguyen Nhat To, Muhammad Shaban, Talha Qaiser, Navid Alemi Koohbanani, Syed Ali Khurram, Jayashree Kalpathy-Cramer, Tianhao Zhao, et al. Methods for segmentation and classification of digital microscopy tissue images. *Frontiers in bioengineering and biotechnology*, 7:53, 2019.

[37] James A Diao, Wan Fung Chui, Jason K Wang, Richard N Mitchell, Sudha K Rao, Murray B Resnick, Abhik Lahiri, Chirag Maheshwari, Benjamin Glass, Victoria Mountain, et al. Dense, high-resolution mapping of cells and tissues from pathology images for the interpretable prediction of molecular phenotypes in cancer. *bioRxiv*, 2020.

[38] Germán Corredor, Xiangxue Wang, Yu Zhou, Cheng Lu, Pingfu Fu, Konstantinos Syrigos, David L Rimm, Michael Yang, Eduardo Romero, Kurt A Schalper, et al. Spatial architecture and arrangement of tumor-infiltrating lymphocytes for predicting likelihood of recurrence in early-stage non–small cell lung cancer. *Clinical cancer research*, 25(5):1526–1534, 2019.

[39] Johan E Korteling, Anne-Marie Brouwer, and Alexander Toet. A neural network framework for cognitive bias. *Frontiers in psychology*, 9:1561, 2018.

[40] Agustinus Kristiadi, Matthias Hein, and Philipp Hennig. Being bayesian, even just a bit, fixes overconfidence in relu networks. In *International Conference on Machine Learning*, pages 5436–5446. PMLR, 2020.

[41] Marcin Możejko, Mateusz Susik, and Rafał Karczewski. Inhibited softmax for uncertainty estimation in neural networks. *arXiv preprint arXiv:1810.01861*, 2018.

[42] Neeraj Kumar, Ruchika Verma, Sanuj Sharma, Surabhi Bhargava, Abhishek Vahadane, and Amit Sethi. A dataset and a technique for generalized nuclear segmentation for computational pathology. *IEEE transactions on medical imaging*, 36(7):1550–1560, 2017.

[43] Navid Alemi Koohbanani, Mostafa Jahanifar, Neda Zamani Tajadin, and Nasir Rajpoot. Nuclick: A deep learning framework for interactive segmentation of microscopic images. *Medical Image Analysis*, 65:101771, 2020.

[44] Peter Naylor, Marick Laé, Fabien Reyal, and Thomas Walter. Nuclei segmentation in histopathology images using deep neural networks. In *2017 IEEE 14th international symposium on biomedical imaging (ISBI 2017)*, pages 933–936. IEEE, 2017.

[45] Gabriele Campanella, Matthew G Hanna, Luke Geneslaw, Allen Miraflor, Vitor Werneck Krauss Silva, Klaus J Busam, Edi Brogi, Victor E Reuter, David S Klimstra, and Thomas J Fuchs. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature medicine*, 25(8):1301–1309, 2019.