New Bayesian Focal Loss Targeting Aleatoric Uncertainty Estimate: Pollen Image Recognition

Natalia Khanzhina[1] Maxim Kashirin[1]
Andrey Filchenkov[1,2]
[1] ITMO University, [2] ISP RAS Research Center for Trusted Artificial Intelligence
nehanzhina@itmo.ru

Abstract

In biological image recognition, different species might look similar resulting in a small margin, which causes errors in labeling images. Pollen grain image classification heavily suffers from both problems preventing from building well-calibrated recognition models. In this research, we aim to filter out aleatoric uncertainty caused by noisy labeling and similar shape of pollen species. To estimate aleatoric uncertainty, we propose a new Bayesian Focal Softmax loss function. It uses the softmax activation, which is more convenient for a single-label tasks compared to the original Focal loss based on the logistic function. The proposed loss function better estimates aleatoric uncertainty increasing the overall model performance. For evaluation, we used two datasets, POLLEN13L-det containing 13 classes of allergic pollen and POLLEN20L-det containing additional honey plant pollen species. We achieved the state-of-the-art results for both of them by applying the proposed loss function on RetinaNet. It improved the mAP and significantly reduced the variance compared to the regular Focal loss with softmax and provided much better aleatoric uncertainty estimate compared the Bayesian Focal loss with sigmoid activation.

1. Introduction

Seed plants produce pollen that consists of tiny grains looking like a dust. Each plant produces its own pollen resulting in an endless variety of possible pollen species, their distinguishing is required in agriculture, immunology, archaeology, and criminology.

Today, almost 30% of people suffer from pollen allergy that is also known as pollinosis. Without timely measures, it is likely to turn into asthma, which can have a fatal outcome without treatment. In immunology, counting of pollen grains is provided by aeropalinological monitoring [39, 48]. Aeropalinological monitoring provides information about the annual start and duration of the pollen release season to indicate the start time of treatment to control symptoms.

Honey quality control also uses pollen recognition. Today, according to various data, 50 to 80% of the honey market is adulterated [15]. The pollen analysis of honey is currently recognized as the most accurate regarding detecting counterfeit [46].

Automated pollen recognition can be applied in paleopalynology — the study of fossil pollen grains and spores to form a view of the flora of the explored period [47]. The expert can face a difficulty to recognize dried, deformed and destroyed pollen grains. Criminology can also be a field to apply pollen recognition as pollen data is actively used in forensic palynology [28].

All the mentioned fields require automated pollen grain classification based on microscope images. Within a taxon pollen grains vary in their appearance, while pollen of different taxa looks the same due to the similar shape, usually round one. This leads both to labeling errors and a small margin between classes in the latent space, which hampers building precise classifiers.

This is a widespread problem in biological images, where subjects tend to have similar characteristics even when they belong to different classes. The corresponding images are called overlapping samples for they are usually located in overlapping regions in any feature space. They also cause the class overlapping problem, which is considered one of the most complicated in machine learning [30]. In computer vision, the problem becomes even trickier as it causes labeling inaccuracies of the images as the objects from different classes may look similar to an assessor [38]; this also complicates classification.

This problem is natural in multi-label setting [33] when classes may even include other classes. In this case, researchers aim to train networks so they can assign high probability to many classes simultaneously. On contrast, this is unlikely to help in single-label overlapping class problem, which is more widespread in biological image recognition. In this case, overlapping classes increases

so-called aleatoric uncertainty of the models trained on the corresponding data. This uncertainty is also increased by the erroneous labeling caused by the class similarity. Bayesian deep learning [21] provides tools to estimate and reduce aleatoric uncertainty. The only work applying Bayesian deep learning to the pollen grain recognition problem is [23]. It adopted RetinaNet [31] for a Bayesian framework and proposed the modified loss functions, Bayesian Focal loss and Bayesian Smooth L_1 loss. However, Focal loss and Bayesian Focal loss use logistic activation function suitable for multi-label class overlapping problems such as in COCO dataset [32]. To tackle these problems, we propose a new Bayesian Focal Softmax loss based on softmax activation, which is more suitable for the most of computer vision tasks as they are single-labeled.

More precisely, we aimed to answer the following research questions:

RQ1: Can a loss function designed for single-label overlapping classes improve the detection score of pollen in images compared to a multi-label overlapping class loss function?

RQ2: Can this loss function effectively scale when the number of classes increases?

To answer these questions, we propose a novel loss function to model homoscedastic aleatoric uncertainty for the detection task (i.e. joint localization and classification) designed for single-label overlapping classes problem. The paper contributions are the following:

1. A new loss function for the single-label classification task with overlapping classes for modeling the aleatoric uncertainty called **Bayesian Focal Softmax Loss**.

2. The state-of-the-art result for the task of pollen detection on two benchmarks: *POLLEN13L-det* and its extended version, *POLLEN20L-det*.

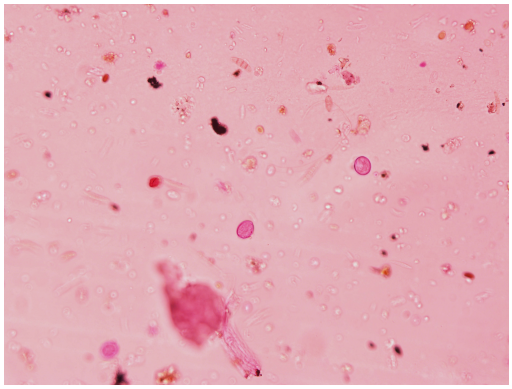

Figure 1. Example of an image from a light microscope slide taken from a pollen sampler. The slide contains *Urtica* pollen and non-pollen fractions.

2. Related Work

2.1. Pollen Recognition

The problem of automation of pollen recognition was first identified in 1968 [13] based on processing pollen images obtained using a microscope. Since then, researchers have been trying to solve the problem suggesting various methods. In works [3, 5–8, 41, 42], the task is solved by extracting of pollen image natural features and applying classical machine learning methods to them, such as Support Vector Machine, Linear Discriminant Analysis, Random Forest, k-Nearest Neighbors, and others.

Recent papers on the problem are mostly based on deep learning. In papers [9, 10, 14, 23, 43, 44], authors adapted the existing state-of-the-art convolutional neural networks for pollen recognition, such as VGG-16 [45], AlexNet [27], DenseNet [19] for classification.

The main drawback of most works on this task is the ignorance of the pollen detection step. This step is the key for automating recognition in general, since images from a pollen sampler usually contain several pollen grains as well as objects that are not pollen. Fig. 1 shows an image of an pollen sampler obtained using a light microscope. Only a few authors [14, 23] addressed this problem via adopted RetinaNet [31] and Faster R-CNN [16] for detection and achieved superior results.

The lack of open pollen datasets labeled for the detection task complicates the building of models. The only benchmarks for the tasks of pollen detection and classification were presented in paper [23]. The authors collected two datasets, *POLLEN13L-det* containing 13 classes of allergic pollen and *POLLEN20L-det* containing additional 7 honey plant pollen species. However, the reported results were only on *POLLEN13L-det*, studying of *POLLEN20L-det* remains topical.

However, these datasets are still quite small, thus, *POLLEN20L-det* consists of only 2413 images of 20 classed. This existing lack leads to high epistemic uncertainty of trained models. Moreover, here is a risk of high aleatoric uncertainty because of the overlapping classes problem inherent in pollen data and possible labeling errors discussed in Sec. 1, which requires the use of Bayesian deep learning.

2.2. Bayesian Deep Learning

Different types of uncertainty influencing predictive models errors in computer vision tasks were first identified by Gal *et al*. [20]. Aleatoric uncertainty reflects the noise level in the training sample and can be used at the prediction stage. Aleatoric uncertainty is divided into homoscedastic — homogeneous for the entire distribution of data, and heteroscedastic — heterogeneous for different data objects. Although the estimation of heteroscedastic uncertainty is

generally more useful for computer vision problems [20], its modeling requires changes in the architecture of the neural network. Also, heteroscedastic uncertainty application in practice requires the development of methods to postprocess prediction for a specific object.

As studies show [21], modeling of homoscedastic aleatoric uncertainty can be performed based on the modification of loss functions only, and not of the whole architecture; which is less laborious. Its modeling leads to increase the accuracy of solving computer vision problems [21]. Work [21] considers the application of accounting for this type of uncertainty for a multitasking architecture that solves the problems of semantic, instance segmentation, and image depth prediction.

Recently, Bayesian deep learning has been widely used for object detection [4, 17, 26, 35–37, 40]. These works mostly focus on the other type of uncertainty, epistemic. Little research is devoted to the evaluation of aleatoric uncertainty [26, 29]. However, the existing works almost do not study the modeling of homoscedastic aleatoric uncertainty for the detection problem, although this may help isolate noise from the data and increase the reliability of the model.

The only work that studies homoscedastic aleatoric uncertainty for pollen recognition is [23], described in Sec. 2.3.

2.3. Bayesian Deep Pollen Recognition

Work [23] is based on [24] adopted RetinaNet [31] for a Bayesian framework by proposing the modified loss functions, Bayesian Focal loss and Bayesian Smooth L_1 loss, to model homoscedastic aleatoric uncertainty.

First, we describe these loss functions, since they are the baseline to our work. Let $f^W(x)$ denote the output of a neural network with weights W on input x and ϵ be the norm of difference between the ground truth value and model prediction, or the model error:

$$\epsilon = \|y - f^W(x)\|.$$

The following loss function was obtained for the localization task, called Bayesian Smooth L_1 loss:

$$BSmooth_{L1}(\epsilon) = \\ = \begin{cases} \frac{\epsilon^2}{2\sigma^2} + \log\sigma, & \text{if } \epsilon < \frac{1}{\beta^2} \\ -\beta^2\epsilon\log\tau + \log\tau + \frac{1}{2\sigma^2\beta^4} + \log\sigma, & \text{otherwise,} \end{cases} \tag{1}$$

where $\tau = 1 - erf\left(\frac{1}{\beta^2\sqrt{2\sigma^2}}\right)$ with erf the Gauss error function [2] and σ^2 is the variance that represents the homoscedastic aleatoric uncertainty, $1/\beta^2$ is the threshold for switching from the L_1 to L_2 function.

Next, the likelihood function was built for the classification task, which is the Bayesian Focal loss modeling aleatoric uncertainty defined as:

$$BFL\left(p(y|f^W(x),\sigma\right) = -\left(\frac{1}{\sigma}\left(1-p_t\right)^{\sigma^{-2}}\right)^{\gamma} \times \\ \times\left(\frac{1}{\sigma^2}\log p_t - \log\sigma\right), \tag{2}$$

where σ^2 is the variance that represents homoscedastic aleatoric uncertainty, γ is the modulating factor, BFL is Bayesian Focal loss and $p_t = \begin{cases} p, & y = 1, \\ 1-p, & \text{otherwise.} \end{cases}$

The detailed derivation of the loss functions is described in [24].

In [23], a classification subnetwork is based on a logistic activation as a likelihood function, as suggested in the original RetinaNet [31], authors of which have chosen it instead of the classical softmax activation function because classes of COCO dataset [32] are overlapping (e.g. "face" class is a part of "person" class). However, it is designed for multi-label setting, while each pollen grain belongs to one plant species only. Thus, we propose to use the softmax activation function in a classification subnetwork for Focal loss. Furthermore, we propose a new Bayesian Focal Softmax loss to model homoscedastic aleatoric uncertainty and combat the small margin issue.

3. Bayesian Focal Softmax Loss

This section presents the derivation of the Bayesian Focal Softmax loss function for the single-label classification task, which models aleatoric uncertainty using the softmax activation function, $p = Softmax\left(f^W(x)\right)$.

The Focal loss function [31] is used for classification and defined as:

$$FL(p_t) = -\alpha\cdot\left(1-p_t\right)^{\gamma}\cdot\log p_t, \tag{3}$$

where

$$p_t = \begin{cases} p & y = 1 \\ 1-p & \text{otherwise} \end{cases},$$

and α and γ are constants.

The Bayesian version of the Focal Softmax loss function can be written as:

$$BFSL\left(p_t\right)) = -\alpha\left(\frac{1}{\sigma^2}\log p_t - \log\sigma\right)\left(1 - \frac{1}{\sigma}p_t^{\frac{1}{\sigma^2}}\right)^{\gamma}. \tag{4}$$

Let us derive it. As in [21, 23, 24], the softmax activation function taken as a likelihood function can be interpreted as

the Boltzmann distribution where the input is scaled by σ^2:

$$p(y|f^W, \sigma) = Softmax\left(\frac{1}{\sigma^2} f^W(x)\right). \quad (5)$$

Here y is a class label, $f^W(x)$ is a class predicted by the model for given x, σ^2 reflects aleatoric uncertainty. Let $f_c^W(x)$ denote the c-th element in vector $f^W(x)$.

As we aim to maximise the likelihood, we can use Focal loss, which it co-directed with the log likelihood. To use it effectively, we need to release $f^W(x)$ in the softmax function from the scaling factor $1/\sigma^2$ and obtain Bayesian Focal Softmax loss, $BFSL$:

$$BFSL(p_t) = FL\left(p\left(y = c|f^W(x), \sigma\right)\right) = \quad (6)$$

$$= -\alpha\left(1 - p(y = c|f^W(x), \sigma)\right)^\gamma \cdot \log\left(p(y = c|f^W(x), \sigma)\right). \quad (7)$$

Let us derive the first multiplier in Eq. 7:

$$1 - p(y = c|f^W(x), \sigma) =$$

$$1 - \frac{\exp\left(f_c^W(x)\frac{1}{\sigma^2}\right)}{\sum_{c'}\exp\left(f_{c'}^W(x)\frac{1}{\sigma^2}\right)} =$$

$$= \frac{\frac{1}{\sigma}\left(\sum_{c'}\exp\left(f_{c'}^W(x)\frac{1}{\sigma^2}\right) - \exp\left(f_c^W(x)\frac{1}{\sigma^2}\right)\right)}{\frac{1}{\sigma}\sum_{c'}\exp\left(f_{c'}^W(x)\frac{1}{\sigma^2}\right)} =$$

$$= \frac{\frac{1}{\sigma}\left(\sum_{c'}\exp\left(f_{c'}^W(x)\frac{1}{\sigma^2}\right) - \exp\left(f_c^W(x)\frac{1}{\sigma^2}\right)\right)}{\left(\sum_{c'}\exp\left(f_{c'}^W(x)\right)\right)^{\frac{1}{\sigma^2}}} =$$

$$= 1 - \frac{\frac{1}{\sigma}\left(\exp f_c^W(x)\right)^{\frac{1}{\sigma^2}}}{\left(\sum_{c'}\exp\left(f_{c'}^W(x)\right)\right)^{\frac{1}{\sigma^2}}} = 1 - \frac{1}{\sigma}p_t^{\frac{1}{\sigma^2}}$$

Substitute Eq. 9 and Eq. 10 in Eq. 3:

$$1 - p(y = c|f^W(x), \sigma) = 1 - \frac{1}{\sigma}p_t^{\frac{1}{\sigma^2}}. \quad (8)$$

Now, let us derive the second multiplier in Eq. 7:

$$\log\left(p(y = c|f^W(x), \sigma)\right) =$$

$$= \frac{1}{\sigma^2}f_c^W(x) - \log\left(\sum_{c'}\exp\left(\frac{1}{\sigma^2}\cdot f_{c'}^W(x)\right)\right) =$$

$$= \frac{1}{\sigma^2}\cdot f_c^W(x) - \frac{1}{\sigma^2}\cdot\log\left(\sum_{c'}\exp\left(f_{c'}^W(x)\right)\right) -$$

$$- \log\left(\frac{\sum_{c'}\exp\left(\frac{1}{\sigma^2}\cdot f_{c'}^W(x)\right)}{\left(\sum_{c'}\exp\left(f_{c'}^W(x)\right)\right)^{\frac{1}{\sigma^2}}}\right).$$

Note that the following statements are fulfilled:

$$\frac{1}{\sigma^2}\cdot f_c^W(x) - \frac{1}{\sigma^2}\cdot\log\left(\sum_{c'}\exp\left(f_{c'}^W(x)\right)\right) = \frac{1}{\sigma^2}\cdot\log p_t, \quad (9)$$

$$\log\left(\frac{\sum_{c'}\exp\left(\frac{1}{\sigma^2}\cdot f_{c'}^W(x)\right)}{\left(\sum_{c'}\exp\left(f_{c'}^W(x)\right)\right)^{\frac{1}{\sigma^2}}}\right) \approx \log\sigma. \quad (10)$$

As a result, we obtain:

$$\log\left(p(y = c|f^W(x), \sigma)\right) = \frac{1}{\sigma^2}\log p_t - \log\sigma. \quad (11)$$

Let us substitute Eq. 8 and Eq. 11 in Eq. 7 and obtain Eq. 4. Finally, we replace $s = \log\sigma^2$ to simplify Eq. 4:

$$BFSL(p_t) = -\alpha\left(e^{-s}\log p_t - \frac{s}{2}\right)\left(1 - e^{-0{,}5s}p_t^{e^{-s}}\right)^\gamma.$$

As a result, we obtain a Bayesian modification of the Focal loss function for the classification problem with activation function $Softmax$. As for the modification of the cross entropy, the equality is fulfilled for $s = 0$.

Thus, for the multi-task Bayesian RetinaNet on softmax with output y_1 for a localization task and y_2 for a classification task, we obtain the following minimization objective:

$$L\left(f^W(x), y_1, y_2, \sigma_1, \sigma_2\right) =$$

$$= BSmoothL_1\left(f^W(x), y_1, \sigma_1\right) + \quad (12)$$

$$+ BFSL\left(f^W(x), y_2, \sigma_2\right),$$

where $BSmoothL_1\left(f^W(x), y_1, \sigma_1\right)$ is the Bayesian Smooth L_1 loss for y_1 from Sec. 2.3, $BFSL\left(f^W(x), y_2, \sigma_2\right)$ is the Bayesian Focal Softmax loss for y_2. This multi-task loss is optimised with respect to W as well as σ_1 and σ_2.

The code implementing the proposed loss is available at https://github.com/NatalieHanzhina/bayesian_retinanet_tf2/tree/softmax.

4. Data

For experimental evaluation, we use a public dataset *POLLEN20L-det* available at https://www.kaggle.com/nataliakhanzhina/pollen20ldet. It contains 20 classes of plant species including 13 allergenic plant species and 8 honey plants with the one being both allergenic and honey. Each image can contain approximately from 1 to 30 pollen grains. In total, it has 2413 images with 7745 pollen grains. Thus, our research covers two applications of pollen recognition, aeropalynology and melissopalynology.

Examples of images are presented in Fig. 2. As seen, pollen grains vary depending on their view (equatorial, polar), observed layer (exine, intine), focal, and angle of rotation. However, pollen of different taxa looks similar due to their shape. For example, two species belong to one genus, *Angelica archangelica* and *Angelica sylvestris* with almost the same shape even to most of experts, which significantly complicates the recognition. The dataset is also labeled by one annotator that can potentially lead to labeling errors. These two factors can cause high homoscedastic aleatoric uncertainty. In Sec. 5 we describe how to filter out this type of uncertainty using our proposed loss function.

5. Experiments and Results Discussion

5.1. Evaluation

We base our study on *POLLEN20L-det* dataset, described in Sec. 4. As no results were reported, no direct comparison is possible with other works besides [23].

To better align the study with the baseline [23], in our experiments we also use *POLLEN13L-det* dataset, which is de-facto a subset of *POLLEN20L-det* containing its 13 classes of allergenic pollen. Also, we evaluated our models only on honey plant pollen to examine the generalization ability of our proposed loss.

We evaluated the proposed loss function, Bayesian Focal Softmax loss, and RetinaNet trained with the softmax activation function and the baseline models, the original RetinaNet and Bayesian RetinaNet (both based on sigmoid activation function) on all three datasets. For bravety, we call the model with the proposed loss Bayesian RetinaNet on softmax.

As a backbone for both RetinaNets and both Bayesian RetinaNets, we used ResNet-50 [18]. We assume that heavier backbones, such as ResNet-101 and ResNet–152, may lead to overfitting because the datasets we use are small. The architectures of Bayesian RetinaNets were the same as the original RetinaNet model and RetinaNet on softmax. The changes were made only for the losses replaced with the Bayesian objectives, baseline and the proposed one. For all the models, we used image scale equal to 800.

Experiments were conducted on two GeForce TITAN RTX GPU with 24 Gb VRAM. The original implementation of the RetinaNet model was taken from the github repository [12] based on the `tensorflow-keras` [1] framework. The implementation of Bayesian RetinaNet was taken from [22] also based on the `tensorflow-keras` [1] framework.

First, we trained the original RetinaNet model applying the Adam optimizer [25] with an initial learning rate of 0.00001 and early stopping. We applied the standard augmentation techniques such as horizontal flips, shifts, rotations.

Next, following [23], we trained our models, which are RetinaNet on softmax and Bayesian RetinaNet on softmax, using the same optimizer and hyperparameters. We initialized $s_1 = \log \sigma_1^2$ for the localization task with 1.0, $s_2 = \log \sigma_2^2$ for the classification task with 0.0.

On the entire dataset, *POLLEN20L-det*, Bayesian models training took less than 30 epochs, which is about an hour and a half in average, while non-Bayesian models training took more than 30 epochs, about two hours in average. We trained 5 distinct models with different random seeds for all the networks to collect statistics on the test set. This method was preferred to the cross-validation because of the small dataset size. The dataset train/test split was 75/25, respectively.

5.2. Quantitative Results

Table 1 compares RetinaNet on softmax, Bayesian RetinaNet on softmax, and the same models on sigmoid activation function on *POLLEN20L-det*. Table 2 compares the same four models on *POLLEN13L-det*. The comparison of these models performance on honey pollen data can be seen in Table 3.

5.3. Qualitative Results

Here, we visualize the feature spaces obtained with Bayesian RetinaNet models trained with Bayesian Focal loss on sigmoid and the proposed Bayesian Focal Softmax loss. Fig. 3 demonstrates the clusters of pollen image embeddings, mapped to 2-dimensional space using UMAP [34]. The plots show that the proposed loss results in a more untangled latent space, where the points of one taxon are clustered more tightly than in the latent space generated by the sigmoid-based loss. On both plots *Angelica archangelica* and *Fagopyrum esculentum* taxa are closely located as some of their pollen views looks similar, while on the Bayesian Focal Softmax loss plot *Angelica archangelica* and *Angelica sylvestris* clusters are far from each other. *Angelica sylvestris* and *Salix* are better clustered, the *Acer* cluster becomes more convex, which is consistent with the quantitative results, as their mAP increased. The *Corylus* points also better clustered resulting in an increase in mAP by 5.44%.

One can conclude that the proposed loss function provides a bigger margin and more separate clusters compared to Bayesian Focal loss on sigmoid, which leads to higher performance.

5.4. Result Analysis

The Bayesian models outperformed the non-Bayesian models in terms of mAP across all three datasets. The difference between mAPs of sigmoid-based models is 0.2–2.8%, while the difference between mAPs of softmax-based models is 0.6–7%. Thus, the impact of modeling aleatoric

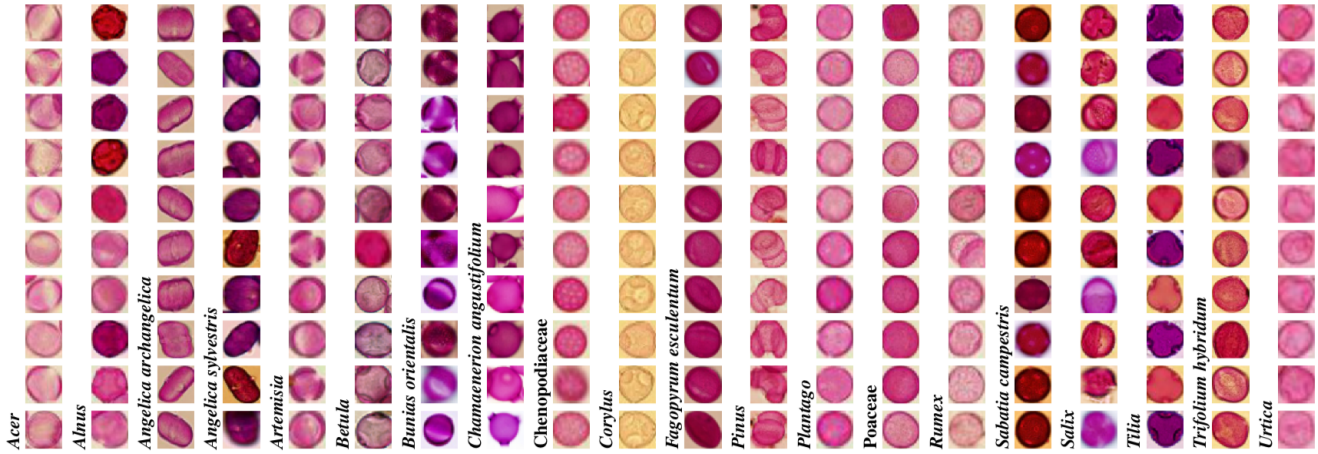

Figure 2. Examples of different pollen classes images from *POLLEN20L-det* dataset. Names written with italic are species, names written with regular font are genus.

Table 1. Comparison of RetinaNet trained with original loss functions, RetinaNet trained with the softmax activation of Focal loss, Bayesian RetinaNet trained with Focal sigmoid-based loss and Bayesian RetinaNet trained with the proposed Focal softmax-based loss functions on ***POLLEN20L-det* dataset**. Focal sigmoid-based and Focal softmax-based losses model homoscedastic aleatoric uncertainty. Here, mAP is the mean average precision presented for different IoU thresholds and object sizes (small, medium, large). The results are presented with a standard deviation of 5 distinct models.

Taxa	RetinaNet, AP, %	Bayesian RetinaNet on sigmoid, AP, %	RetinaNet on softmax (our), AP, %	Bayesian RetinaNet on softmax (our), AP, %
Acer	94.29±0.98	92.87±1.52	92.44±1.50	**94.52±0.86**
Alnus	93.77±1.61	93.72±1.54	90.97±3.72	**95.33±1.57**
Angelica archangelica	98.76±0.10	**98.81±0.01**	98.76±0.10	98.78±0.03
Angelica sylvestris	98.10±0.64	98.22±0.55	98.06±0.37	**98.34±0.37**
Artemisia	98.17±1.60	**99.12±0.46**	94.97±4.64	98.82±2.27
Betula	78.37±3.54	**80.37±2.84**	77.27±0.88	78.61±2.02
Bunias orientalis	98.40±0.35	**98.85±0.36**	98.07±0.47	98.37±0.41
Chamaenerion angustifolium	98.56±0.44	**98.73±0.36**	98.05±0.26	98.49±0.26
Chenopodiaceae	97.53± 0.69	**97.68±1.39**	96.31±0.78	96.95±2.50
Corylus	75.48 ±3.54	77.17±4.98	72.65±8.80	**82.61±3.80**
Fagopyrum esculentum	**100.00**	99.81±0.31	99.93±0.15	**100.00**
Pinus	97.90±1.02	**98.86±0.23**	98.53±0.52	98.12±1.00
Plantago	97.50±0.94	97.77±0.36	**98.20±0.68**	97.29±2.00
Poaceae	99.93±0.12	99.82±0.14	99.78±0.30	**99.95±0.08**
Rumex	**99.70±0.30**	96.22±1.96	99.47±0.45	98.03±1.28
Sabatia campestris	**100.00**	**100.00**	**100.00**	**100.00**
Salix	95.71±0.54	96.47±0.62	94.87±1.48	**97.16±0.45**
Tilia	99.76±0.30	**99.90±0.08**	99.55± 0.31	99.49±0.63
Trifolium hybridum	97.54±0.92	**98.70±0.52**	98.19±0.62	97.84±0.77
Urtica	97.98±0.69	98.44±0.79	97.96±0.74	**98.62±0.75**
mAP	95.87±0.43	96.08±0.53	95.20±0.99	**96.37±0.11**
Aleatoric uncertainty	-	$6.19 \cdot 10^{-5}$	-	$\mathbf{7.56 \cdot 10^{-6}}$

uncertainty for softmax-based RetinaNet is much higher, than for the original RetinaNet.

At the same time, the original RetinaNet based on sigmoid activation surprisingly superpasses RetinaNet based on softmax activation on all the datasets. This result requires further investigation.

The proposed Bayesian Focal Softmax loss outperforms Bayesian Focal loss with sigmoid in most of the cases in

Table 2. Comparison of RetinaNet trained with original loss functions, RetinaNet trained with softmax activation of Focal loss, Bayesian RetinaNet trained with Focal sigmoid-based loss and Bayesian RetinaNet trained with the proposed Focal softmax-based loss functions on ***POLLEN13L-det*** **dataset**. Focal sigmoid-based and Focal softmax-based losses model homoscedastic aleatoric uncertainty. Here, mAP is the mean average precision presented for different IoU thresholds and object sizes (small, medium, large). The results are presented with a standard deviation of 5 distinct models.

Taxa	RetinaNet, AP, %	Bayesian RetinaNet on sigmoid, AP, %	RetinaNet on softmax (our), AP, %	Bayesian RetinaNet on softmax (our), AP, %
Acer	97.82±0.53	98.32±0.62	97.09±0.55	**98.87±0.33**
Alnus	85.97±2.74	90.32±1.79	76.72±5.56	**92.07±0.95**
Artemisia	84.83±1.77	96.01±0.97	67.10±8.47	**97.14±1.13**
Betula	83.67±1.68	84.64±1.13	84.56±1.19	**86.29±1.47**
Chenopodiaceae	93.66± 4.74	**97.58±1.33**	89.38±5.70	97.01±0.85
Corylus	98.36 ±2.02	**100.00**	90.16±7.73	99.56±0.61
Pinus	97.24±1.31	**99.17±0.30**	97.89±0.68	98.85±0.46
Plantago	86.48±10.62	97.67±0.65	76.49±10.78	**98.36±0.45**
Poaceae	**100.00**	**100.00**	99.98±0.05	**100.00**
Rumex	97.46±1.15	97.34±1.46	94.15±1.68	**97.58±1.04**
Salix	97.57±0.32	97.43±0.56	97.69±0.16	**98.04±0.42**
Tilia	**97.61±0.18**	97.51±0.50	96.81±0.30	97.33±0.12
Urtica	95.54±1.19	**96.16±1.00**	95.59±1.07	95.84±0.82
mAP	93.56±1.45	96.32±0.29	89.51±2.73	**96.69±0.15**
Aleatoric uncertainty	-	0.95	-	**0.01**

Table 3. Comparison of RetinaNet trained with original loss functions, RetinaNet trained with softmax activation of Focal loss, Bayesian RetinaNet trained with Focal sigmoid-based loss and Bayesian RetinaNet trained with the proposed Focal softmax-based loss functions on **8 classes of honey plant pollen**. Focal sigmoid-based and Focal softmax-based losses model homoscedastic aleatoric uncertainty. Here, mAP is the mean average precision presented for different IoU thresholds and object sizes (small, medium, large). The results are presented with a standard deviation of 5 distinct models.

Taxa	RetinaNet, AP, %	Bayesian RetinaNet on sigmoid, AP, %	RetinaNet on softmax (our), AP, %	Bayesian RetinaNet on softmax (our), AP, %
Angelica archangelica	98.79±0.034	**98.81±0.03**	98.64±0.18	**98.81±0.02**
Angelica sylvestris	98.12±0.27	**98.29±0.20**	97.73±0.41	98.19±0.43
Bunias orientalis	98.22±0.25	98.59±0.30	97.54±0.60	**98.81±0.43**
Chamaenerion angustifolium	98.56±0.54	98.62±0.44	97.98±0.41	**98.76±0.13**
Fagopyrum esculentum	99.90±0.08	99.91±0.20	99.03±1.03	**100.00**
Sabatia campestris	**100.00**	**100.00**	99.76±0.31	**100.00**
Tilia	99.78±0.12	**99.79±0.20**	99.50±0.43	99.41±0.45
Trifolium hybridum	99.02±0.60	99.55±0.26	98.47±1.02	**99.58±0.13**
mAP	99.05±0.11	99.19±0.07	98.58±0.43	**99.20±0.06**
Aleatoric uncertainty	-	0.04	-	**0.01**

average. Despite it is marginally better than the latter one on the honey plant dataset of 8 classes, on 6 out of 8 classes the proposed loss performed at the same level or better, than the baseline loss.

With the dataset growth the impact of Bayesian Focal Softmax loss is more explicit compared to Bayesian Focal loss with sigmoid: while on honey plant pollen the difference between results is 0.1%, on *POLLEN13L-det* and *POLLEN20L-det* the difference is 0.37% and 0.29% respectively.

The following effect can be seen on *POLLEN20L-det* dataset: while in average the proposed loss is better than the sigmoid one, the latter majorizes on a half of the classes. However, Bayesian Focal Softmax loss leads due to the great boost in the average precision on *Corylus* class: more than 5% better, than Bayesian Focal loss with sigmoid and 10% better, than non-Bayesian models.

Bayesian models provide much smaller deviation than non-Bayesian ones. Furthermore, the Bayesian Focal Softmax loss decreases the deviation much better than the

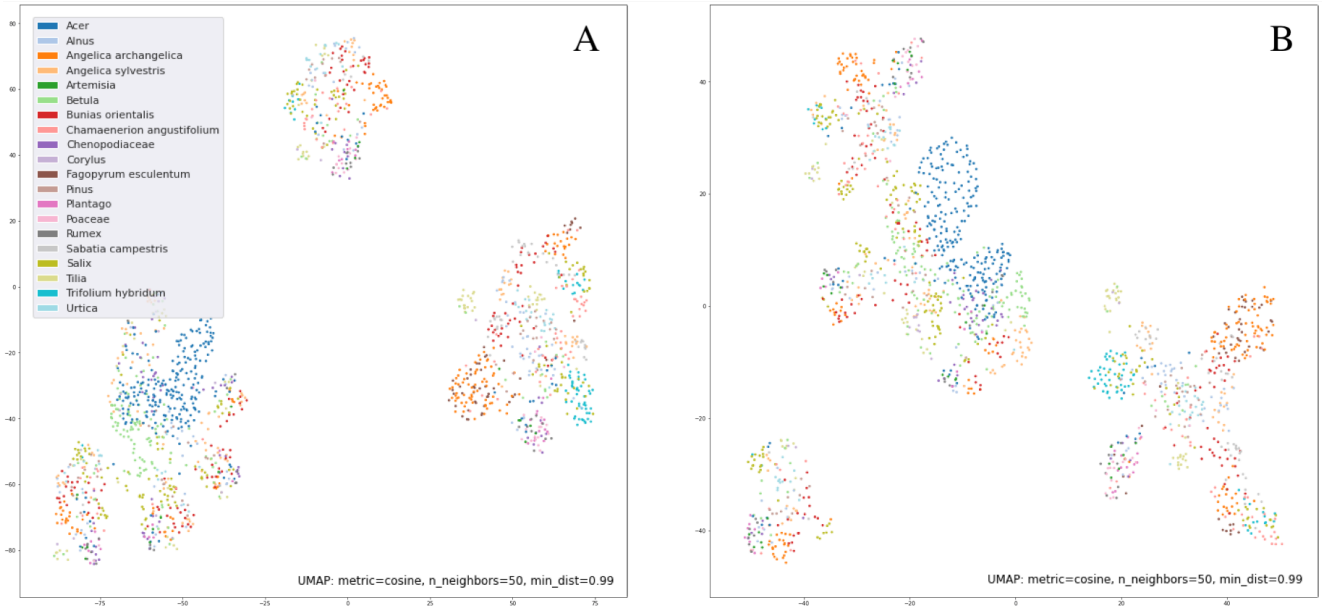

Figure 3. (A) UMAP visualization of embeddings obtained using Bayesian RetinaNet trained with Bayesian Focal loss on sigmoid (baseline); (B) UMAP visualization of embeddings obtained from Bayesian RetinaNet trained with Bayesian Focal Softmax loss (proposed).

Bayesian Focal loss with sigmoid, 7–18 times compared to 2–5 times. Thus, we can conclude, that the proposed loss function better calibrates the model.

Finally, the estimates of aleatoric uncertainty obtained with the proposed loss function are 4–95 times smaller than with the baseline function.

Based on these conclusions, we can answer both **RQ1** and **RQ2** positively.

6. Conclusion

In this work, we have proposed the novel loss function for the pollen detection task, Bayesian Focal Softmax loss. The proposed loss function fits single-label tasks, models homoscedastic aleatoric uncertainty while model training, filters out this uncertainty and increases the model performance.

The proposed loss was studied based on open benchmarks *POLLEN20L-det* and *POLLEN13L-det*. Pollen classes are highly overlapping, which causes aleatoric uncertainty and the small margin. Using this function, we achieved the new state-of-the-art on *POLLEN13L-det* detection task, 96.69% of mAP. Furthermore, we first obtained the result in another pollen detection benchmark *POLLEN20L-det*, achieving 96.37% of mAP on 20 classes.

The proposed loss increases mAP on different pollen datasets compared to the Bayesian Focal loss with sigmoid activation. It also significantly increases mAP and improves model calibration compared to both original RetinaNet and RetinaNet on softmax. More importantly, it reduces model variance and improves aleatoric uncertainty estimate resulting in a bigger margin between classes.

The proposed loss function can be generalized to any other object detection task and applied to other models based on the RetinaNet architecture, for example, SpineNet [11], ATSS [49]. The developed loss function can be interesting to apply for modeling heteroscedastic aleatoric uncertainty to increase the interpretability of detection per object.

Acknowledgements

We thank Alexey Lapenok for their help on implementation, Inna Anokhina for the proofreading and Larisa Novoselova for their help on biological issues. This work was supported by a grant for research centers in the field of artificial intelligence, provided by the Analytical Center for the Government of the Russian Federation in accordance with the subsidy agreement (agreement identifier 000000D730321P5Q0002) and the agreement with the Ivannikov Institute for System Programming of the Russian Academy of Sciences dated November 2, 2021 No. 70-2021-00142.

References

[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga,

Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org. 5

[2] Milton Abramowitz, Irene A Stegun, and Robert H Romer. Handbook of mathematical functions with formulas, graphs, and mathematical tables, 1988. 3

[3] Gary Allen. An automated pollen recognition system: a thesis submitted to massey university, turitea, palmerston north, new zealand in fulfilment of the requirements for the degree of master of engineering. 2008. 2

[4] Abhijit Bendale and Terrance E Boult. Towards open set deep networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1563–1572, 2016. 3

[5] Alain Boucher, Pablo J Hidalgo, Monique Thonnat, Jordina Belmonte, Carmen Galan, Pierre Bonton, and Régis Tomczak. Development of a semi-automatic system for pollen recognition. *Aerobiologia*, 18(3-4):195–201, 2002. 2

[6] Chun Chen, Emile A Hendriks, Robert PW Duin, Johan HC Reiber, Pieter S Hiemstra, Letty A de Weger, and Berend C Stoel. Feasibility study on automated recognition of allergenic pollen: grass, birch and mugwort. *Aerobiologia*, 22(4):275–284, 2006. 2

[7] Manuel Chica. Authentication of bee pollen grains in bright-field microscopy by combining one-class classification techniques and image processing. *Microscopy research and technique*, 75(11):1475–1485, 2012. 2

[8] Celeste Chudyk, Hugo Castaneda, Romain Leger, Islem Yahiaoui, and Frank Boochs. Development of an automatic pollen classification system using shape, texture and aperture features. In *LWA 2015 Workshops: KDML, FGWM, IR, and FGDB*, 2015. 2

[9] Amar Daood, Eraldo Ribeiro, and Mark Bush. Sequential recognition of pollen grain z-stacks by combining cnn and rnn. In *The Thirty-First International Flairs Conference*, 2018. 2

[10] André R de Geus, Celia AZ Barcelos, Marcos A Batista, and Sérgio F da Silva. Large-scale pollen recognition with deep learning. In *2019 27th European Signal Processing Conference (EUSIPCO)*, pages 1–5. IEEE, 2019. 2

[11] Xianzhi Du, Tsung-Yi Lin, Pengchong Jin, Golnaz Ghiasi, Mingxing Tan, Yin Cui, Quoc V. Le, and Xiaodan Song. Spinenet: Learning scale-permuted backbone for recognition and localization. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 11589–11598. IEEE, 2020. 8

[12] Fizyr. Keras-retinanet. https://github.com/fizyr/keras-retinanet, 2021. 5

[13] JR Flenley. The problem of pollen recognition. *Problems in Picture Interpretation*, pages 141–145, 1968. 2

[14] Ramón Gallardo-Caballero, Carlos J García-Orellana, Antonio García-Manso, Horacio M González-Velasco, Rafael Tormo-Molina, and Miguel Macías-Macías. Precise pollen grain detection in bright field microscopy using deep learning techniques. *Sensors*, 19(16):3583, 2019. 2

[15] Norberto L García. The current situation on the international honey market. *Bee World*, 95(3):89–94, 2018. 1

[16] Ross B. Girshick. Fast R-CNN. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 1440–1448. IEEE Computer Society, 2015. 2

[17] Ali Harakeh, Michael Smart, and Steven L Waslander. Bayesod: A bayesian approach for uncertainty estimation in deep object detectors. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 87–93. IEEE, 2020. 3

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 5

[19] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 2261–2269. IEEE Computer Society, 2017. 2

[20] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5574–5584, 2017. 2, 3

[21] Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 7482–7491. IEEE Computer Society, 2018. 2, 3

[22] Natalia Khanzhina. Bayesian retinanet. https://github.com/NatalieHanzhina/bayesian_retinanet_tf2, 2021. 5

[23] Natalia Khanzhina, Andrey Filchenkov, Natalia Minaeva, Larisa Novoselova, Maxim Petukhov, Irina Kharisova, Julia Pinaeva, Georgiy Zamorin, Evgeny Putin, Elena Zamyatina, et al. Combating data incompetence in pollen images detection and classification for pollinosis prevention. *Computers in Biology and Medicine*, page 105064, 2021. 2, 3, 5

[24] Natalia Khanzhina, Alexey Lapenok, and Andrey Filchenkov. Towards robust object detection: Bayesian retinanet for homoscedastic aleatoric uncertainty modeling. *Bayesian Deep Learning Workshop at NeurIPS*, 2021. 3

[25] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. 5

[26] Florian Kraus and Klaus Dietmayer. Uncertainty estimation in one-stage object detection. In *2019 IEEE Intelli-*

gent Transportation Systems Conference (ITSC), pages 53–60. IEEE, 2019. 3

[27] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012. 2

[28] Andrew R Laurence and Vaughn M Bryant. Forensic palynology and the search for geolocation: Factors for analysis and the baby doe case. *Forensic science international*, 302:109903, 2019. 1

[29] Michael Truong Le, Frederik Diehl, Thomas Brunner, and Alois Knol. Uncertainty estimation for deep neural object detectors in safety-critical applications. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, pages 3873–3878. IEEE, 2018. 3

[30] Han Kyu Lee and Seoung Bum Kim. An overlap-sensitive margin classifier for imbalanced and overlapping data. *Expert Systems with Applications*, 98:72–83, 2018. 1

[31] Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 2999–3007. IEEE Computer Society, 2017. 2, 3

[32] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In David J. Fleet, Tomás Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, volume 8693 of *Lecture Notes in Computer Science*, pages 740–755. Springer, 2014. 2, 3

[33] Weiwei Liu, Haobo Wang, Xiaobo Shen, and Ivor W Tsang. The emerging trends of multi-label learning. *IEEE transactions on pattern analysis and machine intelligence*, 44(11):7955–7974, 2021. 1

[34] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018. 5

[35] Dimity Miller, Feras Dayoub, Michael Milford, and Niko Sünderhauf. Evaluating merging strategies for sampling-based uncertainty techniques in object detection. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 2348–2354. IEEE, 2019. 3

[36] Dimity Miller, Lachlan Nicholson, Feras Dayoub, and Niko Sünderhauf. Dropout sampling for robust object detection in open-set conditions. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3243–3249. IEEE, 2018. 3

[37] Dimity Miller, Niko Sünderhauf, Michael Milford, and Feras Dayoub. Uncertainty for identifying open-set errors in visual object detection. *arXiv preprint arXiv:2104.01328*, 2021. 3

[38] Curtis G Northcutt, Anish Athalye, and Jonas Mueller. Pervasive label errors in test sets destabilize machine learning benchmarks. *arXiv preprint arXiv:2103.14749*, 2021. 1

[39] Larisa Viktorovna Novoselova and Nataliya Minaeva. Pollen monitoring in perm krai (russia)-experience of 6 years. *Acta Agrobotanica*, 68(4), 2015. 1

[40] Janis Postels, Francesco Ferroni, Huseyin Coskun, Nassir Navab, and Federico Tombari. Sampling-free epistemic uncertainty estimation using approximated variance propagation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2931–2940, 2019. 3

[41] Kimberly C Riley, Jeffrey P Woodard, Grace M Hwang, and Surangi W Punyasena. Progress towards establishing collection standards for semi-automated pollen classification in forensic geo-historical location applications. *Review of Palaeobotany and Palynology*, 221:117–127, 2015. 2

[42] Olaf Ronneberger, Hans Burkhardt, and Eckart Schultz. General-purpose object recognition in 3d volume data sets using gray-scale invariants-classification of airborne pollen-grains recorded with a confocal laser scanning microscope. In *Pattern Recognition, 2002. Proceedings. 16th International Conference on*, volume 2, pages 290–295. IEEE, 2002. 2

[43] Víctor Sevillano and José L Aznarte. Improving classification of pollen grain images of the polen23e dataset through three different applications of deep learning convolutional neural networks. *PloS one*, 13(9):e0201807, 2018. 2

[44] Víctor Sevillano, Katherine Holt, and José L Aznarte. Precise automatic classification of 46 different pollen types with convolutional neural networks. *Plos one*, 15(6):e0229751, 2020. 2

[45] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. 2

[46] Sónia Soares, Joana S Amaral, Maria Beatriz PP Oliveira, and Isabel Mafra. A comprehensive review on the main honey authentication issues: Production and origin. *Comprehensive Reviews in Food Science and Food Safety*, 16(5):1072–1100, 2017. 1

[47] Alfred Traverse. *Paleopalynology*, volume 28. Springer Science & Business Media, 2007. 1

[48] Ulas Uguz, Aykut Guvensen, and Nedret Sengonca Tort. Annual and intradiurnal variation of dominant airborne pollen and the effects of meteorological factors in çeşme (izmir, turkey). *Environmental monitoring and assessment*, 189(10):1–18, 2017. 1

[49] Shifeng Zhang, Cheng Chi, Yongqiang Yao, Zhen Lei, and Stan Z. Li. Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 9756–9765. IEEE, 2020. 8