ALE : active learning extension for object detection

Theo Oriol, Jerôme Pasquet, Jérome Cortet

ALE : active learning extension for object detection

T. Oriol[1,3] J. Pasquet[1,2] J. Cortet[3]

[1] AMIS, Université de Montpellier 3, Montpellier, France

[2]TETIS - Inrae, AgroParisTech, Cirad, CNRS, Univ. Montpellier, Montpellier, France

[3]UMR 5175 Centre d'Ecologie Fonctionnelle et Evolutive, Université Paul Valéry Montpellier 3, Université de Montpellier, EPHE, CNRS, IRD, CEFE UMR, Montpellier, France

{theo.oriol,jerome.pasquet,jerome.cortet}@univ-montp3.fr

Abstract

Monitoring human activities impact on soil biodiversity over time is a costly and resource-intensive challenge. Modern technologies like deep learning offer a promising solution because they can analyze large datasets much faster than humans. However, deep learning relies on extensively annotated datasets, and annotating these samples is both time-consuming and expensive, complicating its application. This paper introduces a novel active learning approach called Active Learning Extension (ALE), which aims at improving model performance in object detection tasks while minimizing the need for extensive data annotation. Traditional active learning methods typically rely solely on prediction uncertainty to select images for annotation, which can be suboptimal when introducing new classes. ALE addresses this limitation by considering both the uncertainty and the number of predictions. This dual consideration leads to significant improvements, particularly in scenarios like Collembola detection, where creating and updating datasets is highly time-intensive. Our evaluation demonstrates that ALE significantly enhances model performance compared to state-of-the-art methods. The results underscore the importance of selecting challenging examples and accounting for the number of predictions to optimize active learning in object detection.

keywords

Deep learning, Active learning, Object detection.

1 Introduction

With the rise of environmental concerns, the need for tools to monitor the impact of human activities on soil over time has become urgent. Various metrics have been developed to assess soil quality [1, 2], one of which is biodiversity [3]. Collembola, commonly known as springtails, are a class of arthropods that, like other soil organisms, are sensitive to changes in soil properties such as pH, temperature, soil moisture, and nutrient availability. Consequently, they are currently used as a biodiversity indicator, particularly in agricultural and forest practices [4, 5]. Collembola play a crucial role in nutrient cycling and soil aggregation within their ecosystems, and soils can contain thousands of these individuals per square meter [6]. However, using Collembola as an indicator generates a substantial amount of data [7], which can take months to process due to the specialized expertise and the identification having to be done using a microscope [8]. This makes the process very time-consuming. Over the last few years, deep learning models have emerged as promising tools in ecology. The identification of Collembola using deep learning has already been demonstrated [9, 10]. However, to enhance the performance of these tools and enable them to identify a larger pool of species, there is a need to add more data, which conflicts with the time-consuming nature of manual Collembola identification. Given that expert identification is so time-consuming, optimizing the annotation process using active learning is a solution [11, 12, 13, 14, 15]. The challenge is that state-of-the-art active learning for object detection has been designed to improve models on already existing classes, not on new ones. In this paper, we introduce a new active learning technique to add new species to the Collembola datasets. The premise of this technique is that when adding new species to the datasets, it is more efficient to add more annotations with less uncertainty than fewer annotations with more uncertainty. Since the new species, have at first no annotations in the datasets, adding even a small amount of annotations can have a significant impact on the model results.

2 Related

2.1 Active learning

The use of deep learning requires extensive training data, but annotating new samples can often be time-consuming. Active learning aims to maximize model performance while minimizing the number of samples that need to be annotated. This is especially relevant for Collembola detection, where creating and updating datasets is very time-consuming. State-of-the-art active learning for object de-

tection typically follows this process : first, a model is trained on a base dataset. Then, the model makes predictions on a pool of unannotated images. Each prediction is evaluated based on its uncertainty, as it is more beneficial to provide the model with challenging examples that bring new information rather than easy predictions that do not significantly enhance model performance. After each prediction is evaluated, the images receive a score by aggregating the evaluated prediction scores. The top-scoring images are then annotated by an expert.

2.2 Metrics

Least confidence. The least confidence is one of the two main metrics used to evaluate the uncertainty in a model's predictions. It is based on the premise that the smaller the difference between the highest probability and the second highest probability, the greater the uncertainty. The formula is as follows :

$$LC = 1 - (p_1 - p_2) \tag{1}$$

Here, p_1 is the highest probability and p_2 is the second highest probability. The higher the least confidence value, the greater the uncertainty.

Entropy. Entropy is the second main metric used to evaluate the uncertainty. It considers that the flatter the distribution of probabilities is, the more unsure the model is, to do that the entropy is calculated using the following formula :

$$entropy = -\sum_{i=0}^{N} p_i log(p_i) \tag{2}$$

where N represents the length of the probability distribution, and p_i the probability p at the index i.

Aggregation of Detection Metric

The scoring for each image is determined by aggregating the scores S of their predictions. The state-of-the-art aggregation methods include the sum, mean, and maximum of the scores. Images without predictions receive a score of 0.

$$A_{\text{sum}} = \sum_{i=1}^{N} S_i \tag{3}$$

The sum aggregation method tends to prefer images with more annotations but does not consider the number of annotations.

$$A_{\text{max}} = \max_{i \in 1,2,\ldots,N} S_i \tag{4}$$

The maximum aggregation method focuses on identifying images with the most challenging predictions and ignores the number of predictions.

$$A_{\text{mean}} = \frac{1}{N} \sum_{i=1}^{N} S_i \tag{5}$$

The mean aggregation method, similar to the maximum method, does not consider the number of predictions and favors images with only difficult predictions.

2.3 Model

We used Yolov5x6, yolov5 biggest version. Yolov5 is an advanced object detection model developed by Ultralytics as an extension of Yolov3 [16]. It is a one-step detector, meaning it simultaneously detects and classifies objects. It comprises a backbone (CSPDarknet), a neck, and a prediction head (Figure 1). The backbone extracts features from the image, which are then mixed and combined by the neck for prediction. The detection head uses these features to propose bounding boxes and classes. To generate these proposals, the image is divided into multiple grids of various scales, with each cell proposing N objects. Yolov5 achieves precise detection by using anchors to predict box coordinates and different scale aspect ratios. Anchors facilitate coordinate prediction by using different ratios and sizes tailored to fit the data, in this case, Collembola. Instead of directly predicting Collembola coordinates, Yolov5 identifies which cell contains the center of the Collembola and predicts the height and width ratio of the anchor used for the prediction. This approach simplifies the task, greatly improving the accuracy of the model's coordinate predictions.

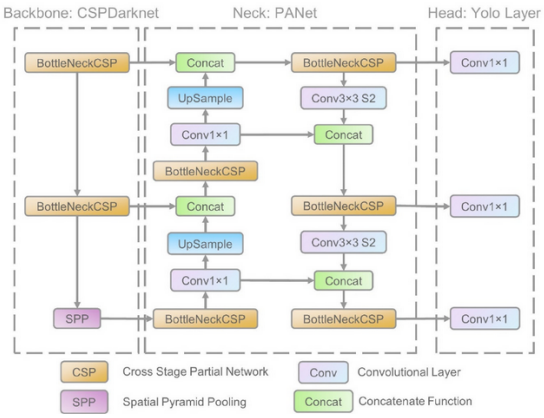

FIGURE 1 – *Yolov5 architecture [17].*

3 Proposed Method

3.1 Datasets

The data used in this paper was initially developed to benchmark object detection for Collembola on microscope images using deep-learning techniques [9]. It includes seven taxa of interest : *Ceratophysella denticulata* and *Ceratophysella gibbosa* (CER), *Hemisotoma ther-*

mophila (HEM-THE), *Hypogastrura manubrialis* (HYP-MAN), *Lepidocyrtus cyaneus* and *Lepidocyrtus lanuginosus* (LEP), *Metaphorura affinis* (MET-AFF), *Isotomiella minor* (ISO-MIN), and *Parisotoma notabilis* (PAR-NOT), along with a category "Other" for Collembola from unannotated species. All Collembola specimens were sourced from 12 different projects, more details on the projects are available in the original paper [9]. Eleven of these projects were used for training, while BISES, the 12th and largest project, was reserved for evaluation to ensure the model did not train on previously seen projects. It is important to note that no instances of *Hypogastrura manubrialis* HYP-MAN were available in the BISES project so they won't be part of the experiments. Within the "Other" category, two species were extracted and annotated to create the new classes, *Pseudosinella alba* (PSE-ALB) and *Sphaeridia pumilis* (SPH-PUM), with 37 and 54 annotations on BISES , respectively. The Collembola of these species were hidden with a white square (Figure 2) in the training set to prevent the model from training on these new taxa while preserving other annotations due to limited data. (Table 1) refers to the number of annotations per species (including the new one) in the base dataset.

FIGURE 2 – *Example of obscured collembola using a white square.*

3.2 Active learning

This paper presents a novel active learning approach called "ALE" (Active Learning Extension), which enhances the current object detection active learning paradigm that typically focuses solely on prediction uncertainty. ALE introduces the consideration of the number of predictions, leading to substantial improvements, especially when new classes are introduced through active learning. Increasing the number of annotations, even with just a few additional ones, on new classes with very few annotations, can significantly improve the model's performance.

ALE. ALE core idea is that when incorporating a new class during active learning, solely relying on prediction uncertainty for image selection is not optimal. It is more effective to consider both uncertainty and the number of predictions. In state-of-the-art object detection active learning, prediction confidence is typically measured using metrics like entropy or least confidence. Each image is then assigned a score using aggregation techniques such as sum, maximum, or mean, and images with the highest score are selected. However, when adding new classes, prioritizing annotations with lower uncertainty might be more efficient than fewer annotations with higher uncertainty. To incorporate the number of predictions into image selection, an extension to the aggregation of uncertainty metrics is employed, as shown in Equation 6.

$$ALE = A(S) \times \sqrt{N} \tag{6}$$

Where $A(S)$ represents the aggregation function of uncertainty scores, and N is the total number of predictions on the image.

This paper introduces the square root as an extension. It was chosen empirically because it rapidly prioritizes more annotations. We tested various combinations of evaluation metrics, and aggregation techniques to determine the most suitable approach for active learning.

4 Experimental protocol and evaluation

4.1 Experiment

The experiment aimed to test our models as follows : First, we trained a base model using the eleven projects of Collembola as training data and the twelfth (BISES) as validation data. Then, we built multiple datasets using ALE with different aggregation and metric combinations, as well as state-of-the-art active learning techniques for comparison. Additionally, we created 3 random selections and used the one with the best overall performance to compare our technique against random sampling. All of these methods selected images from the validation set, which were added to the training set. The models were fine-tuned on the updated training datasets and evaluated on the new validation sets. The experiment was conducted three times, each with a different selection of images to be added to the training set : 20, 50, and 100 images.

4.2 Evaluation

Since the models were evaluated on different validation datasets, comparing them using their own validation results was not feasible. To accurately compare all models, we would need to use only the common images across all validation datasets, ensuring no model is evaluated on images it was trained on. However, this approach is impractical due to the number of models and the limited number of images. To address this, we evaluated the models in pairs by creating new datasets that include only the common evaluation images for each pair. This method allowed us to directly compare each model against another, determining which

TABLE 1 – *Distribution of Annotations in Training and Validation on the Base Dataset.*

	OTHER	CER	CRY THE	HYP MAN	ISO MIN	LEP	MET AFF	PAR NOT	PSE ALB	SPH PUM
train	410	228	126	109	87	144	106	111	0	0
valid	163	92	96	0	62	121	23	151	37	54

one performs better on their shared evaluation dataset. The evaluation metric used for these comparisons was the mean average precision (mAP).

4.3 Training protocols

Every model was trained until convergence with these parameters : an initial learning rate of 0.01 with a weight decay of 0.005, and the Adam optimizer with a beta1 of 0.937. Data augmentation techniques were applied during training to artificially increase the dataset size, enhancing the model's ability to generalize and improve accuracy on the test dataset. Various transformations were employed, including random crop, mosaic, and color distortions such as brightness, contrast, saturation, hue, Gaussian blur, random scaling, random rotation, and random horizontal flipping. These transformations introduce variations that enhance the model's performance and adaptability to different input scenarios.

4.4 Results

The results of the experiment, presented in Tables 2, 3, and 4, illustrate the following comparisons : The left column lists the baseline model trained on the original dataset, the best-performing model from random sampling, and various state-of-the-art models combining multiple metrics and aggregation techniques. the top row represent every ALE versions. Each model is labeled with an evaluation metric followed by an aggregation technique. For example, a model may be labeled as "Entr Sum" or "ALE LC Max," indicating, respectively, the use of the entropy metric with the "Sum" aggregation and the ALE approach with LC (Least Confidence) as the metric, using a maximum-based aggregation and square root extension.

The columns display the average results and variances obtained after ensemble training for each version of the active learning model, ensuring a more robust evaluation. The percentages presented indicate the average improvement in mean accuracy (mAP) compared to other techniques and are accompanied by the standard deviation of the results. This methodology helps capture the stability and robustness of each approach.

TABLE 2 – *Comparison of methods with 20 images active learning.*

	ALE Entr Max	ALE Entr Sum	ALE LC Max	ALE LC Sum
Baseline Model	**12.80%**	**14.80%**	**16.10%**	**11.50%**
Rand	**4.42%±0.48**	**4.44%±2.05**	**6.78%±1.27**	**2.96%±0.38**
Entr Max	**0.43%±0.26**	**0.73%±2.06**	**5.00%±1.68**	-0.73%±0.53
Entr Sum	-1.77%±0.30	-1.97%±2.17	**0.57%±1.40**	-3.53%±0.50
LC Max	**0.17%±0.31**	**0.10%±2.30**	**2.87%±1.56**	-1.03%±0.81
LC Sum	-0.43%±0.30	-0.60%±2.12	**2.00%±1.42**	-2.27%±0.50

TABLE 3 – *Comparison of methods with 50 images active learning.*

	ALE Entr Max	ALE Entr Sum	ALE LC Max	ALE LC Sum
Baseline Model	**22.10%**	**24.20%**	**25.10%**	**25.20%**
Rand	**5.16%±5.97**	**11.23%±4.66**	**7.44%±6.06**	**10.35%±5.05**
Entr Max	-0.65%±2.34	**4.72%±1.42**	**2.08%±1.07**	**4.02%±1.07**
Entr Sum	-3.45%±2.21	**1.75%±1.58**	-1.20%±1.07	**1.10%±1.19**
LC Max	-3.57%±2.22	**2.10%±1.47**	-0.72%±1.03	**1.45%±1.11**
LC Sum	-2.85%±2.22	**2.73%±1.51**	-0.40%±1.08	**1.90%±1.21**

TABLE 4 – *Comparison of methods with 100 images active learning.*

	ALE Entr Max	ALE Entr Sum	ALE LC Max	ALE LC Sum
Baseline Model	**32.60%**	**35.90%**	**33.20%**	**31.30%**
Rand	**5.22%±1.14**	**8.80%±0.54**	**6.32%±0.66**	**6.24%±1.42**
Entr Max	**2.33%±0.66**	**4.83%±0.12**	**3.77%±1.07**	**2.90%±2.19**
Entr Sum	-2.60%±0.75	**0.43%±0.13**	-1.43%±1.18	-2.20%±1.70
LC Max	-1.87%±0.84	**1.90%±0.15**	**0.20%±1.30**	**0.93%±1.20**
LC Sum	-1.60%±0.93	**1.77%±0.17**	0.00%±1.42	**0.50%±0.70**

When 20 images are selected (Table 2), "ALE LC Max" shows the best performance with a score of 16.10%, making it more effective than other ALE methods compared to the Baseline model. "ALE Entr Sum" (14.80%) and "ALE Entr Max" (12.80%) also demonstrate strong results. "ALE LC Sum," with 11.50%, is the least effective among the ALE methods tested. "ALE LC Max" is the only model that outperforms all state-of-the-art models and random sampling. The results from the 20-image selection indicate that even with a small image set, model performance increases significantly.

For the selection of 50 images (Table 3), "ALE LC Sum" achieves a score of 25.20%. "ALE LC Max" (25.10%) and "ALE Entr Sum" (24.20%) closely follow, suggesting that increasing the number of images significantly enhances performance. Two models outperform both state-of-the-art methods and random sampling : "ALE LC Sum" and "ALE Entr Sum."

With the selection of 100 images (Table 4), "ALE Entr Sum" achieves the best performance with a score of 35.90%, surpassing "ALE LC Max" (33.20%) and "ALE LC Sum" (31.30%). Once again, it outperforms all other state-of-the-art models and random sampling.

The results indicate that, in most cases, the ALE versions outperform their state-of-the-art counterparts. However, many of them still fall short of other state-of-the-art models, particularly "Entr Sum." Notably, "ALE Entr Sum" stands out, outperforming every other model in nearly every comparison, except in the 20-image sample, where it is surpassed by "Entr Sum" and "LC Sum."

We would expect the improvement of "ALE Entr Sum"

over "Entr Sum" to follow a linear trend as sample size increases. However, while there is a noticeable improvement from 20 to 50 samples, at 100 samples, even though "ALE Entr Sum" continues to lead, the difference in results between the two models becomes less pronounced. This can be explained by the fact that, in the BISES project, the number of images containing three or more annotations is 62 (Table 5). Consequently, "ALE Entr Sum" maximizes its effectiveness with a selection of 50 images (Table 3), capturing diversity without introducing redundancy. However, when the selection increases to 100 images, this distinction fades, as a larger number of images become common across "ALE Entr Sum" and "Entr Sum" (Table 6), reducing the advantage of the ALE.

TABLE 5 – *Number of annotations per image in the BISES project.*

Number of annotations per image	1	2	>=3
Number of image	414	69	62

TABLE 6 – *Progression of result comparison between the "ALE Entr Sum" and "Entr Sum" models, based on sample count and non-common images in the new class dataset.*

Number of samples	Number of non-common images	Performance comparison
20	3	-1,97%
50	6	1,75%
100	4	0,43%

In Figures 3 and 4, we can observe that while "ALE Entr Sum" does not consistently outperform "Entr Sum" across every class, it demonstrates a notably higher performance on both of the new classes, *Pseudosinella alba* (PSE-ALB) and *Sphaeridia pumilis* (SPH-PUM). This suggests that "ALE Entr Sum" effectively adapts to novel data, showing a significant advantage over "Entr Sum" in handling unfamiliar categories by adding more annotations, even if its overall improvement is not uniform across all classes.

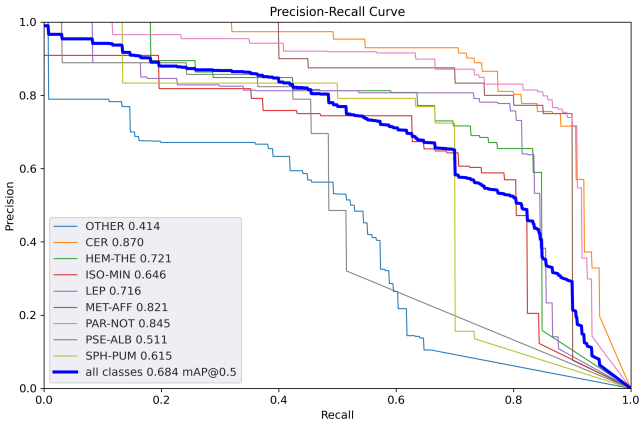

FIGURE 3 – *The AP and mAP of the top-performing "Entr Sum" version from the ensemble, evaluated on its shared dataset with the best "ALE Entr Sum" version from the ensemble (using the 50-image sample version).*

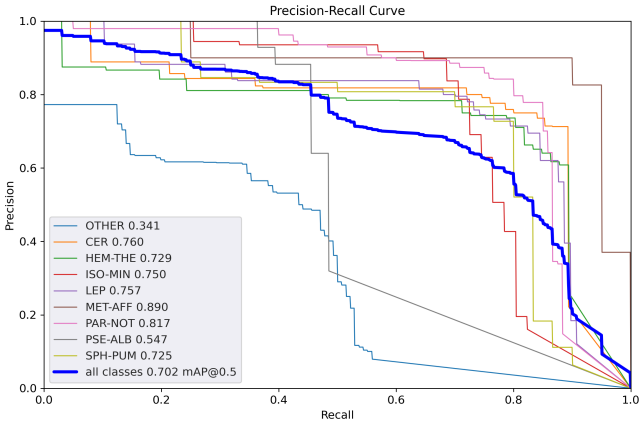

FIGURE 4 – *The AP and mAP of the top-performing "ALE Entr Sum" version from the ensemble, evaluated on its shared dataset with the best "Entr Sum" version from the ensemble (using the 50-image sample version).*

5 Conclusion

In conclusion, this study demonstrates the potential of ALE (Active Learning Extension) to enhance and expand model performance in ecological applications, particularly when adding new classes to an already trained model. ALE's approach, which combines uncertainty with prediction quantity, is especially effective on datasets containing images with varying numbers of predictions, enabling it to outperform traditional active learning methods. This capability allows ALE not only to improve model accuracy but also to extend its applicability by efficiently integrating new classes without necessitating a complete retraining on the entire dataset. In ecological monitoring, where precise species identification is essential, ALE enables the seamless addition of new taxa, allowing researchers to expand bio-

diversity assessments over time. This progressive approach supports adaptive model growth and offers an efficient, scalable solution to the costly, time-intensive process of annotating ecological data, making ALE a valuable tool for advancing biodiversity research and environmental monitoring.

6 Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

7 Acknowledgment

Théo Oriol is thankful for the financial support provided by the University of Paul Valery and the Occitanie Region. This work was granted access to the HPC resources of IDRIS under the allocation 20XX-AD011014028 made by GENCI

Références

[1] Dorothy Stone, Karl Ritz, BG Griffiths, Alberto Orgiazzi, et RE Creamer. Selection of biological indicators appropriate for european soil monitoring. *Applied Soil Ecology*, 97 :12–22, 2016.

[2] Anne Turbé, Arianna De Toni, Patricia Benito, Patrick Lavelle, Perrine Lavelle, Nuria Ruiz Camacho, Wim H van Der Putten, Eric Labouze, et Shaleindra Mudgal. Soil biodiversity : functions, threats and tools for policy makers. 2010.

[3] Jack H Faber, Rachel E Creamer, Christian Mulder, Jörg Römbke, Michiel Rutgers, J Paulo Sousa, Dorothy Stone, et Bryan S Griffiths. The practicalities and pitfalls of establishing a policy-relevant and cost-effective soil biological monitoring scheme. *Integrated environmental assessment and management*, 9(2) :276–284, 2013.

[4] Jean-François Ponge, Servane Gillet, Florence Dubs, E Fedoroff, Lucienne Haese, José Paulo Sousa, et Patrick Lavelle. Collembolan communities as bioindicators of land use intensification. *Soil biology and biochemistry*, 35(6) :813–826, 2003.

[5] José Paulo Sousa, Thomas Bolger, Maria Manuela Da Gama, Tuomas Lukkari, Jean-François Ponge, Carlos Simón, Georgy Traser, Adam J Vanbergen, Aoife Brennan, Florence Dubs, et al. Changes in collembola richness and diversity along a gradient of land-use intensity : a pan european study. *Pedobiologia*, 50(2) :147–156, 2006.

[6] T Larsen, Per Schjønning, et J Axelsen. The impact of soil compaction on euedaphic collembola. *Applied Soil Ecology*, 26(3) :273–281, 2004.

[7] Pascal Querner et Alexander Bruckner. Combining pitfall traps and soil samples to collect collembola for site scale biodiversity assessments. *Applied Soil Ecology*, 45(3) :293–297, 2010.

[8] Philippe JANSSEN, Marc FUHR, et Jean-Jacques BRUN. Effets de l'ancienneté du couvert forestier et de la maturité des peuplements sur la biodiversité des forêts de chartreuse, 2015.

[9] Théo Oriol, Jerome Pasquet, et Jérôme Cortet. Automatic identification of collembola with deep learning techniques. *Ecological Informatics*, 81 :102606, 2024.

[10] Stanislav Sys, Stephan Weißbach, Lea Jakob, Susanne Gerber, et Clément Schneider. Collembolai, a macrophotography and computer vision workflow to digitize and characterize samples of soil invertebrate communities preserved in fluid. *Methods in Ecology and Evolution*, 13(12) :2729–2742, 2022.

[11] Clemens-Alexander Brust, Christoph Käding, et Joachim Denzler. Active learning for deep object detection. *arXiv preprint arXiv :1809.09875*, 2018.

[12] Weiping Yu, Sijie Zhu, Taojiannan Yang, et Chen Chen. Consistency-based active learning for object detection. Dans *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3951–3960, 2022.

[13] JR van Bommel. Active learning during federated learning for object detection. *University of Twente Enschede : Enschede, The Netherlands*, 2021.

[14] Jiwoong Choi, Ismail Elezi, Hyuk-Jae Lee, Clement Farabet, et Jose M Alvarez. Active learning for deep object detection via probabilistic modeling. Dans *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10264–10273, 2021.

[15] Ying Li, Binbin Fan, Weiping Zhang, Weiping Ding, et Jianwei Yin. Deep active learning for object detection. *Information Sciences*, 579 :418–433, 2021.

[16] Joseph Redmon et Ali Farhadi. Yolov3 : An incremental improvement. *arXiv preprint arXiv :1804.02767*, 2018.

[17] Renjie Xu, Haifeng Lin, Kangjie Lu, Lin Cao, et Yunfei Liu. A forest fire detection system based on ensemble learning. *Forests*, 12(2) :217, 2021.