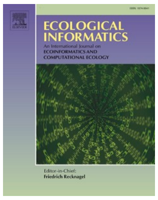
Automatic identification of Collembola with deep learning techniques

Théo Oriol, Jerome Pasquet, Jérôme Cortet

Contents lists available at ScienceDirect

Ecological Informatics

journal homepage: www.elsevier.com/locate/ecolinf



Automatic identification of Collembola with deep learning techniques

Théo Oriol [a,b,*], Jérôme Pasquet [b,c], Jérôme Cortet [a]

[a] Univ Paul Valéry Montpellier 3, Univ. Montpellier, EPHE, CNRS, IRD, CEFE UMR 5175, F34000, Montpellier, France
[b] AMIS, Université de Montpellier 3, Montpellier, France
[c] TETIS Inrae, AgroParisTech, Cirad, CNRS, Univ. Montpellier, Montpellier, France

ARTICLE INFO

Keywords:
Deep learning
Object detection
Collembola
Soil quality
Bioindication

ABSTRACT

Collembola are very abundant organisms in soils (several thousand individuals per square meter) and are considered to be good indicators of soil quality. These indicators are mainly based on the number of individuals observed (abundance per square meter of soil), but also the singularity and number of species present (species richness). A limitation that comes with the usage of collembola as an indicator is the complexity of the identification of the species under a microscope, how time-consuming it is, and the morphological similarity between some species. Deep learning approaches have been very successful in the resolution of image-based problems. Still, no work yet exists that uses deep learning in the recognition of collembola on a microscope slide. This could be a valuable tool for experts seeking to use Collembola as a metric on a larger scale. In this work, we explore and evaluate the performance of state-of-the-art deep learning techniques over the identification of Collembola on a new manually annotated dataset.

1. Introduction

Soil biodiversity is a crucial component of terrestrial ecosystems and represents up to 50% of the total biodiversity on earth (Anthony et al., 2023). Organisms found in soils contribute to many ecosystem services, like soil fertilization, crop protection, water cycle regulation, and water and soil decontamination. They thus have a central place in ecosystems, but are also sensitive to environmental modifications, particularly those affecting soils, like agricultural practices. It thus appears necessary to protect soil biodiversity, which requires monitoring it. Collembola, commonly known as springtails, are a class of intriguing tiny arthropods belonging to the subphylum Hexapoda. With several thousand individuals per square meter of soil, they represent a considerable biomass. They are a key component for ecosystem functioning, like nutrient cycling or soil aggregation. Like other soil organisms, they are sensitive to changes in soil properties, such as soil moisture, temperature, pH, and nutrient availability. Their presence and/or diversity indicates soil quality and degradation, making them valuable tools for monitoring the effects of agricultural and forest practices (Cortet et al., 1999) and soil pollution (Fountain and Hopkin, 2004; Heisler and Kaiser, 1995). These indicators may include both the abundance of individuals per square meter of soil and the richness of species present. Unfortunately, to identify Collembola, many steps are required, including collection of soil cores, extraction of living species from the soil cores, separation of Collembola from other taxa, and mounting them on microscope slides for species identification. Due to their morphological similarity, identifying Collembola through a microscope is a highly intricate and time-consuming task that demands significant expertise. This is a major obstacle to the development and mass use of Collembola as bio-indicators since the low number of available taxonomists contrasts with the fact that the datasets available in ecology are getting larger (Deharveng, 2004). To overcome the challenges of identifying Collembola, modern technology offers a promising solution. Deep learning, a branch of artificial intelligence, has emerged as a powerful tool in ecology and biodiversity research. Over the last few years, deep learning models have emerged as the state-of-the-art approach in computer vision, consistently demonstrating superior performance across various tasks and benchmarks. The use of machine learning in ecology is not new (Crisci et al., 2012), it already has been used for tasks such as ecological modeling (Recknagel, 2001), the study of animal behavior (Arablouei et al., 2023), and species identification (Waldchen and Mader, 2018). What makes deep learning such a growing field is the recent availability of powerful hardware and large amounts of training data. By learning from the given example deep learning models can extract important features and resolve a task without being specifically programmed to resolve it (LeCun et al., 2015), which makes

* Corresponding author.
E-mail address: theo.oriol@univ-montp3.fr (T. Oriol).

https://doi.org/10.1016/j.ecoinf.2024.102606
Received 20 December 2023; Received in revised form 15 March 2024; Accepted 17 April 2024
Available online 24 April 2024

it an ideal solution to large amounts of data. Manual analysis of a vast dataset is time-consuming for experts. Automating this task saves a lot of time and enables monitoring to be used on a larger scale, (Rustia et al., 2021; Schneider et al., 2022; Spiesman et al., 2021), helping to compensate for the small number of taxonomists especially for data that can be complex to analyze like images (Minaee et al., 2021) or videos (Liu et al., 2020). The use of deep learning in computer vision for ecology has dramatically improved in the past few years, the PlantCLEF challenges are a good example (Waldchen and Mader, 2018). Every year it provides a large and complex image dataset to uncover and evaluate state-of-the-art machine learning models, and every year the result of the identification performance improves despite the task becoming more complex. By identifying the most frequent and common species of Collembola in agricultural soils, often relatively poor in diversity, deep learning would allow the mass use of this indicator. It could allow experts to save time by focusing their attention on less frequent species. Microarthropods (including Collembola and Acari) identification and analysis of Collembola with deep learning has already been tried before (Kampichler et al., 2000; Sys et al., 2022), but even though optical microscopy is not something new in deep learning, learning, Collembola identification through microscopic slides has never been done before. This study aims at the investigation of the effectiveness of deep learning on species identification via microscope slide images, particularly in the context of Collembola identification. In this context, two hypotheses are presented. The first hypothesis suggests that a state-of-the-art deep learning model can accurately detect and identify most Collembola at the species level mounted on microscope slides. The second hypothesis is that this model relies on distinctive Collembola features rather than excessively fitting to background image elements.

2. Material and methods

2.1. Digitization and annotation

Since no previous work existed on the identification of Collembola on microscope images with deep-learning techniques, creating an image-annotated dataset was required before going any further. To collect Collembola, it is necessary to extract soil cores from the chosen analysis site, then with the use of some extraction device called Macfadyen (Potapov et al., 2020) we extract specimens to some fixation liquid so that we can separate the Collembola from other taxons we find. Finally, we proceed to depigment and mount them on a microscope slide with a mounting medium called "Marc-andré" (Milano et al., 2018). To gain some time, we used already mounted Collembola, which were already available thanks to multiple research projects concerning the same study object (Joimel et al., 2017). The creation of the dataset required identifying, taking photos of Collembola, and annotating them. Ten species of interest, common and known to be abundant in agricultural soils, were chosen to be automatically identified with deep learning as a proof of concept: *Ceratophysella denticulata* (Bagnall, 1941) (CERDEN), *Ceratophysella Gibbosa* (Bagnall, 1941) (CER-GIB), *Hemisotoma thermophila* (HEM-THE) (Axelson, 1900), *Hypogastrura manubrialis* (Tullberg, 1869) (HYP-MAN), *Lepidocyrtus cyaneus* (Schille, 1908) (LEP-CYA), *Lepidocyrtus Lanuginosus* (Gmelin, 1788) (LEP-LAN), *Metaphorura affinis* (Börner, 1902) (MET-AFF), *Isotomiella minor* (Schäffer, 1896) (ISO-MIN) and *Parisotoma notabilis* (Schäffer, 1896) (PARNOT). These species have been chosen to identify if state-of-the-art models of deep learning can detect and identify the interspecies morphological similarity and the intra-species morphological variance of Collembola. To give examples of similarity, the species *Parisotoma notabilis* and *Hemisotoma thermophila* are easy to differentiate using features only visible on a high zoom (x630), but they have very similar morphology, on a low zoom (x50). It is also the case of *Ceratophysella denticulata*, *Ceratophysella Gibbosa*, and *Hypogastrura manubrialis*, which makes both of these groups of species hard to differentiate. For practical reasons, the species *Lepidocyrtus cyaneus* and *Lepidocyrtus Lanuginosus* were fused as LEP since the main difference between these two species is their coloration, which degrades itself over time once they are mounted on slides, and since there is a low number of annotations, it was decided to merge them. *Ceratophysella denticulata*, and *Ceratophysella Gibosa* were fused as CER due to the low amount of data available in our datasets for the models to be able to differentiate between them. *Ceratophysella denticulata* and *Ceratophysella Gibbosa* were merged into a single category, designated as CER, because of the scarcity of annotations for *Ceratophysella Gibbosa* in our datasets, allowing for its use as a species of interest. It is also important to show images of other Collembola than these 10 species to our model of deep learning because, in a realistic scenario, it will have to differentiate them from other unknown species. To do so, we added a new category called "Other", where we regrouped Collembola of multiple species not included in the species of interest. To avoid bias, all species of Collembola were sampled on multiple projects, which allows us to use recent microscope slides, the most recent one being from projects in 2021, and older microscope slides, the oldest projects being from 2003 (Table 1). With these, we can train the model to work on new as well as older samples of lesser quality, allowing experts to work on older projects. To create the dataset, multiple steps were required: first, the Collembola were identified by an expert with a microscope, then taken in photos, digitalized with the software proview that also applies a white balance and an automatic exposure to make the image clearer, and finally annotated with the software labelimg (Fig. 1), it should be noted that there was no specific protocol delineating the position of each specimen to maintain consistency with expert practices. Variability in positioning was not controlled. The microscope used was a Carl Zeiss labscope A1 equipped with an Optika microscope C—P6 camera to take the pictures. All of them have the same dimensions 3072 × 2048 pixels for 3 × 2 mm and were taken with a x50 magnification. In total 1664 different pictures were taken for 2195 different Collembola identified and annotated (since there can be multiple Collembola per image, there are fewer images than annotations).

2.2. State of the art

To assess the ability of deep learning, and object detection models, specifically in detecting and identifying Collembola, multiple state-of-the-art models have been chosen to create a benchmark. Faster R-CNN and Yolov5 represent two different approaches, a two-stage detector, and one stage detector, respectively.

2.2.1. Yolov5

Yolov5 is well known for achieving state-of-the-art performance on several benchmarks including the coco dataset. It is one of the most popular object detection model approaches and was developed by Ultralytics as an extension of Yolov3 (Redmon and Farhadi, 2018) with improved speed and precision. Yolov5 is a one-step detector, meaning that it detects and classifies objects simultaneously. It is composed of a backbone (CSPDarknet), a neck and a prediction head (Fig. 2). The backbone extracts feature from the image that are mixed and combined for prediction by the neck, and then the detection head takes it as input to propose boxes and classes. To generate the proposition, the image is divided into multiple different grids of multiple scales and each cell will propose N objects. Using anchors to make box coordinate predictions and different scale aspect ratios Yolov5 achieves precise detection. Anchors are used to facilitate the prediction of coordinates, which are crafted using different ratios and sizes based on the data they are supposed to fit, in our case Collembola. Instead of directly predicting Collembola coordinates, Yolov5 predicts in which cell the center of the Collembola is along with the height and width ratio of the anchor used to predict it. Doing so simplifies the task by narrowing down the range of prediction from 0 to N (N being the size of the image in pixels) for each coordinate, to a more focused range of 0 to 1, greatly improving the accuracy of the model coordinate predictions. 4 different versions of Yolov5 were used in the benchmark, Yolov5n, Yolov5m6, Yolov5l6 and

Table 1
Number of annotations per species per project.

Project	Date	Location	Publications	CER	HEM-THE	HYP-MAN	ISO-MIN	LEP	MET-AFF	PAR-NOT	OTHER	Total
BISES	2021	200 sites in urban contexts within 4 cities in France (Nancy, Montpellier, Paris, Nantes)	Unpublished	93	93	0	61	121	23	154	255	800
Bioindicateur 2	2009/2010	10 sites (Croplands, pastures and forests) in several regions, France	(Cortet, 2012; Pérès et al., 2011)	83	1	0	4	8	0	54	181	331
RMQS Biodiv	2006/2007	97 sites (croplands, pastures, and forests) within the whole Brittany Region, France	(Cluzeau et al., 2009; Cluzeau et al., 2012; Ponge et al., 2013)	1	16	0	5	8	0	6	0	36
These Celine Pernin	2003	Forest firebreak in the Maures Massif, Provence Region, France	(Pernin et al., 2006)	1	0	0	0	0	0	0	0	1
Ecopic	2018/2019	Forests along an altitudinal gradient in the French Alps (Chamrousse, Auvergne-Rhone-Alpes Region, France)	Unpublished	23	0	0	11	56	0	5	22	117
Howecourt	2003	Brownfield in Homecourt, Grand-Est Region, France	Unpublished	0	0	0	0	8	0	0	1	9
NARBONS	2003	Maize fields in Montesquieu-Lauragais, Occitanie Region, France	(Cortet et al., 2007)	0	0	0	11	6	0	0	0	17
These Benjamin Pey	2007/2018	Brownfield in Homecourt, Grand-Est Region, France	(Joimel et al., 2021)	9	9	0	0	0	0	0	0	18
Parcelles Lysimetriques	2007	Brownfield in Homecourt, Grand-Est Region, France	(Cébron et al., 2011; Ouvrard et al., 2011)	0	30	0	0	16	0	0	6	52
Pompey	2007	Forest on a former settling pond, Frouard, Grand-Est, France	(Huot et al., 2018)	0	68	14	3	6	0	0	93	184
TIDM URBA	2021	83 sites within the city of Dijon, Bourgogne-Franch-Comte, France	Unpublished	77	0	95	53	32	106	48	162	573
VADEBIO	2010	Cropland in La Bouzule, Grand-Est Region, France	Unpublished	33	0	0	0	3	0	0	21	57
Total				320	217	109	148	264	129	267	741	2195

Yolov5x6. Yolov5n has 1.9 M parameters and is pre-trained on images with a dimension of 640 × 640 pixels Yolov5m6, Yolov5l6, and Yolov5x6 have respectively, 35.7 M, 76.8 M, 140.7 M parameters, they are pre-trained on images with a dimension of 1280 × 1280 pixels.

2.2.2. Faster R-CNN

Faster R-CNN (Ren et al., 2015) an extension of the R-CNN series (Girshick et al., 2014). An evolved version that is more precise, faster, and widely used as a reference in object detection benchmarks. Faster R-CNN is a two-step detector, the model first generates propositions and then classifies them. It is composed of a backbone, a region proposal network (RPN), a region of interest pooling (ROI), and a classification head. The backbone extracts feature from images, the RPN generates object candidate, and the ROI connects the backbone and the RPN to the classifier head, which classifies the objects. To generate the final propositions, the RPN uses the backbone feature and an anchor-based system of sliding windows, to generate N object propositions. The ROI connects the image feature extraction backbone and the object classifier head, facilitating the classification of objects by aggregating and processing relevant regions within the images. ROI pooling aids in the selection and analysis of specific image regions for accurate object classification. 2 different versions of Faster R-CNN were created for the benchmark, Faster R-CNN (640) with an input dimension of 640 × 640 and Faster R-CNN (1280) with an input dimension of 1280 × 1280. They both used Resnet50 as a backbone, have 41 m parameters, and are pre-trained on ImageNet. Both those dimensions were chosen to match the Yolov5 input dimensions. Still, after multiple tests, the 640-version had inconclusive results. This can be explained by the important features being too small using this resolution for Faster R-CNN. Only the 1280 version was used in the benchmark.

2.3. Evaluation

The evaluation of our models was a crucial step in analyzing the ability of the state-of-the-art deep learning models to detect and identify Collembola on microscope slides. To do so we proceed as follows, first, we made predictions on an evaluation dataset that we've extracted from the main dataset. This ensures that the model has never encountered this particular dataset during its training phase. The predictions-annotations matching is done by calculating the IoU metric Eq. 1 for each possible pair and matching them based on it. The IoU Eq. 1 is a metric that quantifies the overlap between the annotation and the prediction box. We considered that an IoU of 0.5 between the prediction and the annotation is a match. If a prediction didn't match any annotation human annotation, it was considered to be a background element.

$$IoU = \frac{Area\ of\ Intersection}{Area\ of\ Union} \tag{1}$$

The second step was to obtain the metrics we needed to evaluate those predictions. To do so, we defined 3 variables: True positive, the prediction coordinates match a ground truth and the species predicted is correct. False positive, the prediction coordinates match a ground truth but predict the wrong species or the coordinates are not matched with any ground truth, meaning that a background element was confused with a Collembola. False negative, the ground truth is not matched with any predictions, suggesting that the model confused a Collembola as a background element. Based on these variables we calculated more global metrics. The Recall Eq. 2 is a metric that quantifies the ability of a model to correctly identify all positive instances of a dataset. The higher the recall for a species, the more confident the model is in finding all the Collembola of this species. The precision formula 3 measures the ability of the model to be correct when predicting a species. The higher the precision for a species, the more confident the model is to be correct when predicting it. From these two metrics, we calculated a more advanced metric, the precision-recall curve, which is a graph representing the balance between the recall and the precision, depending on

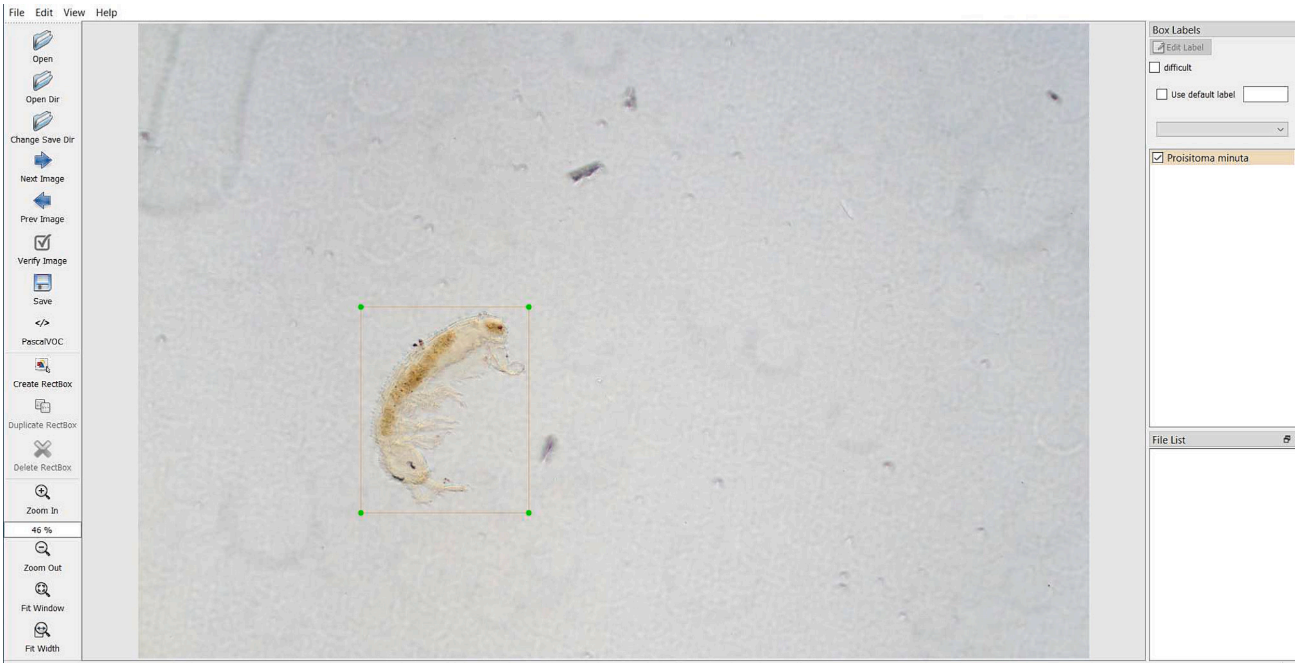


Fig. 1. Example of annotations of Collembola with labelimg.

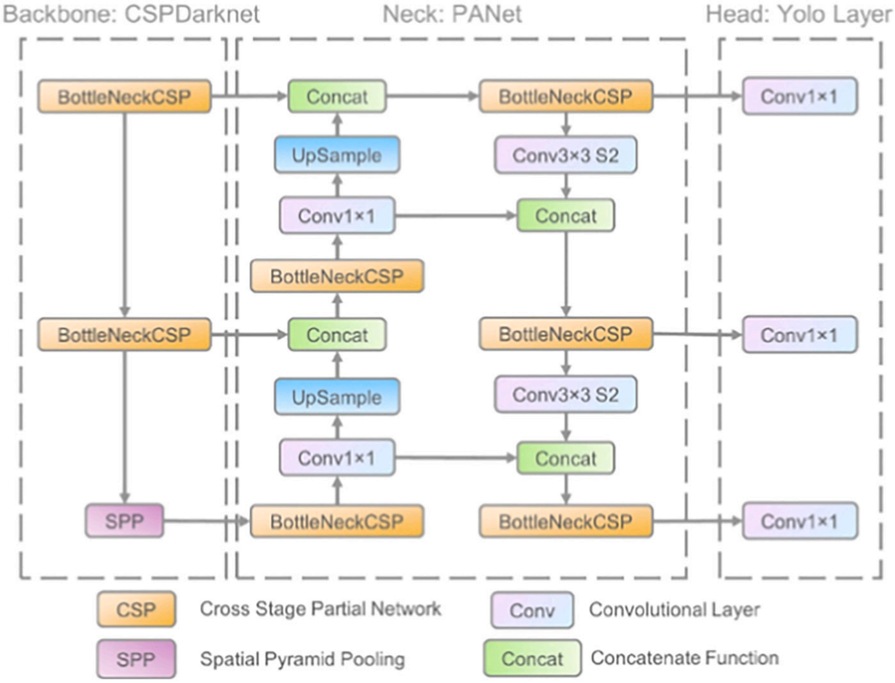


Fig. 2. Yolov5 architecture (Xu et al., 2021).

the confidence of the predictions.

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

With this metric, we can observe the impact of confidence on the results of the model. Using the precision-recall curve, we can calculate the AP Eq. 5 (Average Precision), which represents the area under the curve, and quantify the model result on one species, as a single number. Since we had multiple species, we also used the mAP Eq. 4 (Mean average precision) which gives a score for the whole model whereas the AP will give one for each one of them. The objective of these models is to be reliable enough to be used by experts in real-life conditions. This means that we had to be confident enough in the model prediction to use it. We allowed ourselves 5% of mistakes in the prediction of each species of interest. We created a benchmark to compare the results of state-of-the-art models of deep learning on this task. To create a benchmark of models, we use the mAP to analyze which model does best on average for each species and the AP to analyze how good it is at predicting each species. To ensure the model's ability to reach 95% precision, we used the precision-recall curve that allowed us to find the threshold of confidence with which for a recall of N, there is at least 95% precision per species.

$$mAP = \frac{1}{C}\sum_{i=1}^{C} AP_i \tag{4}$$

where C is the number of species.

$$AP = \int_0^1 precision(r)\, dr \tag{5}$$

where r represents recall, and precision(r) represents the *precision at a given recall level r.*

Because of the low amount of data at our disposal, cross validation was used to ensure the statistical results, and that the model was not over-fitting. Over-fitting happens when a model is too complex for the amount of data given and loses the ability to learn a solution that generalizes well because it fits the training data too much, resulting in poor performance on unseen data. The cross-validation technique we used works as follows: K crosses are created by randomly shuffling the dataset and dividing it into a training and an evaluation set K times, then for each cross, a new model is trained. This allows us to train K different models, each one on a different version of the dataset. Once all the models are trained, we evaluate them and average all of their performances to better understand the model's capability on this dataset. The similarity of each model's evaluation results would indicate that the models are less likely to be overfitting on the data, therefore showing that they are more likely to be capable of generalizing this task on unseen data. The opposite would be if the models achieve highly different results, suggesting that they are very dependent on the data they were trained on and might generalize poorly to unseen data. In our case, each model used a cross-validation of 5 crosses.

2.4. *Training protocols*

The models were all trained with the same parameters: 500 epochs, an initial learning rate of 0.01 with a weight decay of 0.005, the optimizer Adam was chosen, with a beta1 of 0.937, and data augmentation transformations were applied while training. Data augmentation is a technique used in machine learning to artificially increase the size of a dataset by creating new samples from the existing ones. Augmented samples are used to train the model more effectively by increasing its ability to generalize and its accuracy on the test dataset. The advantage of this technique when dealing with a low amount of data such as in our case, is the reduced risk of overfitting, since models are exposed to more variations, they will tend to less memorize the dataset. This becomes essential when you have limited data to train models with, which is a common problem in deep-learning applications. Another advantage is the robustness it gives to the model, by applying variation it makes models more reliable on a wider range of input data, which in our case would make it more reliable on images from old microscope slices which tend to be in a bad state. Finally, it also saves time by reducing the need to collect a bigger dataset. Various transformations can be applied to create new samples. In the case of images, simple modifications can be applied such as flipping or rotating, and more complex ones like color distortion or random crop. Here are the data augmentation techniques used in training: Random crop, Mosaic, and Color distortions such as brightness, contrast, saturation, hue, Gaussian blur, Random scaling, Random rotation, and Random horizontal flipping. Data augmentation helps recreate the state of old microscope slides on new ones (Fig. 3), which improves results on old projects.

2.5. *Model bias*

To test the hypothesis that the models are using the Collembola feature to make a prediction, we created two experiments. In the first one, we used a Grad-Cam (Selvaraju et al., 2017) to observe where the model looks on the image to make a prediction, in the second experiment we modified the image background of Collembola to see if the model is still capable of predicting if the noise from the image can't indicate the specimen.

2.5.1. *Grad-cam*

Grad-cam (Selvaraju et al., 2017) is a technique used to create a heatmap representing where the model focuses on the image to make a decision. It uses the feature maps generated during the inference of the model and the respective gradient to show which part of the image was used in the identification. The objective was to use this technique on the best model of the benchmark to analyze if the model is capable of using features of Collembola when making a prediction.

2.5.2. *Background*

To be sure that the model is not using anything else than the Collembola feature to make a prediction, we changed the Collembola background from their original background to a background from a different project where their species are either non-existent or in minority Fig. 4. We then analyzed if the prediction changes depending on the background change. For time purposes, only 10 specimens per taxon were cropped, so 70 in total.

2.6. *Comparison with experts*

To better understand the model's potential, we compared its prediction ability with experts on images from the validation dataset. Two experts, who identify Collembola daily agreed to participate in the test. The constraint imposed on the expert was to predict with the same image the model uses. For the experts, this differs from the annotation phase since they didn't have direct access to the specimens, they couldn't

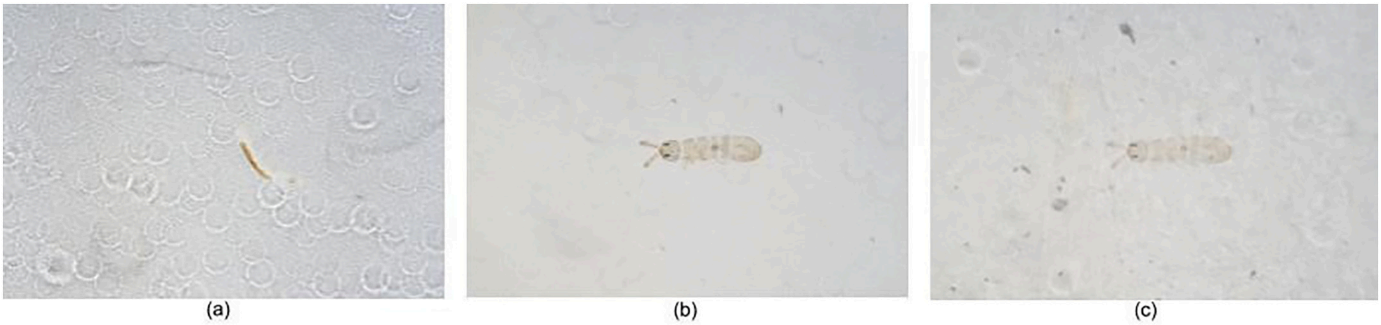


Fig. 3. Image (a) represents a microscope slide photo from an old project, the slide is in a bad state with a lot of noise on the image, and features of the Collembola are not all visible. Image (b) represents a microscope slide photo from a recent project in a good state, thanks to data augmentation we can add noise onto image (b) to create image (c), which will create more images of slide in a bad state and improve model results on old projects.

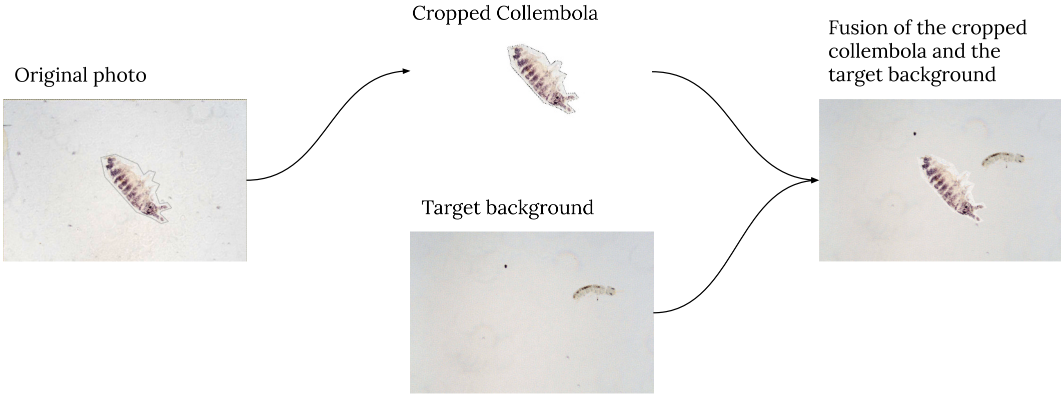


Fig. 4. We change the Collembola from its original background to the target background by cropping the Collembola from the original image and pasting it to the target background.

zoom on specific details (they only had access to the x50 magnification of the image), they couldn't modify the light or know from which project the Collembola come from.

2.7. Hardware specification

All the models were trained on a server equipped with 256Go of RAM 4 GeForce RTX 2080Ti and Processor Intel(R) Xeon(R) Gold 5222 CPU @3.80GHz. The server OS is Ubuntu20.04.6 LTS, and the version of PyTorch used was 1.13.1.

3. Results

Table 2 results highlight the performance of each model with their respective mAP and AP per class. Yolov5x6, the largest Yolov5 version, with a map of 0.894, outperforms all other models, all species combined, while Faster R-CNN with a mAP of 0.656 is the worst-performing model. It is interesting to note that the inference time of Yolov5x6 on one image with our hardware is 0.11 s including the post-processing of the prediction, while an expert takes between 10 s on average for a common species to 5 min or more for a less common one.

3.1. Analysis Yolo

Table 2 shows that Yolov5x6 is the best benchmark model on average and for each species of interest. Against intuition, the strength of the model is on species with fewer annotations. Metaphorura affinis has an AP of 0.991 when the total amount of annotation for this species was only 129, compared to Parisotoma notabilis with 267 annotations and an AP of 0.910. On the other hand, *Isotomiella minor*, which, like Metaphorura affinis, has a limited number of annotations, performs comparatively less well, with an AP of 0.807. Overall, it is evident that the model excels in identifying Collembola from the species of interest, especially when considering the challenge posed by the varying number of annotations per species. With the matrix of confusion shown in Fig. 5, we can analyze what kind of mistakes the model makes when making identification. When identifying *Isotomiella minor*, its weakness, the model makes 2 types of mistakes. It either confuses the Collembola with the background or with a Collembola from the category “Other”, this represents 15% of the error made, and the second type of mistake is the confusion between, *Isotomiella minor*, Parisotoma notabilis, and Hemisotoma thermophila which represents 6% of mistakes made. Unknow Collembola from the category “Other” are also confused with species of interest, up to 18% of unknown Collembola are miss-classified, and the errors are spread over all the species of interest. Since we aim to obtain a precision of at least 95% per species, we use Fig. 6 to analyze the recall we can obtain depending on the precision. Table 3 refers to different levels of recall and the precision per species associated with them. Since there is a balance between precision and recall, more recall is equal to less precision, with the necessary precision of 95% the recall will be of at least 20%. Augmenting the recall drastically reduces the precision for each species except for Metaphorura affinis and the taxon CER. The precision mostly depends on the level of recall expected from the scientists.

3.2. Model bias

3.2.1. Grad-cam

The following results (Fig. 7) are achieved using a Grad-Cam on Yolov5x6 the best model of the benchmark 2, on every image from the validation dataset. The model tends to focus on key features of Collembola identification like their antennas, ocular fields, torso or anal spines. It also focuses a lot of attention on the background, specifically for the species Parisotoma notabilis and *Isotomiella minor*.

3.2.2. Background

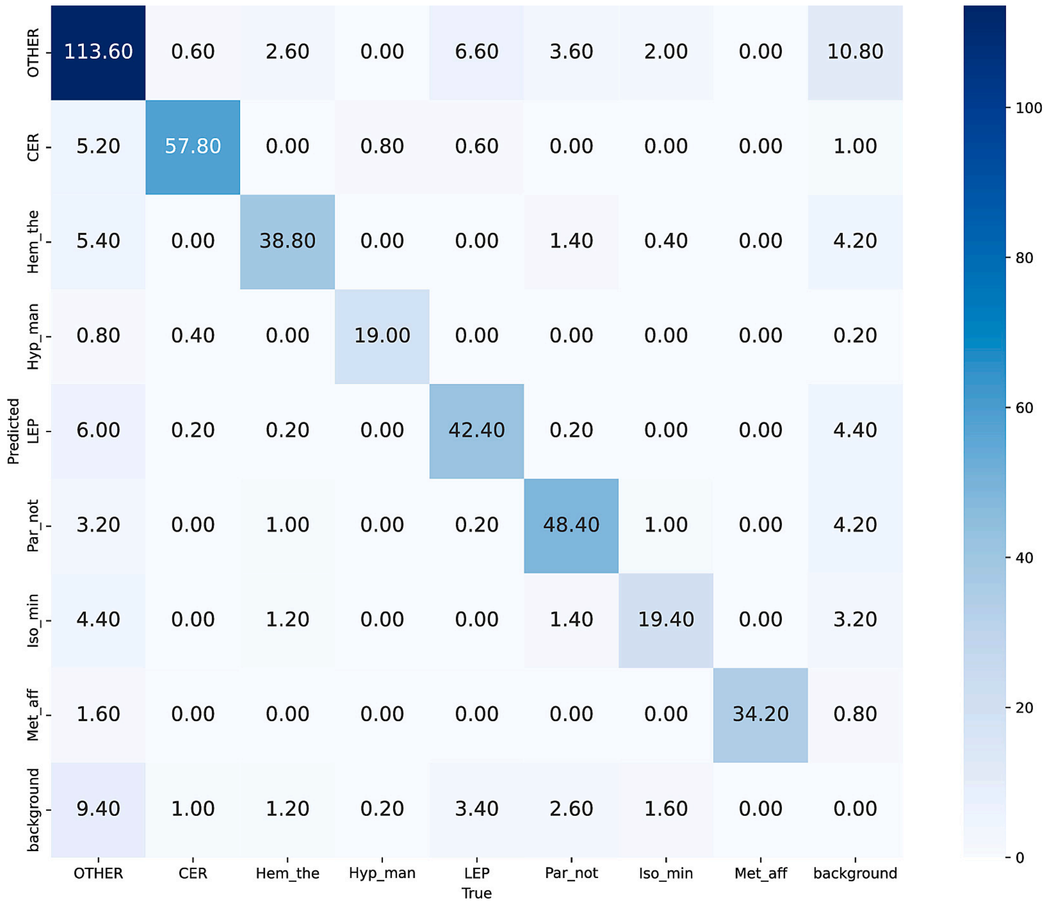
70 specimens were cropped to a different background Fig. 4 and identified. For most of them, the model is capable of correctly identifying the species without using the background noise, but two species of interest have worse results than the other if we prevent the model from using it. Parisotoma notabilis and *Isotomiella minor* both represent 7 out of 13 mistakes the model made as we can see in Table 4. They seem to be harder to identify without the original project background.

3.2.3. Comparison with experts

During the expert analysis of the dataset images, attempts were made

Table 2
Benchmark of the different models. The best mAP and the best AP per species were written in bold.

	mAP	“OTHER” AP	“CER” AP	“HEM-THE” AP	“HYP-MAN” AP	“LEP” AP	“PAR-NOT” AP	“ISO-MIN” AP	“MET-AFF” AP
Faster R-CNN (1280)	0.656	0.632	0.83	0.587	0.718	0.652	0.525	0.417	0.887
Yolov5n	0.802	0.709	0.916	0.772	0.894	0.772	0.778	0.659	0.916
Yolov5m6	0.869	0.779	0.926	0.841	0.943	0.837	0.885	0.775	0.966
Yolov5l6	0.870	0.793	0.944	0.845	0.935	0.843	0.843	0.794	0.966
Yolov5x6	**0.894**	**0.803**	**0.947**	**0.891**	**0.948**	**0.855**	**0.910**	**0.807**	**0.991**



Predicted \ True	OTHER	CER	Hem_the	Hyp_man	LEP	Par_not	Iso_min	Met_aff	background
OTHER	113.60	0.60	2.60	0.00	6.60	3.60	2.00	0.00	10.80
CER	5.20	57.80	0.00	0.80	0.60	0.00	0.00	0.00	1.00
Hem_the	5.40	0.00	38.80	0.00	0.00	1.40	0.40	0.00	4.20
Hyp_man	0.80	0.40	0.00	19.00	0.00	0.00	0.00	0.00	0.20
LEP	6.00	0.20	0.20	0.00	42.40	0.20	0.00	0.00	4.40
Par_not	3.20	0.00	1.00	0.00	0.20	48.40	1.00	0.00	4.20
Iso_min	4.40	0.00	1.20	0.00	0.00	1.40	19.40	0.00	3.20
Met_aff	1.60	0.00	0.00	0.00	0.00	0.00	0.00	34.20	0.80
background	9.40	1.00	1.20	0.20	3.40	2.60	1.60	0.00	0.00

Fig. 5. Yolov5x6 confusion matrix, each line represents the actual class labels, each column represents the predicted class labels, and the sidebar serves as a visual aid to interpret the intensity or magnitude of the values represented in the matrix. The cells of the matrix show the predictions made and the actual species of the specimens, providing a detailed assessment of a model's performance across multiple classes.

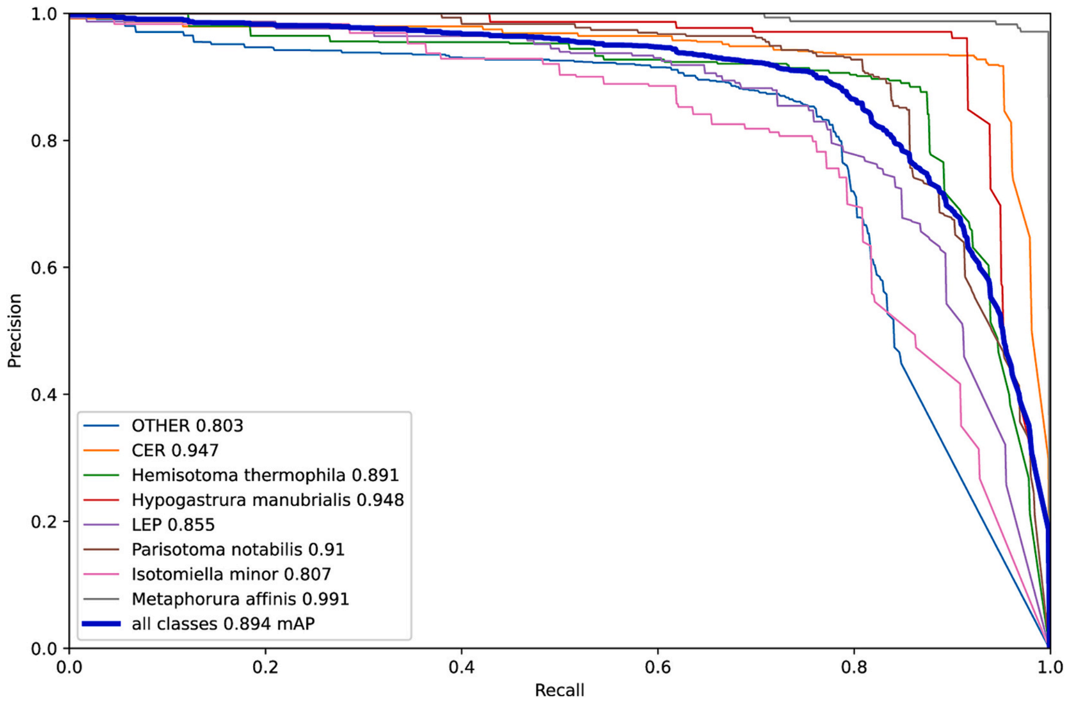


Fig. 6. Yolov5x6 precision-recall curve, the AP of each species is the area under their curve, the MaP is the mean of all the AP.

Table 3
Precision per species for 20%, 50%, 80% and 95% recall.

Species / Recall	20%	50%	80%	95%
OTHER	0.950	0.925	0.717	0.145
CER	0.979	0.968	0.935	0.917
Hemisotoma thermophila	0.964	0.952	0.903	0.448
Hypogastrura manubrialis	1	0.986	0.970	0.573
LEP	0.987	0.939	0.777	0.333
Parisotoma notabilis	1	0.983	0.927	0.472
Isotomiella minor	0.983	0.903	0.696	0.184
Metaphorura affinis	1	1	0.987	0.982

to identify Collembola species. However, it became evident that the identification process encountered a significant obstacle due to the low zoom levels in the images. The stated reason was that crucial details for identification were impossible to discern using this magnification, as shown in Fig. 8.

4. Discussion

In this paper, we achieve Collembola detection through deep-learning models, we delve into the identification of Collembola species on microscope slides, including the training of state-of-the-art models Yolov5 and Faster R-CNN (Ren et al., 2015), and the creation of a dataset of Collembola for object identification. Our results present the challenge of species identification on images taken with a microscope with a focus on Collembola. The primary objective was the evaluation of Yolov5 and Faster R-CNN performance on identifying Collembola on microscope slides. A dataset of 2195 annotations was built to achieve such training, including 9 species of interest and a category “Other”. Results outcome (Table 2) clearly shows the superiority of Yolov5 over Faster R-CNN by a substantial margin and its ability to identify Collembola on microscope slides. The intuition behind it is the use of different resolutions from the image that Yolov5 uses to make identifications, this would make the model use different features, from different size while identifying Collembola. Collembola identification is a complex task due to several unique challenges related to their morphological characteristics

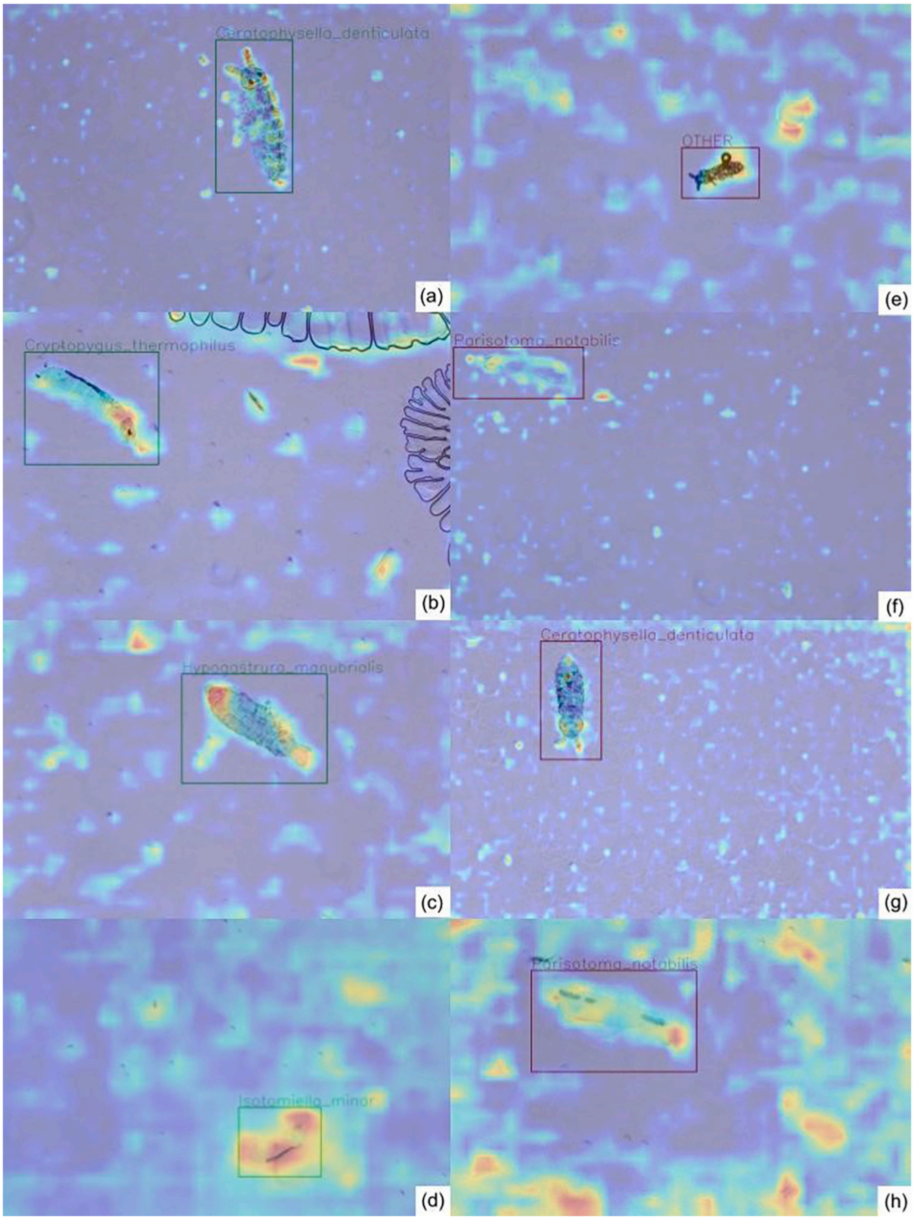


Fig. 7. From (a) to (d) Grad-Cam of correct identifications made by Yolov5x6, from (e) to (h) Grad-Cam of wrong identifications.

Table 4
Number of errors for ten Collembola per taxon when changing the background of specimen.

Species	Number of errors out of 10 specimens
CER	1
Hemisotoma thermophila	0
Hypogastrura manubrialis	1
Isotomiella minor	3
LEP	1
Metaphorura affinis	1
Parisotoma notabilis	4
Total	11

(Deharveng, 2004). The intra-species variance, the interspecies similarity, and the morphological traits making identification possible sometimes being only visible with a high zoom. Yolov5x6, the largest version of Yolov5 in our benchmark (Table 2), exhibited remarkable performance, outperforming Faster R-CNN by a large margin, which is not surprising considering that Yolov5 usually does better than Faster R-CNN (Fang et al., 2021; Hussain et al., 2021; Tan et al., 2022; Wang and Yan, 2021). It could be due to its number of parameters, the dimension of images used as input, and the added resolution output. It is interesting to note that, the difficulty of identification tends to vary with species, some of them being much easier to identify than others. Metaphorura affinis, the least challenging species, achieved an impressive AP of 0.991 despite a low amount of annotation. While Parisotoma notabilis, with more than twice the number of annotations, has a lower AP of 0.910. This demonstrates that some taxa are inherently harder to identify, and their features tend to be more difficult to find by the model, they need more annotations to improve their results. *Isotomiella minor* for example, is the most challenging species of our model with an AP of 0.807 and also is one of the species with the lowest number of annotations, increasing the number of annotations and using techniques that focus on the variance intra-species and similarity inter-species would greatly benefit its results. In addition to the unique challenges posed by species like *Isotomiella minor*, an analysis of the confusion matrix sheds light on common sources of model errors. Misclassifying Collembola from the species of interest into the "Other" category, often due to an inter-species similarity. Some Collembola were also confused with the background, most of the time because of the poor condition of the slide and the specimens. In the context of this paper, the misclassification of species of interest into the category "Other" is only a minor concern, the main objective is to reach a precision of at least 95% on each species of interest. Every specimen classified as "Other" will be verified by an expert, so making each mistake in this category will be corrected by a specialist. The level of precision mainly depends on the threshold of confidence chosen by the experts. The threshold of confidence affects the balance between the precision and the recall, the higher the confidence, the higher the precision, and with more precision the prediction will be more accurate. With lower confidence, the recall is higher, and more Collembola will be found, which will result in a higher gain in time for the experts. The experts choose the balance between gaining time and trusting the predictions. To illustrate the feasible recall while achieving our objective of 95% precision we use the precision-recall curve (Fig. 6). We observe the precision for different thresholds of recall 3. When reaching 95% precision for each species, we obtain at least 20% of recall, meaning 20% of Collembola are found and 80% have to be identified by experts. The experts can choose to gain more time and reduce the level of confidence which will augment the recall accordingly. 80% of Collembola can be identified with at least 70% precision, but some species are more difficult to identify than others. Despite the variability in the identification precision, Yolov5x6 is globally capable of identifying most of the Collembola in a very short time, whereas an expert would not be able to do so with a low zoom of only 50×. To find what are the features used by the model to make a decision we used a Grad-Cam (Selvaraju et al., 2017). The results from Fig. 7 show significant focus from the model on the background which could have been the reason for the model performance being so high despite the low zoom, but experimentation confirms that the model remains robust, even when the Collembola are translated from their background to a different project background where their species doesn't appear, except for two challenging cases, *Isotomiella minor* and Parisotoma notabilis. In these cases, altering the background creates more mistakes when making predictions as illustrated in Table 4, and the model relies on the background much more than any other species as seen on the Grad-Cam Fig. 7. This indicates that most cases don't require it, but the model uses background noise for challenging cases that lack data or mostly have annotations from older project. It is not possible to generalize a pattern of feature attention or compare the model with experts when using Grad-Cam because its interpretation is too subjective, but we can notice

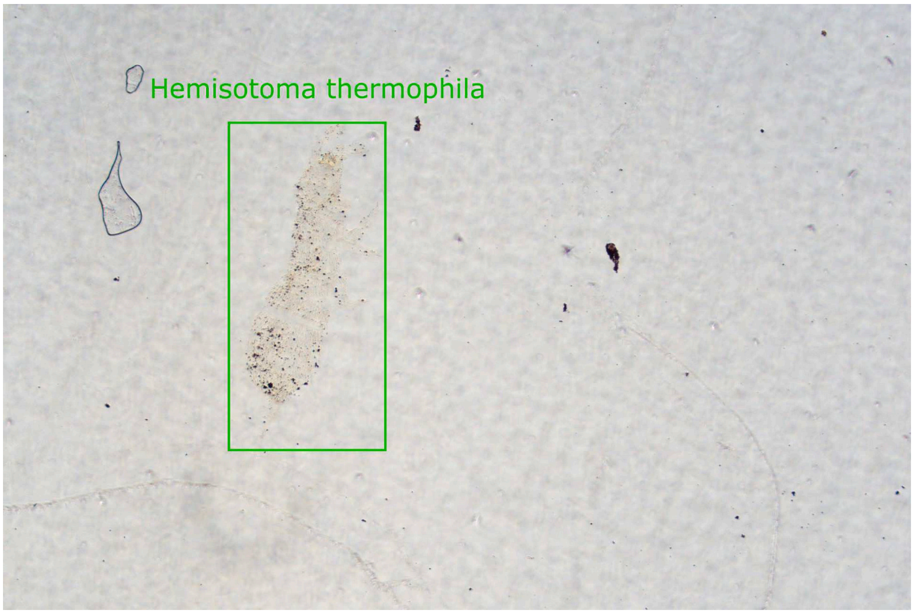


Fig. 8. Hemisotoma thermophila on an old slide. The image of the specimens is very noisy and the features making the identification possible are invisible without a bigger zoom.

the observations of specific features that experts would use, like eye plates ocular fields or antennas when they are clearly visible. Automation of Collembola identification with a deep learning model would allow the use of Collembola on a much higher scale. Our study carries significant implications for the field of Collembola species identification on microscope slides. It shed light on Yolov5, specifically Yolov5x6 on the identification of species on images taken with a microscope, and as a powerful tool in the automation of such identifications. It is noteworthy that the variety of identification difficulties emphasizes the need for more annotations, which would significantly improve model performance. In summary, while some species are harder to identify than others, our model shows potential in the identification of species of interest, and overall, our model is poised to provide substantial time savings in the identification of Collembola species of interest.

5. Conclusion

In conclusion, Collembola can play a crucial role in terrestrial ecosystems as indicators for monitoring soil quality. To use them as metrics, identification is a prerequisite, and it is a time-consuming task. Among the 8700 species of Collembola worldwide, our expertise suggests that 35 species represent the majority found in agricultural soils in France. The automation of identifying these key species using state-of-the-art object detection offers a significant time-saving opportunity for experts. This study evaluates the performance of state-of-the-art deep learning models in identifying Collembola on microscope slides, introducing a new dataset designed for model training. Our leading model outperforms human experts on images with a zoom of x50 and effectively utilizes Collembola features for identifications. While further improvements through annotation and few-shot learning techniques can enhance the model, it has already proven to be effective in substantially reducing the time required by experts. This presents a valuable tool for experts seeking to use Collembola as a metric on a larger scale.

CRediT authorship contribution statement

Théo Oriol: Writing – original draft, Conceptualization. **Jérôme Pasquet:** Writing – review & editing, Supervision. **Jérôme Cortet:** Writing – original draft, Supervision.

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.

Data availability

Data will be made available on request.

Acknowledgment

I would like to express my sincere gratitude to Charlotte Eibes, Mathias Brand, and Nicolas Henon for their invaluable contributions to the data retrieval process. I am also deeply thankful for the financial support provided by the University of Paul Valery and the Occitanie Region.

References

Anthony, M.A., Bender, S.F., van der Eijden, M.G., 2023. Enumerating soil biodiversity. In: Proceedings of the National Academy of Sciences, 120 e2304663120.

Arablouei, R., Wang, L., Currie, L., Yates, J., Alvarenga, F.A., Bishop-Hurley, G.J., 2023. Animal behavior classification via deep learning on embedded systems. Comput. Electron. Agric. 207, 107707.

Cébron, A., Cortet, J., Criquet, S., Biaz, A., Calvert, V., Caupert, C., Pernin, C., Leyval, C., 2011. Biological functioning of pah-polluted and thermal desorption-treated soils assessed by fauna and microbial bioindicators. Res. Microbiol. 162, 896–907.

Cluzeau, D., Pérès, G., Guernion, M., Chaussod, R., Cortet, J., Fargette, M., Martin-Laurent, F., Mateille, T., Pernin, C., Ponge, J.-F., et al., 2009. Intégration de la biodiversité des sols dans les réseaux de surveillance de la qualité des sols: exemple du programme pilote à l'échelle régionale, le rmqs biodiv. Etude et gestion des sols 16, 187–201.

Cluzeau, D., Guernion, M., Chaussod, R., Martin-Laurent, F., Villenave, C., Cortet, J., Ruiz-Camacho, N., Pernin, C., Mateille, T., Philippot, L., et al., 2012. Integration of biodiversity in soil quality monitoring: baselines for microbial and soil fauna parameters for different land-use types. Eur. J. Soil Biol. 49, 63–72.

Cortet, J., 2012. Programme bioindicateurs–phase ii, Analyse de l'Indicateur Basé Sur Les Peuplements de Mésofaune du sol. ADEME, Angers.

Cortet, J., Gomot-De Vauflery, A., Poinsot-Balaguer, N., Gomot, L., Texier, C., Cluzeau, D., 1999. The use of invertebrate soil fauna in monitoring pollutant effects. Eur. J. Soil Biol. 35, 115–134.

Cortet, J., Griffiths, B.S., Bohanec, M., Demsar, D., Andersen, M.N., Caul, S., Birch, A.N., Pernin, C., Tabone, E., De Vaufleury, A., et al., 2007. Evaluation of effects of transgenic bt maize on microarthropods in a european multi-site experiment. Pedobiologia 51, 207–218.

Crisci, C., Ghattas, B., Perera, G., 2012. A review of supervised machine learning algorithms and their applications to ecological data. Ecol. Model. 240, 113–122.

Deharveng, L., 2004. Recent advances in collembola systematics. Pedobiologia 48, 415–433.

Fang, Y., Guo, X., Chen, K., Zhou, Z., Ye, Q., 2021. Accurate and automated detection of surface knots on sawn timbers using yolo-v5 model. BioResources 16, 5390.

Fountain, M., Hopkin, S., 2004. A comparative study of the effects of metal contamination on collembola in the field and in the laboratory. Ecotoxicology 13, 573–587.

Girshick, R., Donahue, J., Darrell, T., Malik, J., 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 580–587.

Heisler, C., Kaiser, E.-A., 1995. Influence of agricultural traffic and crop management on collembola and microbial biomass in arable soil. Biol. Fertil. Soils 19, 159–165.

Huot, H., Cortet, J., Watteau, F., Milano, V., Nahmani, J., Sirguey, C., Schwartz, C., Morel, J.-L., 2018. Diversity and activity of soil fauna in an industrial settling pond managed by natural attenuation. Appl. Soil Ecol. 132, 34–44.

Hussain, A., Barua, B., Osman, A., Abozariba, R., Asyhari, A.T., 2021. Low latency and non-intrusive accurate object detection in forests. In: 2021 IEEE Symposium series on computational intelligence (SSCI). IEEE, pp. 1–6.

Joimel, S., Schwartz, C., Hedde, M., Kiyota, S., Krogh, P.H., Nahmani, J., Peres, G., Vergnes, A., Cortet, J., 2017. Urban and industrial land uses have a higher soil biological quality than expected from physicochemical quality. Sci. Total Environ. 584, 614–621.

Joimel, S., Schwartz, C., Bonfanti, J., Hedde, M., Krogh, P.H., Pérès, G., Pernin, C., Rakoto, A., Salmon, S., Santorufo, L., et al., 2021. Functional and taxonomic diversity of collembola as complementary tools to assess land use effects on soils biodiversity. Front. Ecol. Evol. 9, 630919.

Kampichler, C., Dzeroski, S., Wieland, R., 2000. Application of machine learning techniques to the analysis of soil ecological data bases: relationships between habitat features and collembolan community characteristics. Soil Biol. Biochem. 32, 197–209.

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521, 436–444.

Liu, D., Li, Y., Lin, J., Li, H., Wu, F., 2020. Deep learning-based video coding: a review and a case study. ACM Computing Surveys (CSUR) 53, 1–35.

Milano, V., Maisto, G., Baldantoni, D., Bellino, A., Bernard, C., Croce, A., Dubs, F., Strumia, S., Cortet, J., 2018. The effect of urban park landscapes on soil collembola diversity: a mediterranean case study. Landsc. Urban Plan. 180, 135–147.

Minaee, S., Boykov, Y., Porikli, F., Plaza, A., Kehtarnavaz, N., Terzopoulos, D., 2021. Image segmentation using deep learning: a survey. IEEE Trans. Pattern Anal. Mach. Intell. 44, 3523–3542.

Ouvrard, S., Barnier, C., Bauda, P., Beguiristain, T., Biache, C., Bonnard, M., Caupert, C., Cebron, A., Cortet, J., Cotelle, S., et al., 2011. In situ assessment of phytotechnologies for multicontaminated soil management. Int. J. Phytoremediation 13, 245–263.

Pérès, G., Vandenbulcke, F., Guernion, M., Hedde, M., Beguiristain, T., Douay, F., Houot, S., Piron, D., Richard, A., Bispo, A., et al., 2011. Earthworm indicators as tools for soil monitoring, characterization and risk assessment. An example from the national bioindicator programme (France). Pedobiologia 54, S77–S87.

Pernin, C., Cortet, J.O. Me, Joffre, R., Le Petit, J., Torre, F., 2006. Sewage sludge effects on mesofauna and cork oak (quercus suber l.) leaves decomposition in a mediterranean forest firebreak. J. Environ. Qual. 35, 2283–2292.

Ponge, J.-F., Pérès, G., Guernion, M., Ruiz-Camacho, N., Cortet, J., Pernin, C., Villenave, C., Chaussod, R., Martin-Laurent, F., Bispo, A., et al., 2013. The impact of agricultural practices on soil biota: a regional study. Soil Biol. Biochem. 67, 271–284.

Potapov, A., Bellini, B.C., Chown, S.L., Deharveng, L., Janssens, F., Kovac, L., Kuznetsova, N., Ponge, J.-F., Potapov, M., Querner, P., et al., 2020. Towards a global synthesis of collembola knowledge: challenges and potential solutions. Soil Organisms 92, 161–188.

Recknagel, F., 2001. Applications of machine learning to ecological modelling. Ecol. Model. 146, 303–310.

Redmon, J., Farhadi, A., 2018. Yolov3: An incremental improvement, arXiv preprint arXiv:1804.02767.

Ren, S., He, K., Girshick, R., Sun, J., 2015. Faster r-cnn: towards real-time object detection with region proposal networks. Adv. Neural Inf. Proces. Syst. 28.

Rustia, D.J.A., Chao, J.-J., Chiu, L.-Y., Wu, Y.-F., Chung, J.-Y., Hsu, J.-C., Lin, T.-T., 2021. Automatic greenhouse insect pest detection and recognition based on a cascaded deep learning classification method. J. Appl. Entomol. 145, 206–222.

Schneider, S., Taylor, G.W., Kremer, S.C., Burgess, P., McGroarty, J., Mitsui, K., Zhuang, A., deWaard, J.R., Fryxell, J.M., 2022. Bulk arthropod abundance, biomass and diversity estimation using deep learning for computer vision. Methods in Ecology and Evolution 13, 346–357.

Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D., 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision, pp. 618–626.

Spiesman, B.J., Gratton, C., Hatfield, R.G., Hsu, W.H., Jepsen, S., McCornack, B., Patel, K., Wang, G., 2021. Assessing the potential for deep learning and computer vision to identify bumble bee species from images. Sci. Rep. 11, 7580.

Sys, S., Weißbach, S., Jakob, L., Gerber, S., Schneider, C., 2022. 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, 2729–2742.

Tan, M., Chao, W., Cheng, J.-K., Zhou, M., Ma, Y., Jiang, X., Ge, J., Yu, L., Feng, L., 2022. Animal detection and classification from camera trap images using different mainstream object detection architectures. Animals 12, 1976.

Waldchen, J., Mader, P., 2018. Machine learning for image-based species identification. Methods Ecol. Evol. 9, 2216–2225.

Wang, L., Yan, W.Q., 2021. Tree leaves detection based on deep learning. In: Geometry and Vision: First International Symposium, ISGV 2021, Auckland, New Zealand, January 28-29, 2021, Revised Selected Papers 1. Springer, pp. 26–38.

Xu, R., Lin, H., Lu, K., Cao, L., Liu, Y., 2021. A forest fire detection system based on ensemble learning. Forests 12, 217.