



**HAL**  
open science

## Being confident in confidence scores: calibration in deep learning models for camera trap image sequences

Gaspard Dussert, Simon Chamailé-jammes, Stéphane Dray, Vincent Miele

### ► To cite this version:

Gaspard Dussert, Simon Chamailé-jammes, Stéphane Dray, Vincent Miele. Being confident in confidence scores: calibration in deep learning models for camera trap image sequences. *Remote Sensing in Ecology and Conservation*, In press, 10.1002/rse2.412 . hal-04749328

**HAL Id: hal-04749328**

**<https://hal.science/hal-04749328v1>**





Submitted on 23 Oct 2024

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

## RESEARCH ARTICLE

# Being confident in confidence scores: calibration in deep learning models for camera trap image sequences

Gaspard Dussert<sup>1</sup> , Simon Chamaille-Jammes<sup>2</sup> , Stéphane Dray<sup>1</sup>  & Vincent Miele<sup>1</sup> <sup>1</sup>Université Lyon 1, CNRS, UMR 5558, Laboratoire de Biométrie et Biologie Evolutive, Villeurbanne, France<sup>2</sup>CEFE, Université de Montpellier, CNRS, EPHE, IRD, Montpellier, France**Keywords**

Calibration, camera trap, confidence score, interpretability, machine learning, species classification

**Correspondence**Simon Chamaille-Jammes, CEFE, Université de Montpellier, CNRS, EPHE, IRD, Montpellier, France. Tel: +33 (0)467613218; E-mail: [simon.chamaille@cefe.cnrs.fr](mailto:simon.chamaille@cefe.cnrs.fr)**Funding Information**

This work was granted access to the HPC resources of IDRIS under the allocation 2022AD010113729 made by GENCI.

Editor: Dr. Marcus Rowcliffe  
Associate Editor: Dr. Jorge Ahumada

Received: 26 January 2024; Revised: 26 April 2024; Accepted: 20 May 2024

doi: 10.1002/rse2.412

**Introduction**

Camera traps have become a central tool in the monitoring and conservation of communities and populations. They generate a lot of data that can be used to infer, for instance, species richness, occupancy or activity patterns (Sollmann, 2018). To exploit these data, it is first required to identify the species present in the photos or videos. This manual annotation task is generally long and tedious, but it has been shown in recent years that it can be replaced by an automatic classification made by artificial intelligence (AI; e.g. deep learning models), often with an accuracy of over 90% (Norouzzadeh et al., 2018; Whytock et al., 2021; Willi et al., 2019).

However, a responsible use of AI (Wearn et al., 2019) requires understanding whether results can be trusted or not, generally or per prediction. In a species classification

**Abstract**

In ecological studies, machine learning models are increasingly being used for the automatic processing of camera trap images. Although this automation facilitates and accelerates the identification step, the results of these models may lack interpretability and their immediate applicability to ecological downstream tasks (e.g. occupancy estimation) remains questionable. In particular, little is known about their calibration, a property that allows confidence scores to be interpreted as probabilities that model's predictions are true. Using a large and diverse European camera trap dataset, we investigate whether deep learning models for species classification in camera trap images are well calibrated. Additionally, as camera traps are often configured to take multiple photos of the same event, we also explore the calibration of predictions aggregated across sequences of images. Finally, we study the effect and the practicality of a post-hoc calibration method, i.e. temperature scaling, for predictions made at image and sequence levels. Based on five established models and three independent test sets, we show that averaging the logits over the sequence, selecting an appropriate architecture, and optionally using temperature scaling can produce well-calibrated models. Our findings have clear implication for, for instance, the calculation of error rates or the selection of confidence score thresholds in ecological studies making use of artificial intelligence models.

model, accuracy (i.e., rate of true predictions) provides this information at the model's level, whereas confidence scores should provide this at the prediction level. In many ecological studies, downstream tasks may directly rely on these scores, for instance when subsetting data considering that values above a certain threshold indicate true detections, or when propagating model uncertainty into subsequent statistical models. In these applications, confidence scores are frequently interpreted as probabilities of the predictions being true. However, it is often neglected (as probably unknown) that there is no guarantee that these scores can be interpreted this way, as many deep learning models may return biased confidence scores (Gawlikowski et al., 2022).

In the context of classification models, a model returning confidence scores that can be reliably interpreted as probabilities of the prediction being true is said to be well

calibrated. For instance, if a well calibrated model predicts the label of 100 images with a confidence score of 0.8, we would expect to observe an actual accuracy of 80% on these images. Conversely, models can be miscalibrated in a number of ways. For instance, it is common for models to tend to globally over- or under-estimate confidence scores, and are said to be over/under-confident. Other, more complex forms of miscalibration are possible. For instance, a model may be both over-confident for high confidence scores and under-confident for low scores (Calster et al., 2019). Alternatively, a model may be globally well-calibrated but miscalibrated with respect to a variable of interest (Kelly & Smyth, 2023). There are various reasons why a model may be poorly calibrated, including the architecture, the distribution of the training set and overfitting (Guo et al., 2017; Minderer et al., 2021; Mukhoti et al., 2020). Although the question of calibration of confidence scores has been shown to be crucial in different fields such as autonomous driving (Bojarski et al., 2016) or medical diagnosis (Nair et al., 2018), it has rarely been considered in ecological studies.

In the field of ecology, a poorly calibrated model can induce several issues in use cases where ecological inferences are made using the confidence scores. Indeed, as the score distribution is biased, confidence scores can not be interpreted as probabilities so that it is impossible to control for the accuracy (or error rate) associated to a given threshold value without manually labelling at least part of the dataset. In contrast, good calibration enables the interpretability of the scores as probabilities allowing to control for the error rates in downstream tasks such as occupancy estimation (Gimenez et al., 2022; Rhinehart et al., 2022), inference of species interaction (Nicvert et al., 2024; Parsons et al., 2022), realtime alert to guide law-enforcement (Whytock et al., 2023), confidence-based checking on citizen science platforms (e.g. Zooniverse (Lotfian et al., 2021; Simpson et al., 2014)) or confidence-based uploads to biodiversity inventories (August et al., 2020).

In this paper, we explore the calibration of confidence scores in the context of species classification models for camera trap data. For that task, a common approach, as assessed in recent iWildcam competitions (Beery et al., 2021), consists in two steps: (step 1) detecting animals, humans and vehicles and filtering out empty images using a robust detection model such as MegaDetector (Beery et al., 2019; Mitterwallner et al., 2023) and (step 2) using a convolutional neural network (CNN) classification model to identify the species in the bounding box returned by the detection model, when an animal has been detected. We therefore focus on these species classification models (step 2), which are developed for a large

range of species all over the world. We explore the interplay between accuracy and calibration for five state-of-the-art model architectures applied to camera trap data collected in three out-of-sample test data sets. Also, we consider the calibration of confidence scores at the level of sequences of images. Indeed, camera traps are often configured to take multiple photos at each trigger so that predictions are aggregated at the level of the sequence of images (sometimes called the “observation” or “event”). The issue of the calibration of confidence scores aggregated at the sequence level has not, to our knowledge, been addressed in the literature. Furthermore, we study the relevance of a popular post-hoc calibration method called temperature scaling (Platt, 2000), for both image and sequence levels. Overall, in addition to providing a solution to produce calibrated scores, our work intends to illustrate the benefits and effectiveness of calibration in downstream tasks with a practical use-case. Our findings show that we can estimate accurately the rate of classification errors made by a model and this important step can improve the pipeline of analyses based on camera trap data. Lastly, we use our results to provide a set of best practices for researchers and practitioners in the field.

## Materials and Methods

### The DeepFaune dataset

We use the dataset of the DeepFaune initiative (Rigoudy et al., 2023), which is a collaborative effort involving over 50 partners who, together, have gathered over 2 million camera trap images and twenty thousand camera trap videos that they had manually annotated. These partners are affiliated to a wide range of institutions, such as organizations managing protected areas, hunting federations, and academic research groups. Images and videos were mainly collected in France, but also in a few European countries. Most of the annotations were at the species level, but some were at a higher taxonomic level (e.g. mustelid) and were provided for the whole camera trap event (a single image, an image sequence or a video). Videos were converted into images by extracting frames of the first 4 s, with a time step of 1 s. The dataset provides a great diversity of habitats, elevations and weather conditions, as well as a wide variety of camera trap models with different settings, resolutions, flash type and image processing.

### Training and validation datasets

Recent studies on species classification suggest that two-step approaches may be more efficient than classifiers processing the whole image (Bothmann et al., 2023; Celis

et al., 2024; Norman et al., 2023), and many softwares or cloud-based platforms use this approach (Agouti ([www.agouti.eu](http://www.agouti.eu)), TrapTagger (<https://wildeyeconservation.org/traptagger/>), EcoAssist (<https://addaxdatascience.com/ecoassist/>)). We use MegaDetector v5 (MDv5) (Beery et al., 2019) to extract bounding boxes of animals, human and vehicles. Because MDv5 has already near-perfect accuracy on human and vehicles we only kept, for the training of our classifier, the bounding boxes that predicted the presence of an animal. For each bounding box, we created a cropped image of the original image and propagated the label of the event to which the image belongs, resulting in 429 347 cropped images of 22 different classes (the distribution of the classes is shown in Figures S1 and S2).

To avoid overfitting and shortcut learning between the background of the images (i.e. camera trap site) and the observed species, we designed the training and validation sets to have disjoint pairs of background and species while having the same balance of species and diversity of habitats. The validation set represented about 20% of the images available while being disjoint from the training set at the species level: for each species, the validation set is made of images originating from partners different than the ones used in the training set, while being as close as possible to a 80/20 split. Ultimately, we had 368 786 images in the training set and 60 561 in the validation set.

### Out-of-sample test sets

To demonstrate that the results of the classifier could generalize beyond the images collected in the DeepFaune initiative, three out-of-sample test datasets were used. These datasets originated from ecological programs conducted in three geographically distinct areas. We refer to these datasets by the name of the areas they originate from:

- **Pyrenees:** camera trap study in the national reserve of Orlu in the French Pyrenees, conducted by the French Biodiversity Agency (OFB), 100 266 images and 12 species after preprocessing.
- **Alps:** camera trap study in the Ecrins national park in the French Alps, conducted by S. Chamaillé-Jammes, 8106 images and 12 species after preprocessing.
- **Portugal:** camera trap study in the Peneda-Gerês National Park in Portugal (Zuleger et al., 2023), publicly available. 99 750 cropped images and 16 species after preprocessing.

### Sequences of images

It is common to configure camera traps to take a series of images after each trigger. It is therefore relevant to

have a single prediction for the whole series of images. We thereafter name such series “sequences”. In our test sets, we considered that two consecutive images taken within 10 s, at the same site (i.e. the same camera trap), belonged to the same sequence. We obtained sequences of 1 to 213 images.

### Confidence score at sequence level

A sequence with  $S$  images has  $S$  individual predictions that can be aggregated in many different ways to produce a single prediction for the whole sequence. Formally, for a sequence of  $S$  images  $x_i$ , the model predicts the logits  $z_i = (z_{i1}, \dots, z_{iK})$  for each image, with  $K$  the number of classes. In our framework, only one label is predicted per image, therefore confidence scores are derived using the softmax function:  $p_i = (p_{i1}, \dots, p_{iK}) = \text{softmax}(z_{i1}, \dots, z_{iK})$ . However, it is important to note that another function could be used for other tasks (e.g. sigmoid function could be used in a multi-label classification task). We aimed at predicting the confidence score  $P_{\text{seq}} = (P_{\text{seq}1}, \dots, P_{\text{seq}K})$  as a function of the predictions at the image level. We explored four different aggregation functions (Fig 1):

- **Average Score:** We averaged, over the sequence, the scores for individual pictures of the sequence:

$$P_{\text{seq}} = \left( \frac{1}{S} \sum_{i=1}^S p_{i1}, \dots, \frac{1}{S} \sum_{i=1}^S p_{iK} \right) \quad (1)$$

- **Average Logit:** We averaged, over the sequence, the logits for individual pictures of the sequence, and then applied the softmax function:

$$P_{\text{seq}} = \text{softmax} \left( \frac{1}{S} \sum_{i=1}^S z_{i1}, \dots, \frac{1}{S} \sum_{i=1}^S z_{iK} \right) \quad (2)$$

- **Max Score:** We kept the scores of the image that had the highest score among all scores of all images of the sequence:

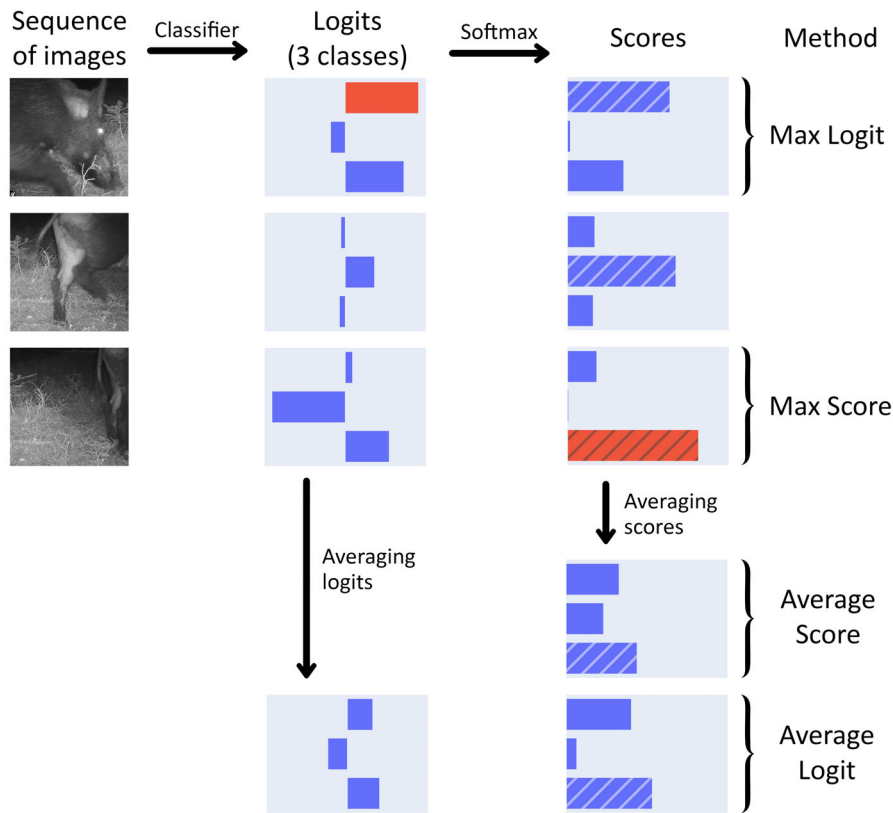
$$P_{\text{seq}} = p_{i^*} \text{ with } i^* = \operatorname{argmax}_{i \in [1, S]} \left\{ \max_{k \in [1, K]} \{p_{ik}\} \right\} \quad (3)$$

- **Max Logit:** We kept the scores of the image that had the highest logit among all logits of all images of the sequence:

$$P_{\text{seq}} = p_{i^*} \text{ with } i^* = \operatorname{argmax}_{i \in [1, S]} \left\{ \max_{k \in [1, K]} \{z_{ik}\} \right\} \quad (4)$$

### Calibration metrics

For a set of  $N$  images, we define the true class of the  $i$ -th image as  $y_i$  and the confidence scores of the  $K$  classes for



**Figure 1.** Illustration of the four aggregation methods. We represent a sequence of three images and a three-class classification problem. For each image, three logits are predicted and represented on the same row. The logits of each image are transformed into confidence scores by the softmax function. In this illustration, each aggregation method (to the right of the brackets) would have given different scores. The greatest overall logit and score are in red. The top-1 score is hatched to emphasize that only this score is used to calculate the calibration for the sequence.

that image as  $p_i = (p_{i1}, \dots, p_{iK})$ . The predicted class  $\hat{y}_i$  is the top-1 classification prediction, that is the class with the greatest confidence score, denoted  $s_i$ :

$$\hat{y}_i = \operatorname{argmax}_{k \in [1, K]} p_i \text{ and } s_i = \max_{k \in [1, K]} p_i \quad (5)$$

We define  $M$  evenly spaced bins: for  $m \in [1, M]$ , the bin  $b_m$  is the set of indices  $i$  such that  $s_i \in ]\frac{m-1}{M}, \frac{m}{M}]$ . From these, we can compute the average bin accuracy and the average bin confidence score:

$$\operatorname{acc}(b_m) = \frac{1}{|b_m|} \sum_{i \in b_m} \mathbf{1}(\hat{y}_i = y_i) \quad (6)$$

$$\operatorname{conf}(b_m) = \frac{1}{|b_m|} \sum_{i \in b_m} s_i \quad (7)$$

The bin-wise accuracy can be plotted to construct the reliability histogram (Guo et al., 2017) (e.g. Figure S3). It facilitates visualization of a model's calibration: the closer the tops of the histogram bars are from the identity line, the better calibrated the model is. In addition, if the tops

of the histogram bars are mostly above (resp. below) the line, the model is said to be under-confident (resp. over-confident).

The most common metric to measure a model's calibration quantitatively is the Expected Calibration Error (ECE) (Guo et al., 2017). ECE is defined as the bin-wise calibration error weighted by the size of the bin:

$$\operatorname{ECE} = \sum_{m=1}^M \frac{|b_m|}{N} |\operatorname{acc}(b_m) - \operatorname{conf}(b_m)| \quad (8)$$

Due to the large number of images in our test sets, we decided to use a greater number of bins, specifically 20 instead of the standard 15, to obtain a more precise measurement of calibration with the ECE. In addition to this metric, we evaluated the classification performance of our classifier with the accuracy metric. These two metrics can also be used to evaluate the classification and the calibration at the sequence level, using the score  $p_{\text{seq}}$  and the associated predicted label  $\hat{y}_{\text{seq}} = \operatorname{argmax}_{k \in [1, K]} p_{\text{seq}}$ .

## Temperature scaling

Temperature scaling (Platt, 2000) is a post-processing method to improve the calibration of the model after the training. The scores predicted by the model are rescaled by a temperature parameter  $T > 0$  using a generalization of the softmax function:

$$p_{ij} = \frac{\exp\left(\frac{z_{ij}}{T}\right)}{\sum_{k=1}^K \exp\left(\frac{z_{ik}}{T}\right)} \quad (9)$$

For  $T = 1$ , the scores obtained are the same as with the standard softmax function.

$T > 1$  leads to lower scores and helps when the model is on average too over-confident. Conversely,  $T < 1$  increases the scores and helps under-confident models. For a given dataset, it is possible to determine the optimal temperature  $T^*$ , that minimize the ECE. However, this optimum temperature may differ from one dataset to another, and determining the optimum requires access to the labels. It is therefore unrealistic to report the performance metrics calculated with these individual temperature  $T^*$ , as it cannot be calculated for a new dataset without manually annotating a fraction of the data. Instead, we propose to look at performance using a single temperature  $\bar{T}$  shared across the three datasets. We define  $\bar{T}$  as the temperature that minimizes the average ECE across the three test datasets. Temperature scaling can be combined with the four aggregation methods (Section 2.5) to calibrate sequence level predictions by simply replacing the standard softmax function with Equation 9.

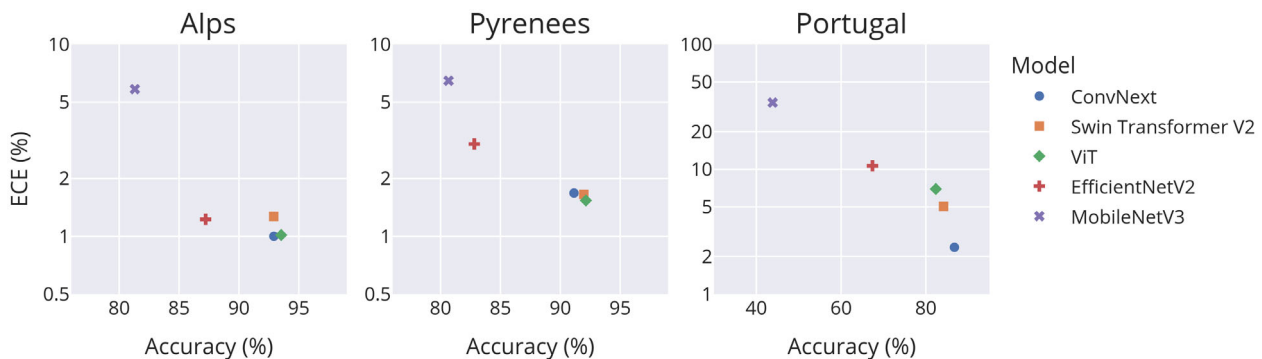
## Deep learning models

To demonstrate the robustness of our findings, we used 5 established machine learning architectures: EfficientNetV2, ConvNext, ViT, Swin Transformer V2, and MobileNetV3

(Dosovitskiy et al., 2021; Howard et al., 2019; Liu et al., 2021, 2022; Tan & Le, 2021). We have selected these architectures to represent CNNs (EfficientNetV2, ConvNext), Transformers (Swin, ViT) and light architectures that could be deployed in camera traps that do the classification at the edge (MobileNetV3). Models were trained using the timm library (Wightman, 2019) with transfer-learning from ImageNet-22 k (Ridnik et al., 2021), the largest publicly available database. Data augmentation was applied using the imgaug library (Jung et al., 2020) using only standard transformations such as flips, crops, conversion to grayscale and affine transformation. The optimization was done using Stochastic Gradient Descent (SGD), with a batch size of 32 and a different learning rate adapted for each architecture. Early stopping was used to avoid overfitting, this process monitors a metric and stops the training prematurely if there was no improvement for a specified patience. In our case, we monitor the validation accuracy with a patience of 10 epochs.

## Error rate estimation

For calibrated models, a predicted score is equivalent to the probability of the prediction being correct. Hence, for a given dataset analyzed with a classification model, the sum of the predicted scores gives an estimate of the number of correct predictions (true positives). The estimated number of errors (incorrect predictions, or false positives) is therefore the difference between the total number of images and this number. The error rate estimate can then be defined as the ratio of the estimated number of incorrect predictions over the total number of images. This calculation can also be restricted to predictions for which confidence scores are above a given threshold, which allows (i) estimation of a threshold-specific error rate (ii) identification of the threshold associated with a given error rate. We evaluated the effectiveness of this approach



**Figure 2.** Scatterplot of ECE vs. accuracy values for five models (colored points) and three test data sets (panels), computed at the image level. Here, the ECE is not postcalibrated with temperature scaling (i.e. the temperature is 1 for all models).

in practice by testing it with the ConvNext model and the three test datasets pooled together. We compared the error rate using different aggregation methods, and a wide range of thresholds and temperature scaling.

## Results

### Calibration at the image level

Generally, we observed that without temperature scaling, more accurate models were better calibrated (Fig. 2), as indicated by lower Expected Calibration Error (ECE). ConvNext was the model providing the best overall performance: by averaging the metrics over the three datasets, it had the best accuracy (90.27%) and the best ECE (1.68%). The models exhibited comparable relative performances across the three datasets. ViT and Swin Transformer V2 performed close to ConvNext, while EfficientNetV2 showed substantially lower performance on both metrics. Additionally, the lightweight model, MobileNet, had bad to very bad (ECE of 34.27% on the Portugal dataset) accuracy and calibration performances.

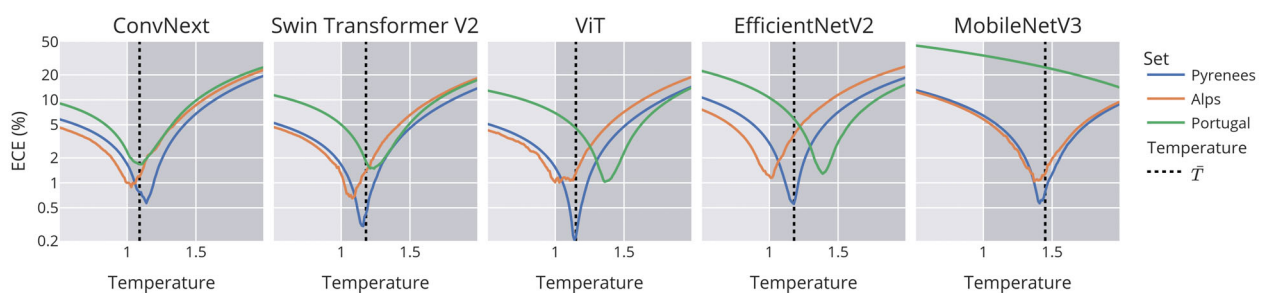
As expected, temperature scaling allowed improving ECE values, for all models and datasets. We almost always observed a V-shaped relationship between ECE and temperature, with ECE increasing quickly and by several percents around the optimum temperature value (Fig. 3). This optimum temperature was generally greater than 1, suggesting that all models were initially overconfident to a greater or lesser extent. Interestingly, the V-shaped curves of the different datasets overlapped well for the most accurate models (ConvNext and transformed-based models, ViT and Swin), and optimum temperature were similar across datasets. This suggested that we find a temperature value that, although common to all datasets, would help to improve the calibration in post-processing for each dataset. Indeed, in the best case (ConvNext), the temperature  $\bar{T}$  reduced the calibration error by 28%

compared to the situation without temperature scaling ( $T = 1$ ) (dashed line in Fig. 3), versus a reduction of 38% of the calibration error with a distinct temperature per datasets ( $T^*$ ). In the worst case (EfficientNetV2)  $\bar{T}$  leads to a reduction of 31% versus 80% for per-dataset  $T^*$ .

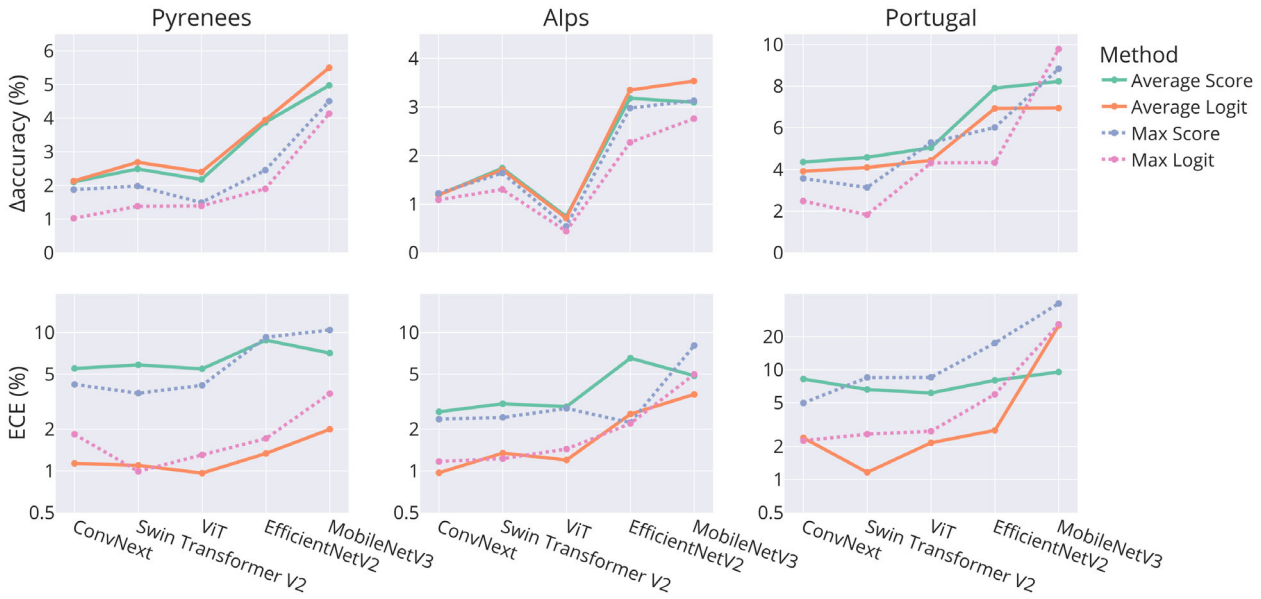
### Calibration at the sequence level

All the sequence aggregation methods led to a model with an overall accuracy much greater than without aggregation (i.e. at the image level) (Fig. 4 top). This was true for all models and all datasets, with up to +10% of accuracy for MobileNetV3 on the Portugal dataset. The Average Score and Average Logit were the two best methods for maximizing accuracy, with a slight advantage for the former (respectively 3.71% and 3.56%  $\Delta$ Accuracy on average). Importantly, of the two aggregation methods that improved accuracy most (Average Score and Average Logit), Average Logit provided better calibrated scores (Fig. 4 bottom, 3.32% and 6.08% ECE on average). Therefore, considering both accuracy and calibration metrics, the Average Logit was the best aggregation method.

We finally studied the interplay between temperature scaling and aggregation methods. While all models demonstrated overconfident predictions at the image-level, some aggregation methods, especially the Average Score method, resulted in under-confident predictions. We observed that the aforementioned V-shaped was more flat for the Average Score method than for the other methods (first column in Figure S4 versus the others). Consequently, Average Score has the worst calibration of the four methods with temperature scaling (3.99% ECE), significantly behind the second-worst method (Max Score, 1.47% ECE). We also noted that the Average Logit method provided the lowest ECE values overall (1.17% on average), and thus remained the best method, with temperature scaling further improving calibration at sequence level. Finally, and as observed at the image level,



**Figure 3.** Calibration transferability using temperature scaling, at the image level. Curves of ECE values along the gradient of temperature values, for five models (panels) and three test data sets (colored curves). An optimum temperature below 1 indicates a model that is on average too underconfident (light gray area), and above 1 indicates a model on average too overconfident (dark gray area). The vertical dashed line shows  $\bar{T}$ , the temperature minimizing the average ECE across the three test datasets.

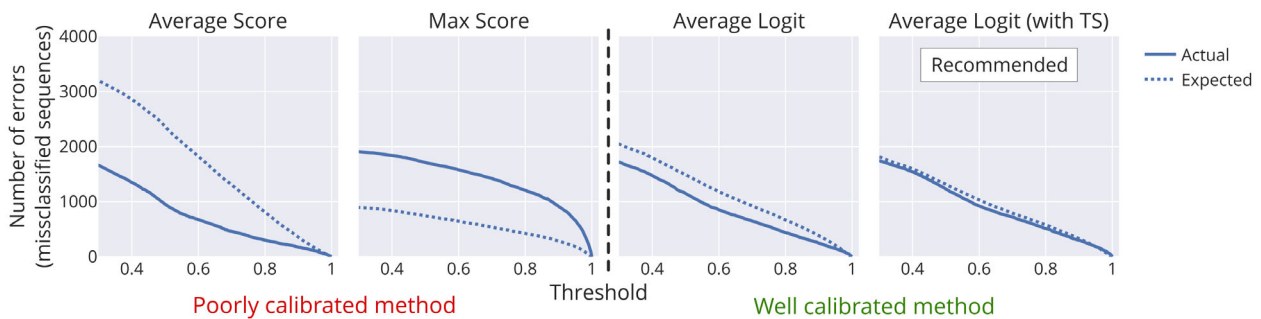


**Figure 4.**  $\Delta$ Accuracy (top, the greater the better) and ECE (bottom, the lower the better) for the four aggregation methods (colored curves) and five models (x-axis) on three test data sets (three panels).  $\Delta$ Accuracy is the difference between the accuracy at the sequence level and the accuracy at the image level. For the sake of clarity, solid lines represents the two best methods in terms of accuracy, Average Score and Average Logit. On the ECE plots (bottom), Average Logit outperforms Average Score in terms of calibration.

a single temperature (possibly close to 1) would be sufficient to improve the calibration with the Average Logit method. Using one temperature for all datasets with the ConvNext model reduces ECE by 25.5% and can be further reduced to 41.4% of the original ECE with individual temperatures. However, using MobileNetV3, we observed a reduction of 25.5%, which is significantly less than the optimal temperature scaling (77.6% with three temperatures), this difference can be attributed to the Portugal dataset having a very different value.

### Calibration use case: controlling the number of errors along the score distribution

A practical implication of calibration is the ability to precisely estimate the number of incorrect predictions on a novel test set without labels. Here we show the impact of calibration on the quality of this estimation, using the ConvNext model (Fig. 5), on the 23 353 sequences gathered after pooling the three data sets. Average Score and Max Score methods provide underconfident and



**Figure 5.** Estimation of the number of errors (i.e. misclassified sequences) when using different thresholds on the confidence scores (dash line), compared to the actual number of errors (solid line), for different aggregation methods (panels). Results shown are for the ConvNext model, the three test sets Pyrenees, Alps and Portugal being pooled together. The expected number of errors is obtained from the scores (see Material and Methods). The actual number of errors is simply the number of incorrect predictions with a score above a certain threshold, and can be known only if species labels are available. An accurate estimation of the number of errors is observed when the two lines overlap. TS means temperature scaling.



overconfident scores respectively, produce inaccurate error estimations (ECE values of 6.67 and 4.38%) and under-estimate (respectively over-estimate) the number of errors by a wide margin for every threshold. For instance, with the Average Score approach, if we consider a threshold of 0.8 (often used in ecological studies (Whytock et al., 2021)), one would predict that the number of errors in the predictions would be approximately 500 images greater than the actual number (from 300 to 800, Fig. 5). Conversely, with the Max Score approach, one would predict that the number of errors in the predictions would be approximately 800 images less than the actual number (from 1200 to 400, Fig. 5). Using Average Logit, the best-calibrated method without temperature scaling (ECE of 1.65%), the estimation of the number of errors is very close to the actual number. Finally, by adding temperature scaling with a temperature  $\bar{T}=0.93$  (ECE of 0.79%), the two curves overlap. These conclusions hold with the other models except MobileNetV3 (Figure S5).

## Discussion

This study assessed the calibration of confidence scores, at image and sequence level, for different deep learning models in the context of species classification in camera trap data. Using five state-of-the-art models and three out-of-sample test datasets, we showed that score calibration can vary greatly across model architectures, in a way that is consistent across test sets. Further, we showed that the different aggregation methods to obtain scores at the sequence level gave very different calibration values, and that the Average Logit method provides better results than the others in terms of both accuracy and calibration. Finally, we showed that temperature scaling can be used both at image level and sequence level, with a single temperature  $\bar{T}$  that, for most models, improves calibration similarly across all datasets.

Our results about the importance of calibration in classification models are crucial for researchers and practitioners dealing with camera trap images. We argue that the integration of machine learning predictions directly into subsequent ecological tasks can be facilitated by achieving calibration. Many ecological downstream tasks (e.g. estimating occupancy, abundance or activity patterns) based on deep learning predictions use an empirical threshold (Krivek et al., 2023; Lonsinger et al., 2023) to consider that a prediction is correct, or test a series of thresholds to determine the optimal one given known species labels (Mitterwallner et al., 2023; Whytock et al., 2021). Fortunately, we showed that calibrated models make it possible to estimate accurate error rates for any threshold. This is critical as it allows one to conduct an analysis with a

known error rate, whose level could depend on the context of the study. Calibrated models will also be central in the development of methods that can handle probabilistic uncertainties (i.e. that directly uses predictions and associated scores rather than relying on thresholding of scores) (Rhinehart et al., 2022). Nevertheless, it's important to recognize the limitations of model calibration in the context of ecological applications. In many scenarios, a gain in calibration can be less profitable than a gain in accuracy. Indeed, if confidence scores are not used in downstream tasks, or if predictions of non-empty images are verified by a human (which is a major use of classification models), then model calibration is not a priority.

Differences in models' performance can be partly explained by model size. Indeed, we found that models with the highest number of parameters (ConvNext, ViT, SwinTransformer) gave the best accuracy and ECE values. On the other hand, the only lightweight model, MobileNetV3, was consistently the worst model. This result suggests that edge computing models are unlikely to be calibrated without post-processing. Despite some literature showing that neural networks can be poorly calibrated, our result shows that this is not always the case (see also Minderer et al., 2021), and that certain families of model architectures, such as ConvNext here, are intrinsically better calibrated than others, independently of the size of the model. The calibration of each model can be further improved on each dataset using temperature scaling as post-processing function. However, determining the optimal temperature requires annotating at least a fraction of the target set of images, a step that practitioners would like to avoid if possible. Fortunately, our empirical findings across three different datasets show that the optimal temperatures remain similar across datasets, both at the image level and sequence level when using the Average Logit aggregation method. This suggests the generalizability of a single temperature that can be determined and fixed for future test sets. That said, we do not claim that the optimal temperatures defined in this paper can be used directly when using one of the studied architectures. Indeed, these temperatures are valid for a given training procedure (datasets, hyperparameters). In practice, it is necessary to estimate the temperature using available test dataset(s) and subsequently maintain this temperature for deployment (since we showed it will be generalizable). This way, when the model will be used to classify new data, the previously estimated temperature will ensure a much better calibration of the predicted scores (though still perfectible using image annotation). We however found that this approach had limitations when applied to the MobileNetV3 model, suggesting that lightweight models for edge computing would be difficult to calibrate even after temperature scaling on an annotated dataset. It

would be interesting to validate this on other lightweight models in future work.

Proper model calibration at the image level is not always sufficient, as many software programs and scientific studies operate at the scale of the sequences that define the relevant “observations” or “events” from an ecological viewpoint. It is therefore extremely important to be able to calibrate the predictions at sequence level. For the first time, we showed that the most intuitive approach, in which scores are averaged, did not provide the best accuracy and had the worst calibration, with largely under-confident predictions. Interestingly, our findings can be confirmed by the analogy with ensemble models. These approaches use  $N$  models to make a prediction on *one* image, whereas we use  $N$  images to make a prediction with *one* model at the sequence level. Wu and Gales (2021) showed that for ensemble models, individual model calibration is not sufficient to yield a calibrated ensemble prediction, and that their own method, which is equivalent to Average Score approach also leads to under-confidence. Moreover, Rahaman and Thiery (2021) show that, thanks to this natural shift in the optimal temperature when models are ensembled, if the individual models were slightly overconfident ( $T > 1$ , as is often the case in deep learning) then the ensemble model was naturally calibrated ( $T \sim 1$ ). Our results strongly support the use of the Average Logit method for aggregating individual scores at the sequence level. It shifts slightly the optimal temperature towards underconfidence, which counterbalanced the overconfident nature of deep learning networks, and resulted in sequence-level prediction that are almost calibrated without post-processing. With Average Logit, it is still interesting to use temperature scaling to improve calibration as much as possible, especially given that the ECE minima are again very close to each other and allow a single temperature to be set.

In this work, we focused on temperature scaling and did not consider other methods that have been shown to sometimes improve calibration, such as label smoothing and mixup (Szegedy et al., 2015; Zhang et al., 2018). We did so because these two methods are debated, as several studies have showed that they can worsen calibration when combined with temperature scaling (Minderer et al., 2021; Wang et al., 2023). As Minderer et al. (2021) state, “label smoothing creates artificially underconfident models and may therefore improve calibration for a specific amount of distribution shift”. Label smoothing also assumes that all incorrect classes are equally likely (Maher & Kull, 2021), which is obviously problematic in ecology (e.g., a wrongly predicted roe deer is much more likely to be a red deer than a wolf). Mixup also deteriorates calibration properties of networks by creating non-realistic images in the training set and leading to substantial

distributional shift (Gawlikowski et al., 2022; Rahaman & Thiery, 2021). Furthermore, we have focused our work on the calibration of the top-1 label on a multiclass classification task, and with a relatively small number of classes. But in future work, it would also be interesting to look at the calibration of the predictions on other tasks of interest for models applied to ecology. For instance, calibration could be explored in multi-label classification, hierarchical classification, object detection or classification over a very large number of classes.

Our work concludes with some recommendations. For certain use cases, we encourage ecologists to consider the implications of calibration as well as accuracy. Secondly, we recommend using the Average Logit method to aggregate information at sequence level, as it performs very well in terms of accuracy and calibration. Finally, to use temperature scaling and make calibration even better, the optimum temperature can be calculated on a test dataset and kept for future datasets. We acknowledge that these considerations may look too difficult to take into account for many practitioners. We therefore hope that developers of camera-trap analysis software or platforms will be able to integrate the knowledge brought by this work into their software solutions.

## Acknowledgements

This work was granted access to the HPC resources of IDRIS under the allocation 2022AD010113729 made by GENCI. We acknowledge the organizations and people contributing to the DeepFaune initiative. Two anonymous reviewers provided in-depth comments that helped us improving the manuscript.

## Author contributions

G.D., S.C.J., S.D. and V.M. conceived the ideas and designed the methodology. G.D., S.C.J. and V.M. gathered the training data. S.C.J collected the data of the Alps test set. G.D. and V.M. coded and performed the analysis. G.D. wrote the first version of the manuscript, S.C.J., S.D. and V.M. contributed critically to the drafts and gave final approval for publication.

## Conflict of interest

None of the authors has a conflict of interest.

## Data availability statement

The five trained models, all derived data used in the analysis, and the code for the inference and metric calculation are available at [10.5281/zenodo.10014376](https://doi.org/10.5281/zenodo.10014376). The Pyrenees

dataset is available upon request only, because of the presence of a sensitive species (brown bear).

## References

- August, T.A., Pescott, O.L., Joly, A. & Bonnet, P. (2020) AI naturalists might hold the key to unlocking biodiversity data in social media imagery. *Patterns*, **1**(7), 100116. Available from: <https://doi.org/10.1016/j.patter.2020.100116>
- Beery, S., Agarwal, A., Cole, E. & Birodkar, V. (2021) The iWildCam 2021 competition dataset (arXiv:2105.03494). arXiv <http://arxiv.org/abs/2105.03494>
- Beery, S., Morris, D. & Yang, S. (2019) Efficient pipeline for camera trap image review (arXiv:1907.06772). arXiv <https://doi.org/10.48550/arXiv.1907.06772>
- Bojarski, M., Del Testa, D., Dworakowski, D., Firner, B., Flepp, B., Goyal, P. et al. (2016) End to end learning for self-driving cars (arXiv:1604.07316). arXiv <https://doi.org/10.48550/arXiv.1604.07316>
- Bothmann, L., Wimmer, L., Charrakh, O., Weber, T., Edelhoﬀ, H., Peters, W. et al. (2023) Automated wildlife image classification: an active learning tool for ecological applications. *Ecological Informatics*, **77**, 102231. Available from: <https://doi.org/10.1016/j.ecoinf.2023.102231>
- Calster, B.V., McLernon, D.J., Smeden, M.v., Wynants, L., Steyerberg, E.W. & Topic Group ‘Evaluating diagnostic tests and prediction models’ of the STRATOS initiative. (2019) Calibration: the Achilles heel of predictive analytics. *BMC Medicine*, **17**, 230. Available from: <https://doi.org/10.1186/s12916-019-1466-7>
- Celis, G., Ungar, P., Sokolov, A., Sokolova, N., Böhner, H., Liu, D. et al. (2024) A versatile, semi-automated image analysis workflow for time-lapse camera trap image classification. *Ecological Informatics*, **81**, 102578. Available from: <https://doi.org/10.1016/j.ecoinf.2024.102578>
- Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T. et al. (2021) An image is worth 16x16 words: transformers for image recognition at scale (arXiv:2010.11929). arXiv <https://doi.org/10.48550/arXiv.2010.11929>
- Gawlikowski, J., Tassi, C.R.N., Ali, M., Lee, J., Humt, M., Feng, J. et al. (2022) A survey of uncertainty in deep neural networks (arXiv:2107.03342). arXiv <http://arxiv.org/abs/2107.03342>
- Gimenez, O., Kervellec, M., Fanjul, J.-B., Chainé, A., Marescot, L., Bollet, Y., & Duchamp, C. (2022). Trade-off between deep learning for species identification and inference about predator-prey co-occurrence. *Computo*. <https://doi.org/10.57750/yfm2-5f45>.
- Guo, C., Pleiss, G., Sun, Y. & Weinberger, K.Q. (2017) On calibration of modern neural networks (arXiv:1706.04599). arXiv <http://arxiv.org/abs/1706.04599>
- Howard, A., Sandler, M., Chu, G., Chen, L.-C., Chen, B., Tan, M. et al. (2019) Searching for MobileNetV3 (arXiv:1905.02244). arXiv <http://arxiv.org/abs/1905.02244>
- Jung, A.B., Wada, K., Crall, J., Tanaka, S., Graving, J., Reinders, C. et al. (2020) Imgaug. <https://github.com/aleju/imgaug>
- Kelly, M. & Smyth, P. (2023) Variable-based calibration for machine learning classifiers (arXiv:2209.15154). arXiv <https://doi.org/10.48550/arXiv.2209.15154>
- Krivek, G., Gillert, A., Harder, M., Fritze, M., Frankowski, K., Timm, L. et al. (2023) BatNet: a deep learning-based tool for automated bat species identification from camera trap images. *Remote Sensing in Ecology and Conservation*, **9**, 759–774. Available from: <https://doi.org/10.1002/rse2.339>
- Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021). Swin Transformer: hierarchical vision transformer using shifted windows. 10012–10022. [https://openaccess.thecvf.com/content/ICCV2021/html/Liu\\_Swin\\_Transformer\\_Hierarchical\\_Vision\\_Transformer\\_Using\\_Shifted\\_Windows\\_ICCV\\_2021\\_paper.html](https://openaccess.thecvf.com/content/ICCV2021/html/Liu_Swin_Transformer_Hierarchical_Vision_Transformer_Using_Shifted_Windows_ICCV_2021_paper.html)
- Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T. & Xie, S. (2022) A ConvNet for the 2020s (arXiv:2201.03545). arXiv <https://doi.org/10.48550/arXiv.2201.03545>
- Lonsinger, R.C., Dart, M.M., Larsen, R.T. & Knight, R.N. (2023) Efficacy of machine learning image classification for automated occupancy-based monitoring. *Remote Sensing in Ecology and Conservation*, **10**(1), 56–71. Available from: <https://doi.org/10.1002/rse2.356>
- Lotfian, M., Ingensand, J. & Brovelli, M.A. (2021) The partnership of citizen science and machine learning: benefits, risks, and future challenges for engagement, data collection, and data quality. *Sustainability*, **13**(14), 14. Available from: <https://doi.org/10.3390/su13148087>
- Maher, M. & Kull, M. (2021) Instance-based label smoothing for better calibrated classification networks (arXiv:2110.05355). arXiv <http://arxiv.org/abs/2110.05355>
- Minderer, M., Djolonga, J., Romijnders, R., Hubis, F., Zhai, X., Houlsby, N. et al. (2021) Revisiting the calibration of modern neural networks (arXiv:2106.07998). arXiv <http://arxiv.org/abs/2106.07998>
- Mitterwallner, V., Peters, A., Edelhoﬀ, H., Mathes, G., Nguyen, H., Peters, W. et al. (2023) Automated visitor and wildlife monitoring with camera traps and machine learning. *Remote Sensing in Ecology and Conservation*, **10** (2), 236–247. Available from: <https://doi.org/10.1002/rse2.367>
- Mukhoti, J., Kulharia, V., Sanyal, A., Golodetz, S., Torr, P.H.S. & Dokania, P.K. (2020) Calibrating deep neural networks using focal loss (arXiv:2002.09437). arXiv <http://arxiv.org/abs/2002.09437>
- Nair, T., Precup, D., Arnold, D.L. & Arbel, T. (2018) Exploring uncertainty measures in deep networks for multiple sclerosis lesion detection and segmentation (arXiv:1808.01200). arXiv <https://doi.org/10.48550/arXiv.1808.01200>

- Nicvert, L., Donnet, S., Keith, M., Peel, M., Somers, M.J., Swanepoel, L.H. et al. (2024) Using the multivariate Hawkes process to study interactions between multiple species from camera trap data. *Ecology*, **105**, e4237.
- Norman, D.L., Bischoff, P.H., Wearn, O.R., Ewers, R.M., Rowcliffe, J.M., Evans, B. et al. (2023) Can CNN-based species classification generalise across variation in habitat within a camera trap survey? *Methods in Ecology and Evolution*, **14**(1), 242–251.
- Norouzzadeh, M.S., Nguyen, A., Kosmala, M., Swanson, A., Palmer, M.S., Packer, C. et al. (2018) Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning. *Proceedings of the National Academy of Sciences of the United States of America*, **115**(25), E5716–E5725. Available from: <https://doi.org/10.1073/pnas.1719367115>
- Parsons, A.W., Kellner, K.F., Rota, C.T., Schuttler, S.G., Millsbaugh, J.J. & Kays, R.W. (2022) The effect of urbanization on spatiotemporal interactions between gray foxes and coyotes. *Ecosphere*, **13**(3), e3993. Available from: <https://doi.org/10.1002/ecs2.3993>
- Platt, J. (2000) Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in Large Margin Classifiers*, **10**, 61–74.
- Rahaman, R. & Thiery, A.H. (2021) *Uncertainty quantification and deep ensembles* (arXiv:2007.08792). arXiv <http://arxiv.org/abs/2007.08792>
- Rhinehart, T.A., Turek, D. & Kitzes, J. (2022) A continuous-score occupancy model that incorporates uncertain machine learning output from autonomous biodiversity surveys. *Methods in Ecology and Evolution*, **13**(8), 1778–1789. Available from: <https://doi.org/10.1111/2041-210X.13905>
- Ridnik, T., Ben-Baruch, E., Noy, A. & Zelnik-Manor, L. (2021) *ImageNet-21K pretraining for the masses* (arXiv:2104.10972). arXiv <https://doi.org/10.48550/arXiv.2104.10972>
- Rigoudy, N., Dussert, G., Benyoub, A., Besnard, A., Birck, C., Boyer, J. et al. (2023) The DeepFaune initiative: a collaborative effort towards the automatic identification of European fauna in camera trap images. *European Journal of Wildlife Research*, **69**(6), 113. Available from: <https://doi.org/10.1007/s10344-023-01742-7>
- Simpson, R., Page, K. R., & De Roure, D. (2014). Zooniverse: observing the world's largest citizen science platform. Proceedings of the 23rd international conference on world wide web, 1049–1054. <https://doi.org/10.1145/2567948.2579215>.
- Sollmann, R. (2018) A gentle introduction to camera-trap data analysis. *African Journal of Ecology*, **56**(4), 740–749. Available from: <https://doi.org/10.1111/aje.12557>
- Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. & Wojna, Z. (2015) Rethinking the inception architecture for computer vision (arXiv:1512.00567). arXiv <http://arxiv.org/abs/1512.00567>
- Tan, M. & Le, Q.V. (2021) EfficientNetV2: smaller models and faster training (arXiv:2104.00298). arXiv <http://arxiv.org/abs/2104.00298>
- Wang, Q., Du, J., Yan, K. & Ding, S. (2023) Seeing in flowing: adapting CLIP for action recognition with motion prompts learning (arXiv:2308.04828). arXiv <http://arxiv.org/abs/2308.04828>
- Wearn, O.R., Freeman, R. & Jacoby, D.M.P. (2019) Responsible AI for conservation. *Nature Machine Intelligence*, **1**(2), 72–73. Available from: <https://doi.org/10.1038/s42256-019-0022-7>
- Whytock, R.C., Suijten, T., van Deursen, T., Świeżewski, J., Mermiaghe, H., Madamba, N. et al. (2023) Real-time alerts from AI-enabled camera traps using the iridium satellite network: a case-study in Gabon, Central Africa. *Methods in Ecology and Evolution*, **14**(3), 867–874. Available from: <https://doi.org/10.1111/2041-210X.14036>
- Whytock, R.C., Świeżewski, J., Zwerts, J.A., Bara-Słupski, T., Koumba Pambo, A.F., Rogala, M. et al. (2021) Robust ecological analysis of camera trap data labelled by a machine learning model. *Methods in Ecology and Evolution*, **12**(6), 1080–1092. Available from: <https://doi.org/10.1111/2041-210X.13576>
- Wightman, R. (2019) PyTorch Image Models. In: *GitHub repository*. San Francisco, CA: GitHub. Available from: <https://doi.org/10.5281/zenodo.4414861>
- Willi, M., Pitman, R.T., Cardoso, A.W., Locke, C., Swanson, A., Boyer, A. et al. (2019) Identifying animal species in camera trap images using deep learning and citizen science. *Methods in Ecology and Evolution*, **10**(1), 80–91. Available from: <https://doi.org/10.1111/2041-210X.13099>
- Wu, X. & Gales, M. (2021) Should ensemble members be calibrated? (arXiv:2101.05397). arXiv <http://arxiv.org/abs/2101.05397>
- Zhang, H., Cisse, M., Dauphin, Y.N. & Lopez-Paz, D. (2018) mixup: beyond empirical risk minimization (arXiv: 1710.09412). arXiv <https://doi.org/10.48550/arXiv.1710.09412>
- Zuleger, A., Perino, A., Wolf, F., Wheeler, H. & Pereira, H. (2023) Long-term monitoring of mammal communities in the Peneda-Gerês National Park using camera trap data. <https://doi.org/10.15468/rah33j>

## Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

**Figure S1.** Number of images in the training and validation sets, for each species. Log scale is used to improve the readability of the rarer classes.

**Figure S2.** Number of images in the three out-of-sample datasets, for each species. Log scale is used to improve the readability of the rarer classes.

**Figure S3.** Reliability histogram of the ConvNext model, using the 3 test sets pooled together, and without temperature scaling.

**Figure S4.** Calibration transferability using temperature scaling, at the sequence level.

**Figure S5.** Estimation of the number of errors (i.e. misclassified sequences) as a function of the confidence score (dashed line), compared to the actual number of errors (solid line), for different methods.