nature machine intelligence

ARTICLES

https://doi.org/10.1038/s42256-021-00393-0

Check for updates

Iterative human and automated identification of wildlife images

Zhongqi Miao[1,2,7] ✉, Ziwei Liu[3,7], Kaitlyn M. Gaynor[4], Meredith S. Palmer[5], Stella X. Yu[1,2] and Wayne M. Getz[1,6] ✉

Camera trapping is increasingly being used to monitor wildlife, but this technology typically requires extensive data annotation. Recently, deep learning has substantially advanced automatic wildlife recognition. However, current methods are hampered by a dependence on large static datasets, whereas wildlife data are intrinsically dynamic and involve long-tailed distributions. These drawbacks can be overcome through a hybrid combination of machine learning and humans in the loop. Our proposed iterative human and automated identification approach is capable of learning from wildlife imagery data with a long-tailed distribution. Additionally, it includes self-updating learning, which facilitates capturing the community dynamics of rapidly changing natural systems. Extensive experiments show that our approach can achieve an ~90% accuracy employing only ~20% of the human annotations of existing approaches. Our synergistic collaboration of humans and machines transforms deep learning from a relatively inefficient post-annotation tool to a collaborative ongoing annotation tool that vastly reduces the burden of human annotation and enables efficient and constant model updates.

In our rapidly changing world, continuous monitoring of natural systems is essential to understand and mitigate the impact of human activity on ecological processes[1–3]. Recent technological innovations now allow for the rapid collection of ecological data across vast spatial and temporal scales. However, the resulting information deluge creates a bottleneck for researchers, who must process the data at management-relevant timescales[4]. Artificial intelligence (AI) offers promising solutions for rapid and high-accuracy data processing[5,6]. However, the dynamic nature of ecological systems poses unique challenges when developing accurate algorithms[7,8]. To overcome these hurdles, we showcase how the integration of limited human labour into the machine learning workflow can greatly increase both the efficiency and accuracy of data processing.

Long-term camera trapping

We are currently experiencing a rapid, human-driven loss of global biodiversity[9–12]. To understand the complex patterns, drivers and consequences of species declines and extinctions, ecologists are increasingly employing emerging technology to assist with data collection and processing. Motion-activated remote cameras (henceforth 'camera traps') have emerged as a popular non-invasive tool for monitoring terrestrial vertebrate communities[13–15]. Their decreasing cost and greater reliability have recently led to the application of camera traps for long-term, continuous deployment aiming to monitor entire wildlife communities across multiple seasons and years[1,16–18]. Compared with one-time or annual surveys, continuous monitoring reveals new insights into wildlife responses to local, regional and global environmental changes, as well as to conservation interventions. This scale of monitoring is particularly valuable for capturing responses to environmental perturbations as they occur[1,2]. The 'Snapshot Serengeti' project (http://www.snapshotserengeti.org), which has operated continuously since 2010, is a flagship example of a long-term camera-trap monitoring programme. Over the past decade, this survey has gathered unprecedented longitudinal data that have substantially enhanced our understanding of the seasonal and inter-annual dynamics of the Serengeti ecosystem[16,19,20]. Projects of this magnitude have become increasingly common across eastern and southern Africa[18] and around the world[1].

The greatest logistical barrier to long-term monitoring with camera traps is the overwhelming amount of human labour needed to annotate thousands or millions of wildlife images for ecological analysis[4,5,16]. This annotation bottleneck creates a considerable mismatch between the paces of data collection and data processing, substantially curtailing the usefulness of camera-trap data for ongoing conservation and monitoring efforts[4]. For example, a relatively modest camera-trap survey (~80 camera traps[1]) captures millions of images a year. We estimate that it would take a single trained expert around 200 full-time working days to annotate one million images. As such, hundreds of human annotators (for example, experts, trained volunteers and citizen scientists) are required to keep pace with image accumulation. This need is likely to grow exponentially over the coming decades as more monitoring sites are set up. Although only one or two experts are needed to validate each wildlife image, it is common practice that multiple (5–20) volunteers or citizen scientists look at each image to produce a high-accuracy 'consensus' classification (~97% accurate compared to expert identifications[16]). This duplication of effort needed to generate accurate results using volunteers further perpetuates the classification bottleneck.

Automatic image-recognition systems

The use of deep learning (a subset of AI technology) to automatically identify animals in camera-trap images has recently drawn considerable attention from the ecological community. Currently,

[1]Department of Environmental Science, Policy, and Management, University of California, Berkeley, Berkeley, CA, USA. [2]International Computer Science Institute, University of California, Berkeley, Berkeley, CA, USA. [3]School of Computer Science and Engineering, Nanyang Technological University, Singapore, Singapore. [4]National Center for Ecological Analysis and Synthesis, University of California, Santa Barbara, Santa Barbara, CA, USA. [5]Department of Ecology and Evolutionary Biology, Princeton University, Princeton, NJ, USA. [6]School of Mathematics, Statistics and Computer Science, University of KwaZulu-Natal, Durban, South Africa. [7]These authors contributed equally: Zhongqi Miao, Ziwei Liu. ✉e-mail: zhongqi.miao@berkeley.edu; wgetz@berkeley.edu

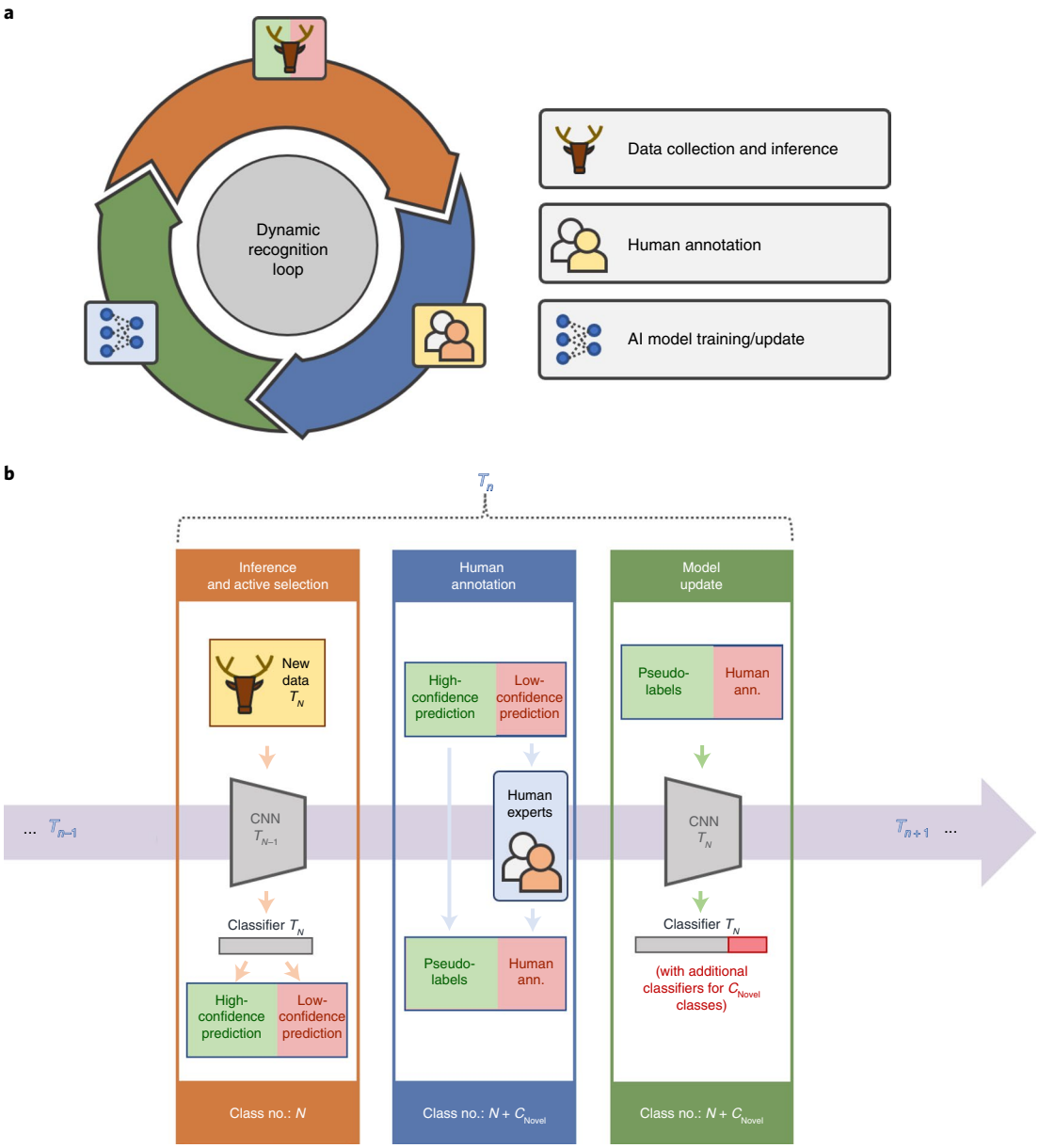


Fig. 1 | Overview of a realistic animal classification system. a, The dynamic recognition loop. In real-world applications, machine learning models do not stop at one training stage. As data collection progresses over time, there is a continuous cycle of inference, annotation and model updating. Every time a tranche of new data are added, pre-trained models are applied to classify the data. When there are novel and difficult samples, human annotation is required and the model needs to be updated to reflect the newly added data. **b**, The progression of a realistic animal classification system. Even if the trained model has high accuracy for the previous validation sets, there may be a difference in the classes between previous validation sets and current inference data (for example, there may be novel categories in the newly collected data that did not exist in previous training and validation sets). Models thus need to be updated over time. Here we present a more practical procedure that can both maximize the utility of modern image-recognition methods and minimize the dependence on manual annotation for model updating. We incorporate an active learning technique that actively selects low-confidence predictions for further human annotation, while keeping highly confident predictions as pseudo-labels. Models are then updated according to both human annotations and pseudo-labels. Ann., annotation; CNN, convolutional neural network; C_{Novel}, number of novel classes at time step T_n; N, total number of classes at time step T_{n-1}; T, time step.

trained deep learning algorithms can classify a million images in a single day running on a desktop computer, a substantial advancement over the months of effort required for human annotators to accomplish the same task[5,21,22].

There have been several attempts to develop robust camera-trap recognition methods for real-world deployment, either tackling distribution shifts (in species numbers and locations) with transfer learning[23–25] or addressing new species emergence with active learning[26–28]. However, before it becomes feasible to rely on deep learning to handle the mass of image data from large-scale, long-term camera-trap projects, two major impediments must be overcome: (1) accounting for temporal changes in species composition at study sites due to migration, invasion, reintroduction and extinction and (2) handling the long-tailed distribution of records across species (that is, extreme imbalance in the number of images of different species; Extended Data Fig. 1). As discussed in the following, these issues limit the ability of current AI to accurately recognize species that are of notable interest to conservation practitioners.

Table 1 | Classification performance comparisons on validation sets of periods 1 and 2

Periods	Methods	Class average accuracy (%)	Class average accuracy on new classes (%)
1	Off-the-shelf model	81.2	-
2	Traditional transfer learning with full human annotation	75.8	63.9
	Our framework without semi-supervision and OLTR	69.2	61.2
	Our framework (semi-OLTR)	**77.2**	**68.1**

Bold indicates higher performance on the same inference set.

Changing species composition. A novel challenge for long-term surveys is that new species may be detected on cameras in subsequent seasons or years, either because the species are rare and undetected in previous survey periods[29] or because they are new to the system. Additionally, the species composition of ecological systems naturally varies through time through the process of succession[30]. Novel species are often of particular conservation concern, as they may represent recolonizing populations[31], reintroduced animals[32] or harmful invasive species[33,34].

In conventional deep learning, researchers focus on the performance of existing test data while ignoring the potential for future changes in data composition[35]. In other words, deep learning models typically require datasets to be fixed in the number of categories (in other words, they are static), whereas, in reality, long-term camera-trap datasets are not constrained to certain numbers of species (they are dynamic).

Fine-tuning models through transfer learning is currently the best solution when new species populate a study area[36]. However, this process requires full annotation of newly collected datasets, requiring a considerable amount of new human effort. This defeats the purpose of deep learning to reduce manual labour for long-term camera-trap monitoring.

Data from wildlife communities are long-tailed. Wildlife communities typically contain many individuals of several common species and few individuals of many rare species, resulting in camera-trap data with a long-tailed distribution. For example, in the dataset used for the project from Gorongosa National Park, Mozambique, ~50,000 images (>60% of animal images) are of baboons, warthogs and waterbucks, while only 22 images are of pangolins (a rare and protected species). This imbalance creates performance inconsistencies, because deep learning success is derived from balanced training datasets (for example, ImageNet[37]). For the Gorongosa dataset, a traditional deep learning approach resulted in only 60% accuracy for a category with only 41 images (serval) versus 88.8% performance for a species with 17,938 images (waterbuck). This is a major issue, because animals of particular conservation concern are typically rare[38], producing fewer images and therefore worse classification accuracy than for common species. If such species are always misclassified, the practical benefits of AI are limited.

An iteratively updating recognition system. To overcome these two major issues of (1) changing species community composition and (2) long-tailed species distributions, we designed a deep learning recognition framework that is updated iteratively using limited human intervention. Human annotation is needed whenever images of species novel to the AI model appear in the data. Our goal, therefore, becomes to minimize the need for human intervention as much as possible by applying human annotation solely on difficult images or novel species, while maximizing the recognition performance/accuracy of each model update procedure (that is, the update efficiency).

Traditionally, a deep learning model is applied to new batches of unannotated data collected during each time period to predict species classes. In our approach, we actively flag images that our model predicts with low confidence as novel or unknown species. These low-confidence predictions are then selected for human annotation, while high-confidence predictions are accepted as accurate and used as pseudo-labels for future model updates. The model is then updated (that is, retrained) based on both human annotations and pseudo-labels. To accommodate changing species communities, this procedure of active annotation and model update repeats each time new data are added to the collection (Fig. 1). In terms of long-tailed distribution, we use the open long-tailed recognition (OLTR)[7] method to balance the learning between abundant and scarce species. This component can reduce the number of predictions with low confidence from scarce species.

As a case study, using this new method we trained a model on a camera-trap dataset collected from Gorongosa National Park, Mozambique (details are provided in the Methods), and produced substantially improved model update efficiency over traditional transfer learning approaches. Specifically, using our approach, more than 80% human effort was saved on annotating new data, without sacrificing classification performance.

The dynamic nature of our algorithm maximizes learning and recognition efficiency by taking the best from both humans and machines within a synergistic collaboration, providing a framework that can be practically deployed for long-term camera-trap monitoring studies.

Iterative human and automated identification

In this section, we introduce the overview of our algorithm, data specification and experiment settings.

Algorithm overview. Our approach has two major components: (1) active selection with humans in the loop and (2) model update using active data annotations. For each time period when new data are collected, categories of images are predicted by deep learning models trained from previous periods with corresponding confidence levels. The model actively picks out low-confidence predictions for human annotation, while we accept high-confidence predictions as accurate, without further human verification. These predictions are used as pseudo-labels and included in the final dataset for further model updates or ecological analyses. The model is then updated (retrained) using both pseudo-labels and the newly acquired human annotations (implementation details are provided in the Methods).

After updating the model, we evaluate the model update efficiency and sensitivity to novel categories on a validation set. Specifically, we examine (1) the overall validation accuracy of each category after the update (that is, update performance), (2) the percentage of high-confidence predictions on validation (that is, saved human effort for annotation), (3) the accuracy of high-confidence predictions and (4) the percentage of novel categories that are detected as low-confidence predictions (that is, sensitivity to novelty). The optimization of the algorithm aims to minimize human efforts (that is, to maximize the high-confidence percentage) and to maximize model update performance and high-confidence accuracy.

Data specifications. *Data categories.* We manually identified a total of 55 categories (that is, species) in our data, including non-animal categories such as 'ghost' (misfired images lacking animals), 'setup' (images with a human setting up the cameras) and 'fire'. There were 630,544 images in total. A full list of these categories is provided in Extended Data Fig. 1, along with the number of images associated

Table 2 | Active selection performances of periods 1 and 2 with and without energy-based function

Periods	Inference sets	Confidence metrics	High-confidence ratio (%)	High-confidence accuracy (%)	Novel detection ratio (%)
1	Group 1 validation	Softmax	**80.9**	**91.5**	59.3
	Group 1 validation	Energy (ours)	79.5	91.1	**90.1**
2	Group 2 training	Energy (ours)	78.7	92.4	75.7
	Group 2 validation	Softmax	71.2	90.1	70.5
	Group 2 validation	Energy (ours)	**72.2**	**90.2**	**82.6**

Bold indicates higher performance on the same inference set.

with each category. Some vague categories that human annotators were unable to label accurately because of the varying quality of camera-trap images were also present, such as 'unknown antelope' and 'unknown bird'.

Two groups of training and validation sets. To ensure sufficient training and validation data, we initially identified 41 of the most abundant categories in our camera-trap dataset. The remaining 14 of the 55 categories were all tagged as 'unknown' and used to improve and validate the model's sensitivity to novel and difficult samples. We randomly split the 41 categories (by trigger events) into two groups of training and validation sets (26 categories in the first group of data and 41 in the second group) to mimic periodic data collection from two sequential time periods. Detailed training and validation split information is provided in the Methods.

Detailed pipeline for experiments. For experimental purposes, we separated our identification pipeline into two steps representing two time periods of data collection and the two groups of data curated in this project (Extended Data Fig. 3). The evaluation is focused on the second period when model update occurs. There are three major technical components in the framework: (1) energy-based loss[39], which improves the sensitivity to possible novel and difficult samples for active selection, (2) a pseudo-label-based semi-supervised procedure[40] for efficient model update from limited human annotations and (3) OLTR[7], which balances the learning of a long-tailed distribution.

Period 1. In the first period, we pre-trained an off-the-shelf model (ResNet-50 model[41]) using the first group of data. After training, we adopted the energy-based loss[39] and data from the 14 'left-out' categories to fine-tune the classifier so it was more sensitive to novel and difficult samples.

Period 2. In the second period, we first used the fine-tuned model from period 1 to produce high- and low-confidence predictions from group 2 training data, which were considered to be 'newly collected'. The confidence was calculated based on the Helmholtz free energy (details are provided in the Methods) of each prediction[39]. Novel and difficult samples were distinguished using a preset energy threshold. Then, low-confidence predictions were annotated by humans, while high-confidence predictions were accepted as pseudo-labels.

To update the model, we applied semi-supervised learning and OLTR, using both human annotations and pseudo-labels. Pseudo-label-based semi-supervised approaches iteratively update both the model and pseudo-labels until the best performance on the validation sets is achieved[40]. The use of pseudo-labels also enables the model to learn from the whole dataset instead of human annotated data only. On the other hand, OLTR approaches balance the learning between abundant and scarce categories through an embedding-space memory-based mechanism, where embedding memories of abundant categories are utilized to enhance the distinguishability of scarce categories that do not have enough samples to otherwise provide discriminative features[7]. The Methods provides details of these methods.

After the model was updated, the training sample from the 14 'left-out' categories was added to fine-tune the model's sensitivity to novel and difficult samples using energy-based loss as in period 1.

Future periods. Because the framework is designed to aid long-term data collection and monitoring projects, the framework does not stop at period 2. As time progresses, new data are collected. Users simply have to repeat the steps in period 2 to pick out and annotate difficult/novel samples to update the model. In addition, because the framework is fully modular, when new techniques are developed, parts of the framework can be easily replaced for better performance. For example, if there are better methods for novel category detection, energy-based loss and confidence calculation can be replaced with no effect on the conceptual framework.

Results

Period 1. In the first period, the model achieved an 81.2% average class accuracy on the validation set of group 1 (Table 1), 79.5% of the image predictions were high-confidence and, of these predictions, the accuracy was 91.1% (Table 2). In terms of novel categories, in the validation phase, the model successfully detected 90.1% of the novel samples belonging to the 14 categories that were left out of the training phase. In other words, 90.1% of the novel samples were predicted with low confidence. By contrast, direct softmax confidence (the most conventional way of calculating prediction confidence[42]) achieved a similar high-confidence accuracy as our model (91.5%), but only detected 59.3% novel samples.

Period 2. On group 2 training data, the model pre-trained from period 1 predicted 78.7% of images with high confidence, where the accuracy was 92.4%, while 75.7% of the new categories in group 2 training data were detected as low-confidence predictions (Table 2). As high-confidence predictions are trusted, 78.7% of human effort was saved in annotating group 2 training data because high-confidence predictions were accepted as accurate in our framework.

To update the model, group 2 training data that had been predicted with low confidence were checked by human experts and provided with manual annotations, and high-confidence samples were assigned model-predicted pseudo-labels. Overall, on the validation set of group 2, the model updated on both human annotations and pseudo-labels had an average class accuracy of 77.2% over the 41 categories. Compared to our method without human annotation (69.2%; Table 1), there was an 8% improvement. The model produced 72.2% high-confidence predictions at 90.2% accuracy in the high-confidence predictions of the validation set (Table 2) (see Table 3 for detailed per-category performances). In addition, it produced an 82.6% novel sample detection rate (that is, flagged as low-confidence predictions) from the validation data of the 14 left-out categories (last column of Table 1).

Table 3 | Classification performance comparisons of period 2 by category between our method and fully annotated transfer learning

		Traditional transfer learning with full human annotation		Our framework (semi-OLTR)	
	Species	**No. of human annotations**	**Accuracy (%)**	**No. of human annotations**	**Accuracy (%)**
Exist in groups 1 and 2	Ghost	20,500	**96.2**	4,248	90.2
	Waterbuck	17,938	**88.8**	2,079	82.4
	Baboon	15,660	**87.3**	2,335	81.1
	Warthog	17,400	**87.4**	4,224	79.7
	Bushbuck	6,622	**84.5**	2,179	72.3
	Impala	7,153	**84.0**	1,306	77.1
	Oribi	3,832	**83.8**	966	76.7
	Elephant	2,471	**88.2**	470	85.1
	Genet	1,976	**85.5**	888	84.0
	Nyala	1,569	73.9	434	**75.1**
	Setup	1,229	**87.4**	389	86.0
	Bushpig	1,040	83.1	377	**83.1**
	Porcupine	1,152	83.9	300	**88.3**
	Civet	699	82.9	123	**83.9**
	Vervet	739	73.2	263	**81.0**
	Reedbuck	740	65.8	203	**75.3**
	Kudu	556	70.9	161	**77.2**
	Buffalo	479	**89.0**	63	84.8
	Sable_antelope	323	85.2	48	**86.1**
	Duiker_red	370	86.8	116	**89.6**
	Hartebeest	394	**91.2**	63	84.6
	Wildebeest	303	**83.5**	44	82.4
	Guineafowl_helmeted	304	64.6	250	**74.4**
	Hare	214	78.8	166	**80.8**
	Duiker_common	194	62.7	92	**80.4**
	Fire	160	100.0	14	**100.0**
Exist in group 2 only	Mongoose_marsh	343	70.6	287	**71.8**
	Aardvark	235	77.6	128	**81.0**
	Honey_badger	234	60.3	190	**63.8**
	Hornbill_ground	203	**80.0**	161	72.0
	Mongoose_slender	165	68.0	157	**72.0**
	Mongoose_bushy_tailed	161	**74.0**	106	72.0
	Samango	99	58.0	48	**70.0**
	Mongoose_white_tailed	84	52.0	79	**64.0**
	Mongoose_banded	70	38.0	62	**52.0**
	Mongoose_large_grey	63	44.0	54	**48.0**
	Bushbaby	39	36.0	31	**50.0**
	Guineafowl_crested	46	95.0	35	**100.0**
	Eland	44	**90.0**	31	70.0
	Lion	42	70.0	32	**75.0**
	Serval	41	45.0	32	**60.0**

Bold indicates higher performance.

Comparison with traditional transfer learning. Our model (Tables 4 and 5) was substantially more data-efficient (that is, fewer data were required for the same performance) than traditional transfer learning methods in several respects (Fig. 2). Compared to traditional transfer learning, which uses full human annotations of group 2 training data, our method only involves human annotation

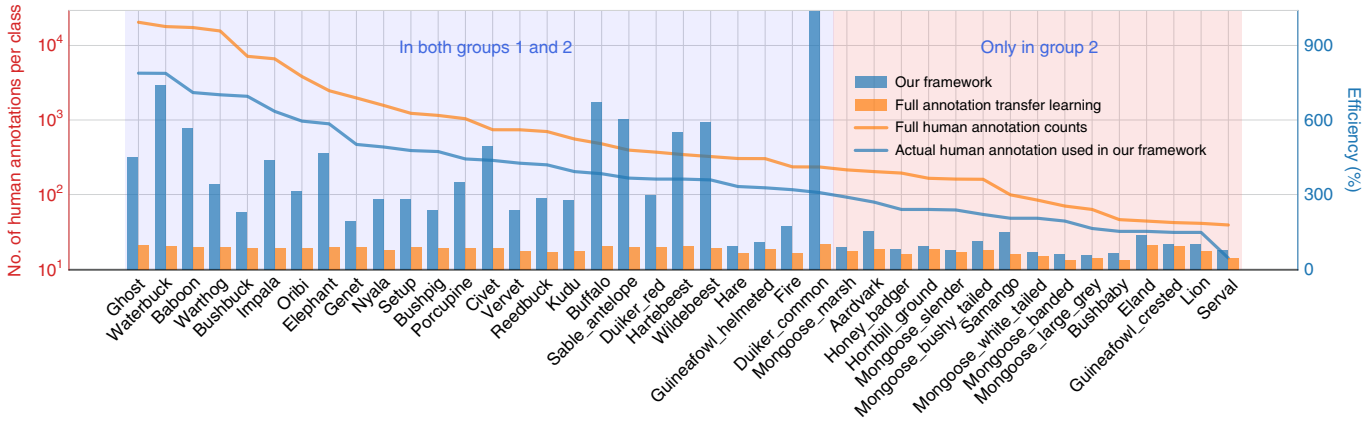


Fig. 2 | Label efficiency comparison with transfer learning on the group 2 validation set (ordered with respect to training sample size). To examine label efficiencies (a measure of accuracy given the number of annotations) after we updated our model in period 2, we calculated the validation accuracy over the percentage of used training annotations for each category. In other words, we define label efficiency as $\text{Efficiency}_i = \text{Validation accuracy}_i/(\text{no. of training annotations}_i/\text{no. of full annotations}_i)$, where i is the category index. The higher the value, the more efficient the model is at learning the corresponding categories and the fewer training data are needed to achieve comparable if not better performance of full manual annotations. We show the label efficiencies of all categories existing in the group 2 training and validation set. The blue bars represent our model's label efficiencies for each category. The orange bars represent baseline efficiencies for comparison, where full annotations were used with the traditional transfer learning method (that is, no. of training annotations$_i$/no. of full annotations$_i$ = 1). The blue and orange lines are annotation counts of each category, where orange represents full annotations and blue represents actually used human annotations in our period 2 model update procedure. For categories that exist in both the group 1 and 2 training sets (that is, known categories; on the left, with a blue background), the efficiency is substantially higher than the baselines across all categories. For categories that only exist in group 2 datasets (that is, they were absent in the group 1 training and validation set and are novel categories; on the right, with an orange background), the model is designed to use as much training data as possible because of the novelty of these categories. In other words, the no. of training annotations$_i$/no. of full annotations$_i$ of these categories is close to 1. Our model still has relatively higher efficiency than the full annotation transfer learning model across all the novel categories because our model had higher validation accuracy with a similar amount of training annotations.

of 21.3% of the group 2 samples. Even with less human annotation, our method still achieved a better overall class average accuracy (77.2% versus 75.8% for traditional transfer learning; Table 1). Our model also performed better than direct transfer learning for classifying the 15 new categories from group 2 (with an average of 4.2% accuracy improvement; Table 3).

Practical deployment. Our new framework showcases the powerful potential of deep learning for long-term ecological applications while employing a novel practical approach that greatly reduces the manual annotation burden. To validate the practical benefits, we deployed the model to classify a new set of data gathered from the same camera-trap monitoring sites (Gorongosa National Park, Mozambique) after group 1 and 2 datasets were collected (details are provided in the Methods). The new dataset is unannotated, unanalysed and contains 623,333 images in total. Images were predicted with the same active selection procedure, and 78.7% of the predictions were considered high-confidence. Thus, only 21.3% of these newly collected data required human annotation (or 78.7% of the human effort; ultimately, annotation cost was saved).

To validate the robustness of the model performance, two experts (K.M.G. and M.S.P.) confirmed the accuracy of 1,000 randomly selected high-confidence predictions (that is, those that were accepted as accurate). Our model predictions are 88.6% accurate with respect to expert classifications. Statistically, ~88% automatic accuracy is already sufficient to help alleviate the data bottleneck encountered in typical camera-trap monitoring projects compared to expert accuracy.

In terms of future model updates, the model can be further updated and validated on the new dataset using the same procedure as for period 2, where a new validation set can be created using a mix of previous validation sets (validation of groups 1 and 2) and the newly acquired human annotations. In addition, the same random verification by human experts on high-confidence predictions can be applied to avoid performance corruptions (that is, increased misclassifications in high-confidence predictions).

Table 4 | List of the augmentation methods and corresponding parameters we used on our training data

Augmentations	Parameters	Values
Random resize crop	Dimension	224 × 224
	Range of crop scale	~0.08–1.0
	Range of crop aspect ratio	~0.8–1.2
Random grey scale	Probability	0.1
Random horizontal flip	Probability	0.5
Random rotation	Probability	0.5
	Rotation degree	45
Colour jittering	Brightness jittering	0.4
	Contrast jittering	0.4
	Saturation jittering	0.4
	Hue jittering	0.1
Normalization	Mean	[0.485, 0.456, 0.406]
	Std	[0.229, 0.224, 0.225]

Invasive and recolonizing species. One of the nontable advances made by our framework is the ability to flag new or rare species that may have particular conservation importance. Our new dataset contained two novel species (leopard and African wild dog) to

Table 5 | List of hyperparameters of our framework used in the two-period experiments

Period	Parameters	Values
Period 1: Training	Baseline architecture	ResNet-50
	Training epochs	40
	Batch size	64
	Initial learning rate (feature)	0.001
	Initial learning rate (classifier)	0.01
	Learning rate decay epochs	10
	Learning rate decay ratio	0.1
	Momentum	0.9
	Weight decay	0.0005
Period 1: Energy fine-tuning	Training epochs	10
	Batch size	96
	Known:unknown ratio	1:2
	Energy loss weight	0.01
	Initial learning rate (feature)	0.00001
	Initial learning rate (classifier)	0.0001
	Confidence threshold (τ)	13.7
	Energy temperature	1.5
Period 2: Updating	Baseline architecture	ResNet-50 + OLTR
	Semi-repeats	3
	Epochs in each repeat	30
	Batch size	64
	Pseudo-label (%)	50
	Initial learning rate of each repeat (feature)	0.0001
	Initial learning rate of each repeat (classifier)	0.01
	Initial learning rate of each repeat (memory)	0.0001
	Learning rate decay epochs	10
	Learning rate decay ratio	0.1
	Momentum	0.9
	Weight decay	0.0005
Period 2: Energy fine-tuning	Training epochs	10
	Batch size	96
	Known:unknown ratio	1:2
	Energy loss weight	0.01
	Initial learning rate (feature)	0.000001
	Initial learning rate (classifier)	0.00001
	Initial learning rate (memory)	0.000001
	Confidence threshold (τ)	6.7
	Energy temperature	0.06

test the model's sensitivity to novel categories. The former naturally recolonized the study area and the latter were reintroduced as a part of ongoing conservation efforts. There were 24 and 5 images for African wild dogs and leopards, respectively. The model successfully detected 20 (83.3%) African wild dog images and four (80.0%) leopard images, demonstrating its capacity to recognize important novel species in continuous monitoring periods.

Discussion

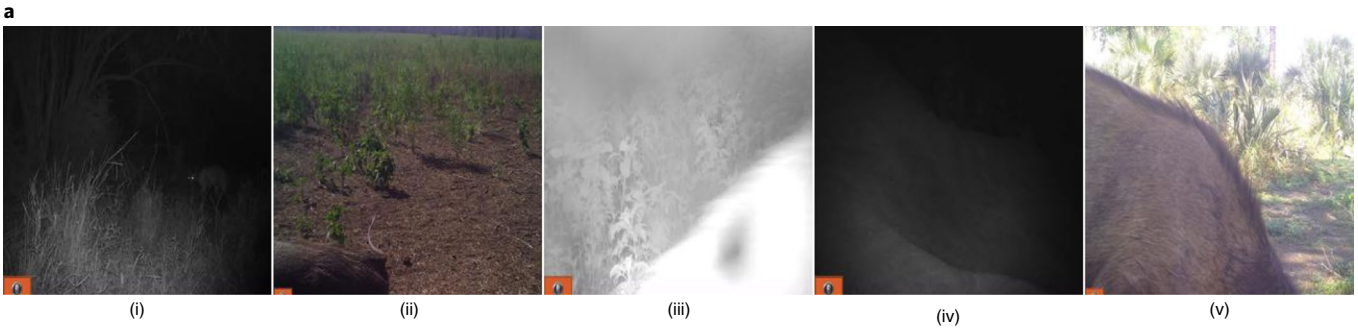
Failure cases. Two types of failure occur in our framework: (1) low-confidence predictions that are not novel species and (2) high-confidence predictions that differ from human-supplied annotations (Fig. 2).

There are several ways in which our model was unable to accurately identify samples from known species with high confidence (Fig. 3a). A common reason for low-confidence predictions was difficulty distinguishing animals from the background. For example, Fig. 3a(i) depicts an antelope obscured by darkness at night, making it difficult for the model to classify with confidence. However, rather than making a misclassification as would occur in traditional AI approaches[41], our model considers the low accuracy of the prediction and flags the image for review or similar. In our approach, these difficult samples are flagged as low-confidence predictions for further human evaluation (annotation) rather than assigned random labels — a practice that can potentially bias further data analysis and inference.

In the second type of model failure, images predicted with high confidence differ from the original annotations (Fig. 3b). We note that these images were originally classified by volunteers who were trained but may not have correctly annotated all samples as accurately as wildlife experts. Surprisingly, most of the confident predictions are proven to be correct after reevaluation by human experts (K.M.G. and M.S.P.). For example, Fig. 3b(iv) was originally labelled as a warthog, although there is no warthog present. However, there is a vervet monkey in the lower left of the frame that was missed by the human classifiers. The model not only detects the previously unobserved animal but also correctly identifies the species.

Thus, these 'failures' actually demonstrate the robustness and flexibility of our framework. As both human annotations and machine predictions can be wrong, a mutual interaction between human and machine can benefit the long-term performance of the recognition system. For example, picking out low-confidence samples like those in Fig. 3b prevents the production of low-quality predictions that can cause bias in camera-trap analyses. Furthermore, applying validated human annotations on these samples can help improve the identification capacity of the model as it needs to recognize more difficult samples during model updates. On the other hand, when the model is highly confident, it can be more accurate than average human annotators, as evidenced by the examples given in Fig. 3b(ii),(iv),(v). In other words, some of the human mistakes are prevented, such that the annotation quality for future model update and camera-trap analyses is improved. On the other hand, as we acknowledge in some cases the model will make incorrect high-confidence classifications, we can apply periodic random verification by human experts on high-confidence predictions (similar to what we did in the 'Practical deployment' section) to ensure that these errors do not propagate through repeated training.

The need for humans in the loop. Our framework demonstrates the unique merit of combining machine intelligence and human intelligence. As Fig. 3c illustrates, machine intelligence, when trained on large datasets to distil visual associations and class similarities, can quickly match visual patterns with high confidence[37]. Human intelligence, on the other hand, excels at being able to recognize fragmented samples based on prior experience, context clues and additional knowledge. Increasingly, we are moving towards applying computer vision systems to real-world scenarios, with unknown classes[7], unknown domains[8] and constantly updating environments. It is therefore crucial to develop effective algorithms that can handle dynamic data streams. Humans in the loop provide a natural and



effective way to integrate the two types of perceptual ability (that is, human and machine), resulting in a synergism that improves the efficiency and the overall recognition system.

Extensions and future directions. Our framework is fully modular and can be easily upgraded with more sophisticated model designs. For example, models with deeper networks can be employed

Fig. 3 | Failure cases. a, Examples of low-confidence predictions. In most cases, the model has low confidence on images with distorted, partially visible (ii–v) or obscured animals (i). It can be incredibly difficult, if not impossible, for either humans or machines to accurately identify the animal species. **b**, Examples of high-confidence predictions that did not match the original annotations. Many high-confidence predictions that were flagged as incorrect based on validation labels (provided by students and citizen scientists) were in fact correct upon closer inspection by wildlife experts (K.M.G. and M.S.P.). For example, in (i), an empty image, originally mislabelled as baboon, was correctly classified by our method as empty. In (ii), although the animal is distant from the camera in a dark environment, the model successfully identifies hartebeest, while the human-supplied label is 'unknown antelope'. In (iii), the model successfully identifies the elephant only based on the trunk and leg, while human volunteers originally classified the image as 'unknown'. In (iv), a vervet monkey is correctly detected and classified in an image originally (incorrectly) labelled as warthog by human annotators. Panel (v) was originally classified as unknown by human annotators, but, based on the body shape and white markings on the rear, the model can correctly recognize the animal as bushbuck. Panel (vi) is an example where multiple species are in the same scene. Although the model does not have the capacity to deal with multi-species samples, as baboon is obviously the major component of this image, the prediction is reasonable. On the other hand, these examples above do not mean that the model always makes correct predictions when highly confident. Panels (vii) and (viii) are two typical examples where the model makes mistakes due to the obscured nature of these images. Red text indicates wrong and green text indicates correct. **c**, Two examples of image retrieval based on feature space similarity. Machine intelligence largely depends on visual similarity associations learned from large-scale datasets to classify animal species. These two examples illustrate image retrieval based on the Euclidean distances of the feature vectors (that is, outputs of the global average pooling layer of the ResNet model used in the project, which is of dimension 2,048 in Euclidean space). For each anchor image (the leftmost image of each row), we show the five closest (that is, most similar) samples in terms of Euclidean distance within the validation set of group 2. Green colour means correct predictions and red means wrong predictions (based on the original annotations). For example, in sequence (i), samples with similar visual appearance are usually from the same species (waterbuck). However, in sequence (ii), the two most similar images (according to our model) to the banded mongoose anchor image are actually not banded mongoose but slender mongoose. The model misclassified these two samples based on their similarities to the other banded mongoose images.

for better classification generalization, more sophisticated semi-supervised training protocols can be adopted for better learning from pseudo-labels, and better novelty detection techniques can be used for better active selection.

Future directions include extending our framework to handle multi-label and multi-domain scenarios. The current approach was developed for single-label recognition (that is, each image only represents one single species). In real-world camera-trap set-ups, it would be desirable to recognize multiple species within the same view. Furthermore, our framework is expected to be deployed in diverse locations with different landscapes. Therefore, our methodology can be more scalable with the ability to handle multiple environmental domains than existing methodologies. In addition, our method will be incorporated in a user-friendly interface, such that users without knowledge of Python can use it.

Methods

Data collection and annotation. The camera-trap data came from the WildCam Gorongosa long-term research and monitoring programme in Gorongosa National Park, Mozambique (18.8154° S, 34.4963° E)[43]. The data used in this study are from 2016–2019. Cameras were located in a mix of grassland, open woodland and closed forest habitats. K.M.G. placed 60 motion-activated Bushnell TrophyCam and Essential E2 cameras in a 300-km^2 area in the southern area of the 3,700-km^2 park. Each camera was mounted on a tree within 100 m of the centre of a 5-km^2 hexagonal grid cell, facing an animal trail or open area with signs of animal activity. Cameras were set in shaded, south-facing sites that were clear of tall grass to reduce false triggers. Cameras took two photographs per detection (henceforth called a 'trigger event') with an interval of 30 s between trigger events. There were 630,544 images in total. The data distribution with respect to categories is reported in Extended Data Fig. 1. In terms of the data split for experimental purposes, detailed distributions of both group 1 and 2 are reported in Extended Data Fig. 2.

Data split. The dataset was randomly split into two groups of training and validation sets to mimic periodic data collection from two sequential time periods, along with an additional 'unknown' set for improving and validating the model's sensitivity to novel and difficult samples. Because we set the cameras to capture one pair of images for each trigger event, image pairs within the same event were usually similar in appearance. To reduce bias, we split the dataset based on camera trigger events, such that both images in a paired trigger event were either in the training or testing set. The training–testing split did not account for camera locations (that is, images from a given camera were present in both testing and training sets). For large-scale, long-term projects, it is more likely that the camera locations are stable. In our study, the cameras cover most of the landscapes in the monitoring area and include a diversity of background types that change seasonally throughout the year. Possible distribution shifts in our dataset solely come from temporal animal community changes instead of spatial landscape/ecosystem changes.

The first group contained the 26 most abundant categories, and the second period contained all 41 categories. We randomly divided each period into training (80% of samples) and validation (20% of samples) sets. For scarce categories that had fewer than 80 images (for example, crested guineafowl, eland, lion and serval), we randomly selected 20 samples instead of 20% of the data to ensure the quality of validation. The labels and distributions of these two groups of data are illustrated in Extended Data Fig. 2.

Within the 14 categories that are tagged 'unknown', we randomly selected 80% of data to fine-tune the model's sensitivity to novel and difficult samples. We then used the rest of the sample from the 14 categories as an extra validation set to evaluate the model's novel image detection capacity.

Implementation details. In this section, we report the implementation details of our method. It was developed with Python as the programming language with PyTorch[44] as the deep learning framework. The detailed experimental pipeline is illustrated in Extended Data Fig. 3.

Data pre-processing. All of the images used in this project were first resized to dimensions of 256×256. For training inputs, these images were randomly cropped and resized to 224×224. For validation and inference inputs, images were centre cropped to 224×224. Table 4 presents the list of data augmentations used for training and corresponding hyperparameters.

Period 1 and baseline model training. There are two steps in this period: (1) baseline model training on group 1 data and (2) classifier fine-tuning using the 14 left-out categories for better sensitivity to novel and difficult samples.

For the baseline model we used ResNet-50[41]. This was pre-trained on ImageNet[37], a generalized object oriented dataset for model weight initialization. The pre-trained model was then trained on group 1 training data, which had 26 categories. All the hyperparameters are provided in Table 5. Model weights with the best validation performance on group 1 validation data were saved as the best model.

After training on group 1 data, we used energy-based loss[39] and the 14 left-out categories (tagged as 'unknown') to fine-tune the classifier for better sensitivity to novel and difficult samples. The energy-based loss was calculated as

$$\begin{aligned} L_{\text{energy}} &= \mathbb{E}_{x_{\text{known}} \sim \mathfrak{D}^{\text{train}}_{\text{known}}} \left(\max\left(0, E(x_{\text{known}}) - m_{\text{known}}\right)^2\right. \\ &+ \mathbb{E}_{x_{\text{unknown}} \sim \mathfrak{D}^{\text{train}}_{\text{unknown}}} \left(\max\left(0, m_{\text{unknown}} - E(x_{\text{unknown}})\right)^2\right. \end{aligned} \tag{1}$$

$$E(x) = -T \log \sum_{i}^{N} \mathrm{e}^{(f(x_i)/T)} \tag{2}$$

where $\mathbb{E}$ is expectation and x_{known} and x_{unknown} are samples from group 1 and samples from 14 unknown categories, respectively. $\mathfrak{D}^{\text{train}}_{\text{known}}$ and $\mathfrak{D}^{\text{train}}_{\text{unknown}}$ represents datasets of group 1 and 14 unknown categories. $E(\cdot)$ is the Helmholtz free energy, calculated as the log sum of outputs from the network. $f(\cdot) : \mathbb{R}^{D \times D} \rightarrow \mathbb{R}^K$ is the network that maps $D \times D$ images to K-dimensional vectors. T is the temperature that regularizes the energy. m_{known} and m_{unknown} are two margins applied on known and unknown energy.

During fine-tuning, both cross-entropy loss and energy-based loss are tuned. Equation (3) is the final loss, where w is the weight applied on energy-based loss:

$$L = L_{\text{cross\_entropy}} + w\,L_{\text{energy}} \quad (3)$$

All hyperparameters are reported in Table 5.

Period 2 and model update. Active selection and confidence calculation. Following ref. [39], confidence for active selection is calculated based on the Helmholtz free energy (equation (2)). Based on a preset energy threshold τ, predictions are separated into high- and low-confidence categories. In other words, predictions are considered confident if $-E(x) > \tau$ and vice versa. Based on prediction confidence, low-confidence predictions are assigned human annotations and high-confidence predictions are utilized as initial pseudo-labels for semi-supervised learning.

Pseudo-labels and semi-supervised learning. Pseudo-label semi-supervision utilizes both human annotations and pseudo-labels to update the model. In the original approach, where models are randomly initialized, pseudo-labels are updated throughout training iterations[40]. In other words, at each iteration, the model predicts samples without human annotations and uses these predictions as pseudo-labels to train the same samples with a stronger set of data augmentations. In our approach, as the pseudo-labels usually have higher quality than random predictions, we set three semi-update repeats and only update the pseudo-labels at the beginning of each repeat using the best model from the last repeat. Specifically, within each semi-update repeat, the model is updated with a fixed set of pseudo-labels and a number of training epochs. Model weights with the best validation performance are saved, and at the end of the repeat, the best model is used to predict samples without human annotations to produce a new set of pseudo-labels, and a new repeat is started. Only model weights with the best validation performance throughout the three repeats are saved, and the number of repeats is a hyperparameter that can be tuned using validation data. Other hyperparameters are provided in Table 5.

OLTR. OLTR is an additional component in our framework targeting the long-tailed distribution of classes in the datasets. Generally speaking, it uses embedding-level memory of each category to enhance the distinguishability of scarce categories. It is based on the idea that a lot of the mid-level visual features (that is, feature embedding) are shared between similar categories (for example, most of the antelopes share similar body shapes). Because the model can usually learn high-quality feature embedding from abundant species, through memory selection techniques the model is able to select relevant feature embedding to help improve the distinguishability of scare categories. We directly apply OLTR into our framework. For a detailed explanation of OLTR, see ref. [7].

Comparison to unsupervised and self-supervised learning. Although unsupervised learning and self-supervised learning have recently made substantial progress[45,46] in learning without human annotations, these learning methods still have difficulties handling novel categories and categories with trivial differences (that is, fine-grained categories)[47]. This is because current unsupervised and self-supervised learning methods rely on human-defined random data augmentation (for example, cropping and rotation) to mimic intra- and interclass variations, while real-world novel and fine-grained categories often possess complex intra- and interclass distributions. In this work we advocate the use of humans in the loop to provide valuable supervision in a data-efficient manner. Together with semi-supervised learning, our framework can reliably recognize new species with only sparse human annotations.

Additional results. Detailed results of model update performance are listed by category in Table 3.

Received: 3 February 2021; Accepted: 22 August 2021;
Published online: 18 October 2021

References

1. Steenweg, R. et al. Scaling-up camera traps: monitoring the planet's biodiversity with networks of remote sensors. *Front. Ecol. Environ.* **15**, 26–34 (2017).
2. Rich, L. N. et al. Assessing global patterns in mammalian carnivore occupancy and richness by integrating local camera trap surveys. *Global Ecol. Biogeogr.* **26**, 918–929 (2017).
3. Barnosky, A. D. et al. Has the Earth's sixth mass extinction already arrived? *Nature* **471**, 51–57 (2011).
4. Ahumada, J. A. et al. Wildlife insights: a platform to maximize the potential of camera trap and other passive sensor wildlife data for the planet. *Environ. Conserv.* **47**, 1–6 (2020).
5. Norouzzadeh, M. S. et al. Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning. *Proc. Natl Acad. Sci.* **115**, E5716–E5725 (2018).
6. Miao, Z. et al. Insights and approaches using deep learning to classify wildlife. *Sci. Rep.* **9**, 8137 (2019).
7. Liu, Z. et al. Large-scale long-tailed recognition in an open world. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition* 2537–2546 (IEEE, 2019).
8. Liu, Z. et al. Open compound domain adaptation. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition* 12406–12415 (IEEE, 2020).
9. Hautier, Y. et al. Anthropogenic environmental changes affect ecosystem stability via biodiversity. *Science* **348**, 336–340 (2015).
10. Barlow, J. et al. Anthropogenic disturbance in tropical forests can double biodiversity loss from deforestation. *Nature* **535**, 144–147 (2016).
11. Ripple, W. J. et al. Conserving the world's megafauna and biodiversity: the fierce urgency of now. *Bioscience* **67**, 197–200 (2017).
12. Dirzo, R. et al. Defaunation in the Anthropocene. *Science* **345**, 401–406 (2014).
13. O'Connell, A. F., Nichols, J. D. & Karanth, K. U. *Camera Traps in Animal Ecology: Methods and Analyses* (Springer Science & Business Media, 2010).
14. Burton, A. C. et al. Wildlife camera trapping: a review and recommendations for linking surveys to ecological processes. *J. Appl. Ecol.* **52**, 675–685 (2015).
15. Kays, R., McShea, W. J. & Wikelski, M. Born-digital biodiversity data: millions and billions. *Divers. Distrib.* **26**, 644–648 (2020).
16. Swanson, A. et al. Snapshot Serengeti, high-frequency annotated camera trap images of 40 mammalian species in an African savanna. *Sci. Data* **2**, 1–14 (2015).
17. Ahumada, J. A. et al. Community structure and diversity of tropical forest mammals: data from a global camera trap network. *Philos. Trans. R. Soc. B Biol. Sci.* **366**, 2703–2711 (2011).
18. Pardo, L. E. et al. Snapshot Safari: a large-scale collaborative to monitor Africa's remarkable biodiversity. *South Africa J. Sci.* https://doi.org/10.17159/sajs.2021/8134 (2021).
19. Anderson, T. M. et al. The spatial distribution of African savannah herbivores: species associations and habitat occupancy in a landscape context. *Philos. Trans. R. Soc. B Biol. Sci.* **371**, 20150314 (2016).
20. Palmer, M., Fieberg, J., Swanson, A., Kosmala, M. & Packer, C. A 'dynamic' landscape of fear: prey responses to spatiotemporal variations in predation risk across the lunar cycle. *Ecol. Lett.* **20**, 1364–1373 (2017).
21. Tabak, M. A. et al. Machine learning to classify animal species in camera trap images: applications in ecology. *Methods Ecol. Evol.* **10**, 585–590 (2019).
22. Whytock, R. C. et al. Robust ecological analysis of camera trap data labelled by a machine learning model. *Methods Ecol. Evol* **12**, 1080–1092 (2021).
23. Beery, S., Van Horn, G. & Perona, P. Recognition in terra incognita. In *Proc. European Conference on Computer Vision* (*ECCV*) 456–473 (IEEE, 2018).
24. Tabak, M. A. et al. Improving the accessibility and transferability of machine learning algorithms for identification of animals in camera trap images: MLWIC2. *Ecol. Evol.* **10**, 10374–10383 (2020).
25. Shahinfar, S., Meek, P. & Falzon, G. How many images do I need? Understanding how sample size per class affects deep learning model performance metrics for balanced designs in autonomous wildlife monitoring. *Ecol. Inform.* **57**, 101085 (2020).
26. Norouzzadeh, M. S. et al. A deep active learning system for species identification and counting in camera trap images. *Methods Ecol. Evol.* **12**, 150–161 (2020).
27. Willi, M. et al. Identifying animal species in camera trap images using deep learning and citizen science. *Methods Ecol. Evol.* **10**, 80–91 (2019).
28. Schneider, S., Greenberg, S., Taylor, G. W. & Kremer, S. C. Three critical factors affecting automated image species recognition performance for camera traps. *Ecol. Evol.* **10**, 3503–3517 (2020).
29. Kays, R. et al. An empirical evaluation of camera trap study design: how many, how long and when? *Methods Ecol. Evol.* **11**, 700–713 (2020).
30. Prach, K. & Walker, L. R. Four opportunities for studies of ecological succession. *Trends Ecol. Evol.* **26**, 119–123 (2011).
31. Mech, L. D., Isbell, F., Krueger, J. & Hart, J. Gray wolf (*Canis lupus*) recolonization failure: a Minnesota case study. *Can. Field-Nat.* **133**, 60–65 (2019).
32. Taylor, G. et al. Is reintroduction biology an effective applied science? *Trends Ecol. Evol.* **32**, 873–880 (2017).
33. Clavero, M. & Garcia-Berthou, E. Invasive species are a leading cause of animal extinctions. *Trends Ecol. Evol.* **20**, 110 (2005).
34. Caravaggi, A. et al. An invasive-native mammalian species replacement process captured by camera trap survey random encounter models. *Remote Sens. Ecol. Conserv.* **2**, 45–58 (2016).
35. Arjovsky, M., Bottou, L., Gulrajani, I. & Lopez-Paz, D. Invariant risk minimization. Preprint at https://arxiv.org/abs/1907.02893 (2019).
36. Yosinski, J., Clune, J., Bengio, Y. & Lipson, H. How transferable are features in deep neural networks? In *Advances in Neural Information Processing Systems* 3320–3328 (IEEE, 2014).
37. Deng, J. et al. ImageNet: a large-scale hierarchical image database. In *Proc. 2009 IEEE Conference on Computer Vision and Pattern Recognition* 248–255 (IEEE, 2009).
38. Pimm, S. L. et al. The biodiversity of species and their rates of extinction, distribution and protection. *Science* https://doi.org/10.1126/science.1246752 (2014).

39. Liu, W., Wang, X., Owens, J. & Li, Y. Energy-based out-of-distribution detection. In *Advances in Neural Information Processing Systems* (eds Larochelle, H. et al.) 21464–21475 (Curran Associates, 2020).
40. Lee, D.-H. Pseudo-label: the simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on Challenges in Representation Learning, ICML*, Vol. 3 (2013).
41. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition* 770–778 (IEEE, 2016).
42. Hinton, G., Vinyals, O. & Dean, J. Distilling the knowledge in a neural network. Preprint at https://arxiv.org/abs/1503.02531 (2015).
43. Gaynor, K. M., Daskin, J. H., Rich, L. N. & Brashares, J. S. Postwar wildlife recovery in an African savanna: evaluating patterns and drivers of species occupancy and richness. *Anim. Conserv.* **24**, 510–522 (2020).
44. Paszke, A. et al. in *Advances in Neural Information Processing Systems* Vol. 32 (eds Wallach, H. et al.) 8024–8035 http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf (Curran Associates, 2019).
45. Chen, T., Kornblith, S., Norouzi, M. & Hinton, G. A simple framework for contrastive learning of visual representations. Preprint at https://arxiv.org/abs/2002.05709 (2020).
46. He, K., Fan, H., Wu, Y., Xie, S. & Girshick, R. Momentum contrast for unsupervised visual representation learning. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition* 9729–9738 (IEEE, 2020).
47. Xiao, T., Wang, X., Efros, A. A. & Darrell, T. What should not be contrastive in contrastive learning. Preprint at https://arxiv.org/abs/2008.05659 (2020).

Acknowledgements

We thank T. Gu, A. Ke, H. Rosen, A. Wu, C. Jurgensen, E. Lai, M. Levy and E. Silverberg for annotating the images used in this study, as well as everyone else involved in this project. Data collection was supported by J. Brashares and through grants to K.M.G. from HHMI BioInteractive, the Rufford Foundation, Idea Wild, the Explorers Club and the UC Berkeley Center for African Studies. We are grateful for the support of Gorongosa National Park, especially M. Stalmans, in permitting and facilitating this research. Z.L. is supported by NTU NAP. K.M.G. is supported by Schmidt Science Fellows in partnership with the Rhodes Trust, and the National Center for Ecological Analysis and Synthesis Director's Postdoctoral Fellowship. M.S.P. is funded by National Science Foundation grant no. PRFB #1810586.

Author contributions

This study was conceived by Z.M., Z.L., K.M.G. and M.S.P. The methods were designed by Z.M. and Z.L. Code was written by Z.M., and the computations were undertaken by Z.M. with help from Z.L. The main text was drafted by Z.M. and Z.L., with contributions, editing and comments from all authors. K.M.G. and M.S.P. collected all data and oversaw annotation. Z.M. created all figures and tables in consultation with W.M.G., Z.L. and S.X.Y.

Competing interests

The authors declare no competing interests.

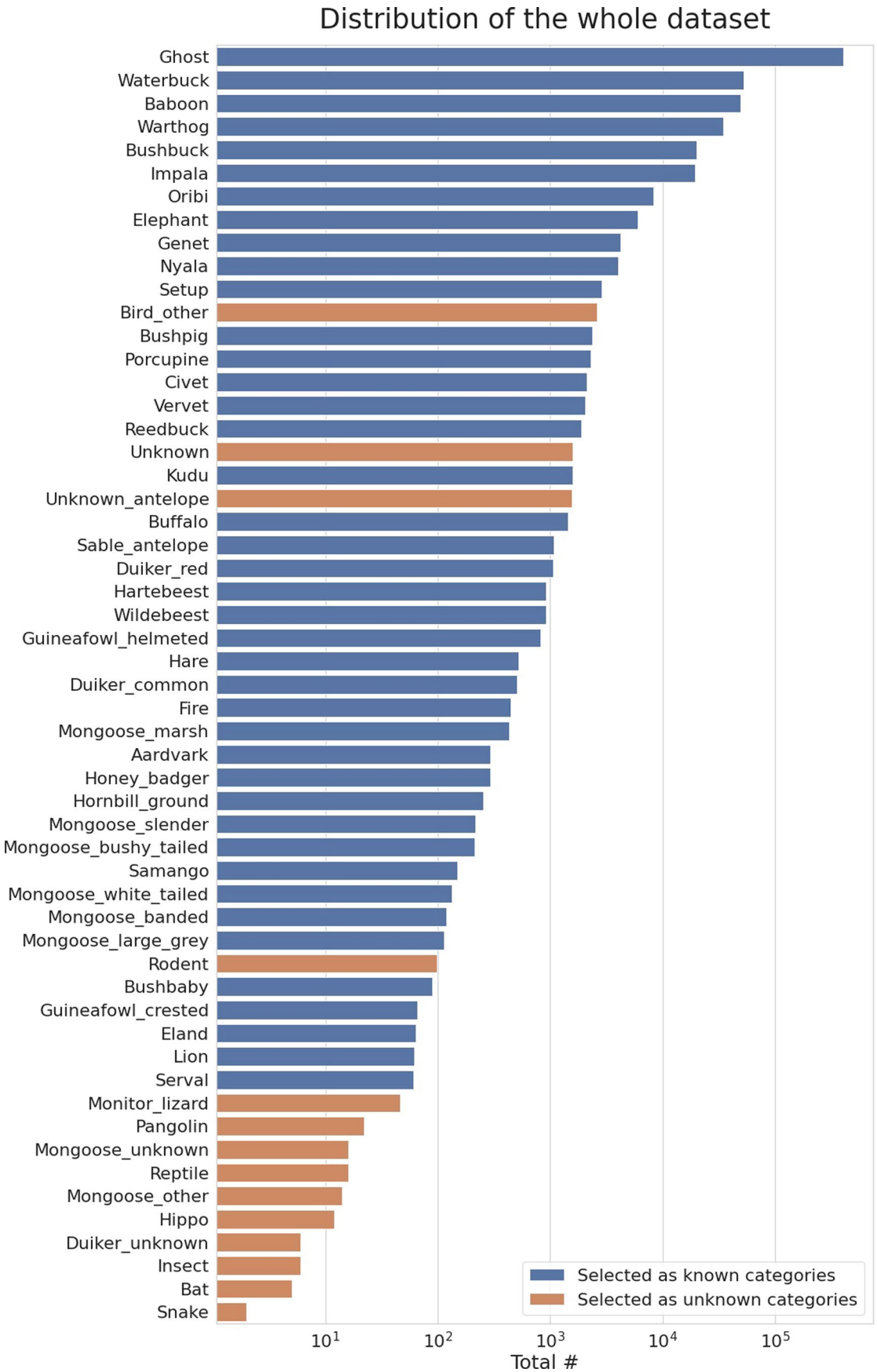
Additional information

Extended data is available for this paper at https://doi.org/10.1038/s42256-021-00393-0.

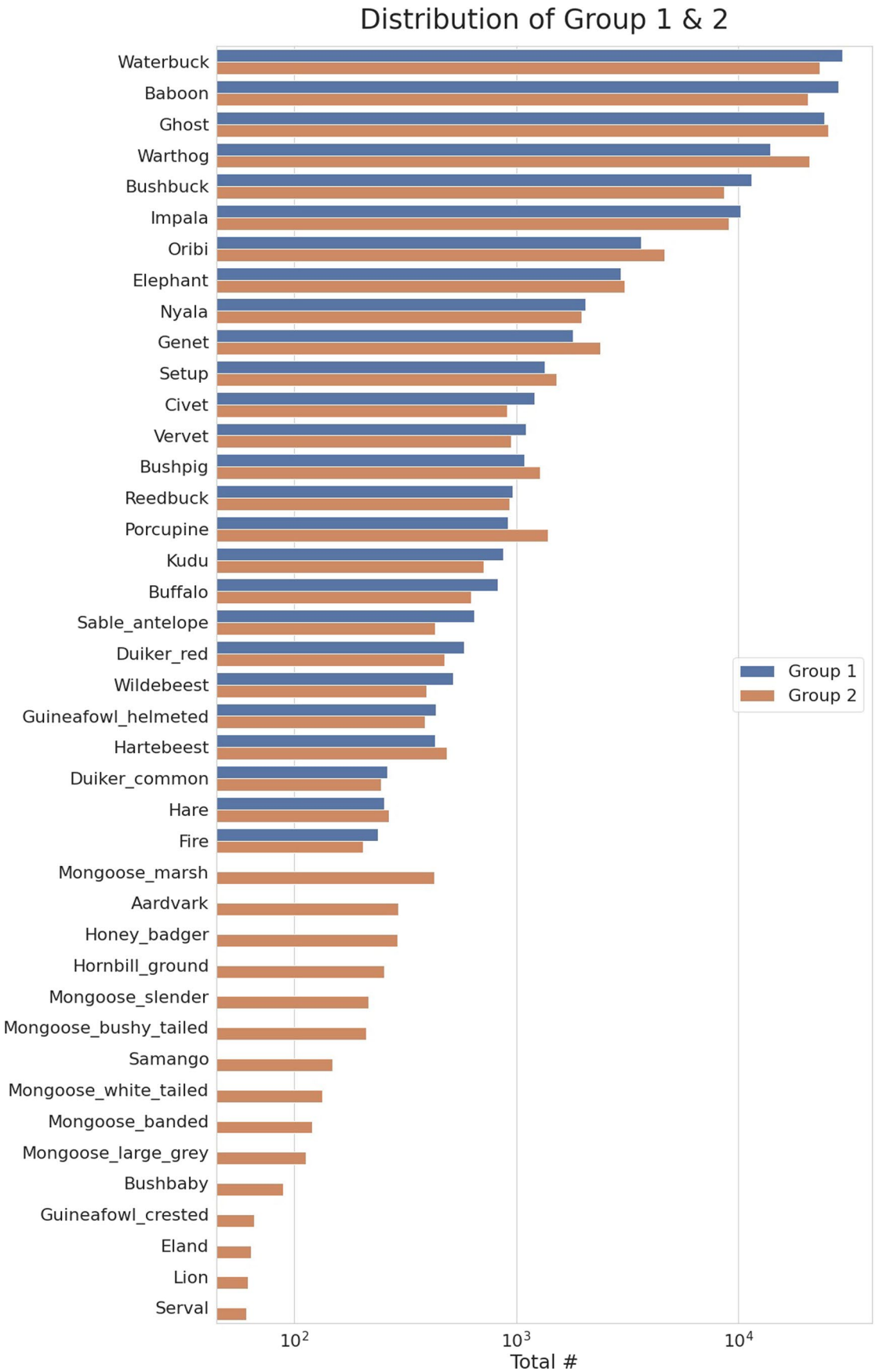
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42256-021-00393-0.

Correspondence and requests for materials should be addressed to Zhongqi Miao or Wayne M. Getz.

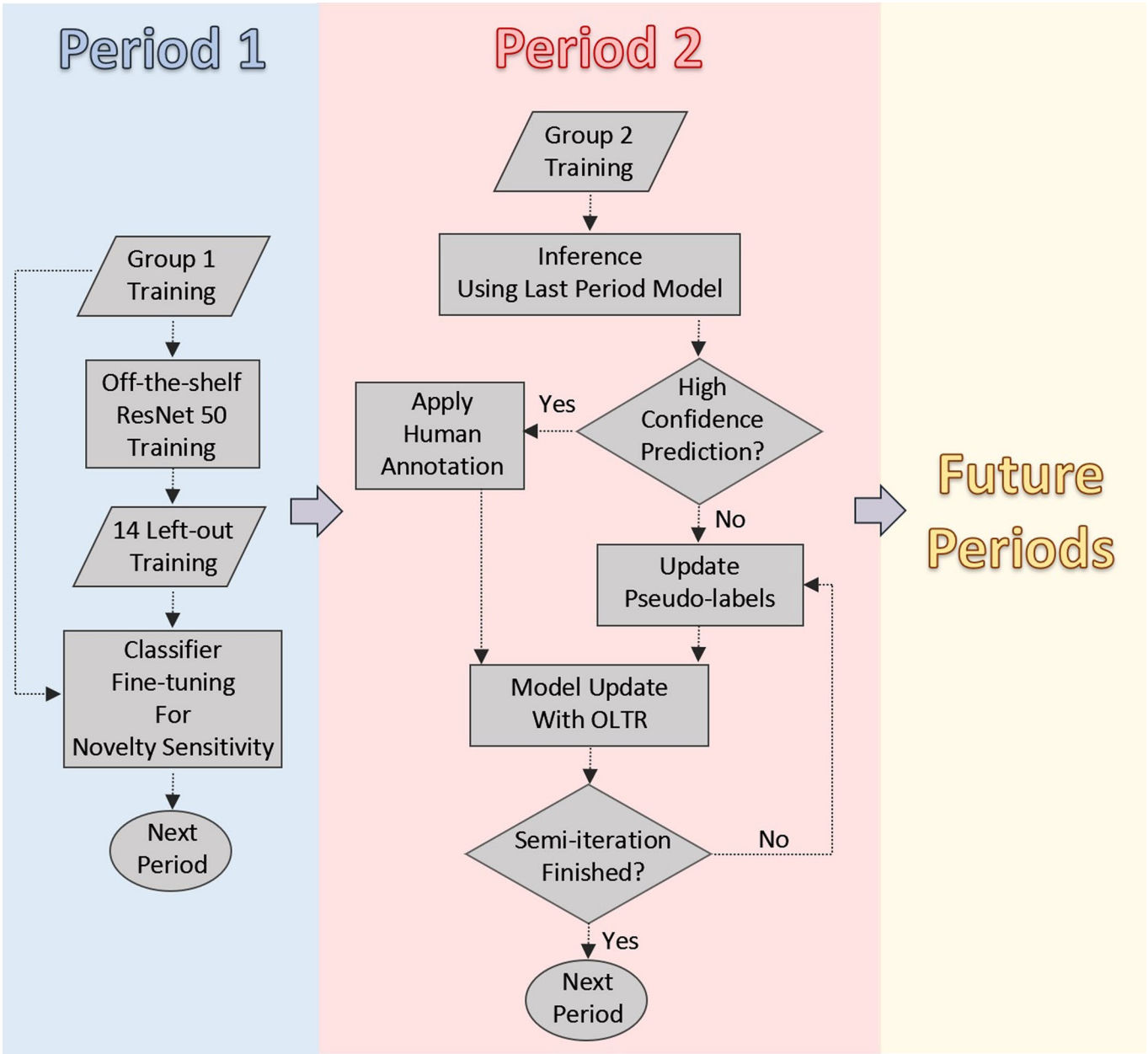
Peer review information *Nature Machine Intelligence* thanks Dan Morris and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.



Extended Data Fig. 1 | The distribution of images across species in the entire camera trap data set. There are 55 categories in total. 14 categories were tagged as "unknown" (colored in orange) and used to improve and validate our model's sensitivity to novel and difficult samples.



Extended Data Fig. 2 | The distribution of species across the two groups of data. We split the data set into two groups to mimic two sequential data collection seasons. In the first group, there are 26 categories (colored in blue). The second group has 41 categories. Group 1 is used in the first period experiment to train a baseline model, and Group 2 is used in the second period experiment to test and update the model.



Extended Data Fig. 3 | The overall experimental workflow of our framework. In the first time step, a baseline model is trained using group 1 training data with only 26 categories. Next, the classifier is fine-tuned using the 14 unknown categories and energy-based loss to increase the sensitivity to out-of-distribution categories. After the classifier is fine-tuned, the classifier is then used to predict classifications for group 2 training data. Here, high-confidence predictions are trusted while low-confidence predictions are flagged for human annotation. In the final step, both machine- and human-annotations are used to update the previous model with OLTR and semi-supervised techniques. Once the model is updated, the classifier is fine-tuned using energy-based loss again for out-of-distribution sensitivity.