Zero-shot animal behavior classification with image-text foundation models

Gaspard Dussert*[1], Vincent Miele[1, 2], Colin Van Reeth[3], Anne Delestrade[3], Stéphane Dray[1] and Simon Chamaillé-Jammes[4]

[1]Université Lyon 1, CNRS, UMR 5558, Laboratoire de Biométrie et Biologie Evolutive (LBBE), Villeurbanne, France
[2]Laboratoire d'Ecologie Alpine (LECA), Université Grenoble Alpes, CNRS, Université Savoie Mont Blanc, France
[3]Centre de Recherches sur les Ecosystèmes d'Altitude (CREA Mont-Blanc), Observatoire du Mont-Blanc, Chamonix, France
[4]Centre d'Ecologie Fonctionnelle et Evolutive (CEFE), Université de Montpellier, CNRS, EPHE, IRD, Montpellier, France

Abstract

In this paper, we investigate whether some recent image-text foundation models are able to perform classification of animal behavior without any fine-tuning. We experiment zero-shot approaches with two types of models: image-text contrastive learning such as CLIP and multimodal LLMs such as CogVLM. Using a new large dataset of European fauna, we demonstrate that some of these models are already very good at predicting behavior, allowing the estimation of behavior-specific activity patterns almost identical to those derived by participatory science annotations.

1. Introduction

Camera traps are a powerful tool for studying wildlife in their natural habitats with minimal disturbance. As they are increasingly deployed throughout the world, a growing need for efficient methods to automatically analyze the vast amount of images they generate emerges. This is for instance true for one common goal of camera-trapping, which is to describe the diel activity patterns and time budgets of some target species, i.e. the distribution of their level of activity and behaviors throughout the day [18].

Many machine learning models are now able to reliably predict the species identity of animals seen in camera trap images [12, 17, 25], an information that can be used in several downstream tasks. For example Whytock *et al*. [24] used such species identification to derive activity patterns. However, one might want to obtain more informative, behavior-specific patterns, for instance to understand when are animals foraging, resting or moving. Also, this would avoid biases such as, for instance, spuriously creating a peak in 'activity' when an animal is resting (see Fig. 2).

Obtaining expert labels of behaviors on large sets of images is however very time-consuming and costly. Also, currently only a few models can predict behavior from camera trap images, and only for a few defined species and behaviors [11, 13, 19]. New models could be trained, but this remains difficult: firstly, a large amount of training data is needed, which is often hard to obtain in ecology; secondly, a lot of computational resources are required for training, which might not be available for ecologists and has a large carbon footprint; and thirdly, the model need to be re-trained each time a new species or behavior class is added.

At the same time, research groups are rapidly publishing "foundation models", capable of executing zero-shot tasks, i.e. making predictions on tasks for which they have not been explicitly trained. These pre-trained foundation models can solve most of the mentioned problems, since they don't require training data, inference requires much less energy than training and the billions of examples seen during training theoretically allow them to predict and generalize on various species and behavior of interest.

In this article, we investigate whether foundation models are already good enough to be used in ecology for behavioral classification without any finetuning. We use a new dataset of camera trap images of European fauna with be-

*gaspard.dussert@univ-lyon1.fr

Figure 1. Sample images from the dataset, of the three species and three behaviors of interest. When an event has multiple image, only the first one is shown. The Zooniverse label and proportion of vote is shown at the top in green, and the prediction of the best zero-shot model (CogVLM) is shown at the bottom in white.

havior labels provided by participatory sciences, and explore the zero-shot capabilities of five recent Vision Language Models (VLM) to predict the behavior of three different species. We hypothesize that the models have not been trained on many images of wild animals active at night and may be biased towards certain behaviors. We therefore also investigate the interest of specifying to the model whether the image was taken during nighttime or daytime.

2. Material and Methods

2.1. Dataset

The CREA Mont-Blanc camera trap dataset is collected by CREA Mont-Blanc (Centre de Recherches sur les Ecosystèmes d'Altitude) in the valley of Chamonix in the French Alps. The dataset is made of 131 537 images collected between 2017 and 2020 during 24 580 events, events being defined as 1-minute periods. Images were labelled on Zooniverse using participatory science. Participants were asked to identify the species as well as five behavior : eating, resting, moving, standing and interacting. Each event had at least 30 votes at the end of the process. We have focused our work on three species that are actively studied by the CREA Mont-Blanc: chamois, red deer and roe deer, and focused on 3 behaviors: eating, moving and resting.

2.2. Zero-shot vision-language models

Vision Language-Models (VLM) are multi-modal deep learning models that can process image and text simultaneously. Often trained on very large datasets (billions of samples), they exhibit very good zero-shot performances. Here we explored the zero-shot capacity of five VLMs: CLIP [15], SigLIP [26], WildCLIP [7], CogVLM [23] and MobileVLM V2 [4]. The first three are based on image-text contrastive learning, while the last two are multimodal Large-Language Model, they process image and text features at the input of a Large-Language Model (LLM) which allows better complex reasoning capacities. We selected CLIP ViT-Base for our baseline, SigLIP-SO for a state-of-the-art contrastive learning VLM, WildCLIP for a model specialized on wildlife and two multimodal LLM : CogVLM and MobileVLM V2 for their better complex reasoning capacities.

CLIP Contrastive Language Image Pre-training (CLIP) is trained to predict the similarity between a text and an image. Formally, a vision encoder (ViT [5]) and a text encoder (transformer [21]) are used to extract the embeddings of images and texts. The similarity of an (image, text) pair is obtained through the cosine similarity of the normalized embeddings.

SigLIP Sigmoid loss for Language-Image Pre-training (SigLIP) uses a sigmoid loss instead of the standard softmax loss, which improves the efficiency and performances of the trained models. In this paper, we use SigLIP-SO, a model with a Shape-Optimized 400M parameters ViT for the image encoder [2], which has better performance than larger models such as EVA-CLIP (5B parameters).

WildCLIP is fine-tuned CLIP on camera trap images. It has been trained on the large Snapshot Serengeti dataset [20], which has seven millions camera trap images captured in the Serengeti national park, Tanzania. The image captions were built from attributes labelled by participatory science, including the animal behavior. The model has been trained with the "Learning without Forgetting" VR-LwF loss [10], which reduces the catastrophic forgetting during the fine-tuning stage, which should helps the model to understand (image, text) pairs of European wildlife. Indeed, such images and captions were seen during WildCLIP's pretraining (i.e. CLIP training) but not during the fine-tuning phase, which focused exclusively on African wildlife.

CogVLM uses a very large vision encoder EVA2-CLIP-E (10B parameters) to extract image embeddings, these embedding are inserted into the input sequence of a LLM Vicuna1.5-7B (7B parameters) which allows the model to process both image and text information at the same time.

MobileVLM V2 is a lighter multimodal LLM, it uses CLIP ViT-L/14 for the vision encoder and Vicuna-7B for the LLM.

2.3. Zero-shot inference

For all methods, images are first cropped using MegaDetector v5 [3] to ensure that the image encoder of each model extracts information related to the animal and not to the rest of the scene.

Image-text contrastive learning based methods such as CLIP are able to perform zero-shot classification [15] using their ability to predict the similarity score between an image and a text. Here, we follow this methodology and match the images with the following captions: "a camera trap picture of an animal {action}". We obtain similarity scores for each action, which are then normalized using the softmax function.

For the multimodal LLMs, we perform a Visual Question Answering task [1]. We prompt the assistant with the following question: "Is the animal in the image eating, moving or resting? Answer the question using a single word.". For each predicted token, we can also retrieve the score of all possible words of the LLM vocabulary. Because the model predicts only one word, only the scores of the first token are considered to compute the confidence score of the predicted behavior. We follow the methodology described by Jiang *et al*. [9]: the score of the predicted behavior is normalized by the sum of the scores of the candidates words only.

We experiment to give additional context to the model to improve the performances: we inform the model whether the image was taken during daytime or nighttime. For CLIP-based methods, we change the caption to "a camera trap picture of an animal {action}, captured during {context}." with *context* being either "nighttime" or "daytime". For LLMs, we add "This image is captured during {context}." before the question.

The classification performance of the model are evaluated with the accuracy of each behavior and the calibration of the scores is measured with the ECE [8]. For these two metrics, we use a subset of the dataset that removes ambiguity in the labels (e.g. sequences with an animal that is seen both eating and moving). This subset is obtained by keeping only the images for which the most voted for action has at least 90% of the votes, and will be called the *cleaned* dataset hereafter. We also aggregate individual images predictions to event predictions to improve the accuracy of the models [6]. Confidence scores at the sequence level are calculated by averaging the scores of the images in the sequence.

2.4. Behavior-specific activity patterns

Activity patterns are computed either on the Zooniverse annotations or the models predictions using the time at which the camera trap events were triggered. Times are first converted to radian and standardized using two anchor time (the average sunrise time and the average sunset time) to account for the variation of the length of the day during the year [22]. A circular kernel density (of the von Mises distribution) is fitted to the radian times to obtain the activity pattern. The overlap index Δ is used to compare two activity patterns by computing the integral of the minimum between the two density function [16]. An index of 1 is obtained for a perfect overlap and 0 for no overlap at all.

Activity patterns are computed using the whole dataset and for our three species and three behaviors of interest. For Zooniverse annotations, we keep only the event for which a behavior has at least 50% of the votes. For models prediction, we used the same threshold of 0.75 on the confidence scores of all the models. We did not apply any filter to ensure independence between event (e.g. 30 minutes interval between two events) as it has been shown to bias the activity patterns [14].

3. Results

Starting with the contrastive learning methods, the baseline CLIP with ViT-Base architecture performs poorly for this behavior recognition task, with a low accuracy of 71.75% (Tab. 1). In particular, it fails to predict correctly the eating behavior. The SOTA model SigLIP is a significant improvement over the CLIP model, with a much better accuracy of 91.14%, although the accuracy of the eating behavior is still lower than the one for the other behaviors. The specialized WildCLIP model offers very similar performance to SigLIP despite a much smaller number of parameters. For the multimodal LLms, CogVLM gives the best performance, with an accuracy of 97.45%, while MobileVLM V2 is on the same level as SigLIP and WildCLIP.

Giving additional day-night context improves the accuracy of all the contrastive learning based methods (Tab. 1), but decreases the accuracy of the LLMs. None of the models exhibited a good score calibration, with the best model (WildCLIP) having an ECE of 6.85%. Moreover, adding context deteriorates the calibration of almost all the models.

The good zero-shot action recognition accuracy of the best models allows them to precisely estimate the behavior-specific activity patterns. Due to limited space, only the chamois activity patterns are shown on Figure 2. Activity patterns computed with CogVLM predictions are very close to those obtained with participatory science labels, and even small variations of the activity are captured. The activity patterns predicted by CLIP are generally unreliable, in particular the model is biased towards predicting "moving" at nighttime.

Quantitatively, measurements of the overlap index of activity patterns derived from Zooniverse annotations and those derived from model predictions are shown in Table 2. In accordance with previous results, CogVLM has the best overall overlap values, with an average overlap index of 90.04%. SigLIP, WildCLIP, and MobileVLM 2 exhibit slightly lower performance compared to CogVLM, while CLIP performs significantly worse. Once again, the additional day-night context helps with contrastive methods but not with LLMs. In particular, SigLIP with context emerges as the second-best model. The additional information re-

Model	Context (day & night)	Accuracy			Macro accuracy	ECE (%)	#Params
		Eating	Moving	Resting			
Contrastive learning							
CLIP [15]	×	31.76	92.39	91.08	71.75	17.15	150M
	✓	48.11	88.67	93.98	76.92	19.61	
SigLIP [26]	×	84.28	94.20	94.94	91.14	10.78	878M
	✓	94.65	89.20	95.18	93.01	14.15	
WildCLIP [7]	×	89.47	**94.70**	93.49	92.55	**6.85**	150M
	✓	90.09	94.17	95.42	93.23	7.80	
LLM							
CogVLM [23]	×	99.37	93.71	**99.28**	**97.45**	11.93	18B
	✓	**99.53**	87.95	99.04	95.50	11.29	
MobileVLM V2 [4]	×	95.75	82.23	**99.28**	92.44	9.55	7B
	✓	96.07	78.89	**99.28**	91.41	10.35	

Table 1. Action recognition accuracy and ECE of the five models on the cleaned dataset.

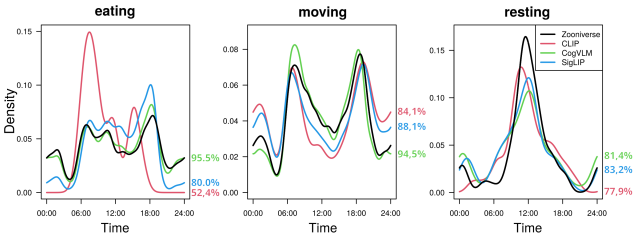

Figure 2. Comparison of Chamois activity pattern using either the participatory science labels from Zooniverse (black) or predictions of three models (colors). Overlap values are indicated on the right of the curves.

Model	Context	Average overlap index			
		Eating	Moving	Resting	All
Contrastive learning					
CLIP	×	36.60	66.97	88.53	64.03
	✓	39.73	75.80	91.27	68.93
SigLIP	×	75.27	86.27	90.47	84.00
	✓	89.97	85.13	**93.73**	89.61
WildCLIP	×	91.03	80.63	92.27	87.98
	✓	91.20	80.23	91.00	87.48
LLM					
CogVLM	×	91.67	87.17	91.30	**90.04**
	✓	93.00	**87.37**	85.27	88.54
MobileVLM V2	×	**93.43**	83.37	87.67	88.16
	✓	93.10	86.86	83.63	87.86

Table 2. Overlap index averaged over the three species of interest on the whole dataset.

duces the performance gap between day and night compared to SigLIP without context, which resulted in poor estimation of activity patterns (Fig. 2, blue line).

4. Discussion

This study evaluates the ability of zero-shot modern VLMs to predict the behavior of three mountain ungulates in camera trap images. We show that these models can have very good accuracies and produce activity patterns almost identical to those obtained using participatory science.

A visual analysis of the disagreement between the predictions of the models and the participatory science annotations reveals that there is a degree of ambiguity and that both may be right in some way. In fact, some sequences display two behaviors (e.g. Fig. 1, bottom right), the most common being eating and moving. While not common, these ambiguities may explain the slight deviation in the activity pattern of CogVLM, despite its near-perfect accuracy on the cleaned dataset. This suggests that exploring a zero-shot multilabel classification approach might be useful.

Given that several of the models achieve very good performance, it is interesting to study the trade-off between model size and performance. When considering multimodal LLMs, CogVLM appears to be the best model if inference time is not a concern. MobileVLM V2 is clearly not interesting, as its performance is similar to the one of WildCLIP, which is 50 times smaller. The difference in performance is highly likely due to the differing sizes of the vision encoders between CogVLM (10B parameters) and MobileVLM (428M parameters). For contrastive methods, WildCLIP seems to be the most efficient model. It performs considerably better than CLIP ViT-Base, despite having the same number of parameters. WildCLIP's performance is impressive given that it has been fine-tuned on African data only, which demonstrates the effectiveness of the VR-LwF "Learning without Forgetting" loss. However, the model is not performing better than SigLIP despite having already been exposed to camera trap images and similar captions depicting the same behaviors. Nonetheless, the release of our data will facilitate the development of specialized models more robust to the changes in species, habitats and behaviors. Finally, SigLIP seems to be the most efficient non-specialized foundation model and has a reasonable inference time (1 hour for 40 000 images on one NVIDIA V100 GPU).

Providing the context of whether the image was taken

during the day or at night only seems to improve the contrastive methods. This can be explained by the fact that giving a more precise caption reduces the distribution gap with the captions in the training dataset [15]. Current multimodal LLMs, on the other hand, are trained with rigid conversation templates and can be very sensitive to the structure of the prompt. Most of the datasets they were trained on do not include context before the question. Hence, in contrast to the contrastive methods, giving this additional context has increased the distribution gap and may have made confused the models.

These findings raise questions regarding the necessity of training new models for ecology, especially considering the rapidly accelerating rate of publication of foundational models. We can expect to see improved performance and new capabilities very soon, such as the ability to analyze multiple consecutive images, as the commercial GPT4-Vision model already does. Knowing how to use these models correctly will allow new information to be extracted from existing camera trap datasets.

Data availability statements

The code and dataset will be made available on Zenodo : https://doi.org/10.5281/zenodo.10925926

References

[1] Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Dhruv Batra, and Devi Parikh. VQA: Visual Question Answering, 2016. arXiv:1505.00468 [cs]. 3

[2] Ibrahim Alabdulmohsin, Xiaohua Zhai, Alexander Kolesnikov, and Lucas Beyer. Getting ViT in Shape: Scaling Laws for Compute-Optimal Model Design, 2024. arXiv:2305.13035 [cs]. 2

[3] Sara Beery, Dan Morris, and Siyu Yang. Efficient Pipeline for Camera Trap Image Review, 2019. arXiv:1907.06772 [cs]. 2

[4] Xiangxiang Chu, Limeng Qiao, Xinyu Zhang, Shuang Xu, Fei Wei, Yang Yang, Xiaofei Sun, Yiming Hu, Xinyang Lin, Bo Zhang, and Chunhua Shen. MobileVLM V2: Faster and Stronger Baseline for Vision Language Model, 2024. arXiv:2402.03766 [cs]. 2, 4

[5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, 2021. arXiv:2010.11929 [cs]. 2

[6] Gaspard Dussert, Simon Chamaillé-Jammes, Stéphane Dray, and Vincent Miele. Being confident in confidence scores: calibration in deep learning models for camera trap image sequences, 2024. Pages: 2023.11.10.566512 Section: New Results. 3

[7] Valentin Gabeff, Marc Rußwurm, Devis Tuia, and Alexander Mathis. WildCLIP: Scene and animal attribute retrieval from camera trap data with domain-adapted vision-language models, 2023. Pages: 2023.12.22.572990 Section: New Results. 2, 4

[8] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On Calibration of Modern Neural Networks, 2017. arXiv:1706.04599 [cs]. 3

[9] Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. How Can We Know When Language Models Know? On the Calibration of Language Models for Question Answering, 2021. arXiv:2012.00955 [cs]. 3

[10] Zhizhong Li and Derek Hoiem. Learning without Forgetting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(12):2935–2947, 2018. Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence. 2

[11] Dan Liu, Jin Hou, Shaoli Huang, Jing Liu, Yuxin He, Bochuan Zheng, Jifeng Ning, and Jindong Zhang. LoTE-Animal: A Long Time-span Dataset for Endangered Animal Behavior Understanding. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 20007–20018, Paris, France, 2023. IEEE. 1

[12] Veronika Mitterwallner, Anne Peters, Hendrik Edelhoff, Gregor Mathes, Hien Nguyen, Wibke Peters, Marco Heurich, and Manuel J. Steinbauer. Automated visitor and wildlife monitoring with camera traps and machine learning. *Remote Sensing in Ecology and Conservation*, n/a(n/a), 2023. 1

[13] Mohammad Sadegh Norouzzadeh, Anh Nguyen, Margaret Kosmala, Alexandra Swanson, Meredith S. Palmer, Craig Packer, and Jeff Clune. Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning. *Proceedings of the National Academy of Sciences*, 115(25), 2018. 1

[14] Christopher Peral, Marietjie Landman, and Graham I. H. Kerley. The inappropriate use of time-to-independence biases estimates of activity patterns of free-ranging mammals derived from camera traps. *Ecology and Evolution*, 12(10):e9408, 2022. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/ece3.9408. 3

[15] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision, 2021. arXiv:2103.00020 [cs]. 2, 3, 4, 5

[16] M. S. Ridout and M. Linkie. Estimating overlap of daily activity patterns from camera trap data. *Journal of Agricultural, Biological, and Environmental Statistics*, 14(3):322–337, 2009. 3

[17] Noa Rigoudy, Gaspard Dussert, Abdelbaki Benyoub, Aurélien Besnard, Carole Birck, Jérome Boyer, Yoann Bollet, Yoann Bunz, Gérard Caussimont, Elias Chetouane, Jules Chiffard Carriburu, Pierre Cornette, Anne Delestrade, Nina De Backer, Lucie Dispan, Maden Le Barh, Jeanne Duhayer, Jean-François Elder, Jean-Baptiste Fanjul, Jocelyn Fonderflick, Nicolas Froustey, Mathieu Garel, William

Gaudry, Agathe Gérard, Olivier Gimenez, Arzhela Hemery, Audrey Hemon, Jean-Michel Jullien, Daniel Knitter, Isabelle Malafosse, Mircea Marginean, Louise Ménard, Alice Ouvrier, Gwennaelle Pariset, Vincent Prunet, Julien Rabault, Malory Randon, Yann Raulet, Antoine Régnier, Romain Ribière, Jean-Claude Ricci, Sandrine Ruette, Yann Schneylin, Jérôme Sentilles, Nathalie Siefert, Bethany Smith, Guillaume Terpereau, Pierrick Touchet, Wilfried Thuiller, Antonio Uzal, Valentin Vautrain, Ruppert Vimal, Julian Weber, Bruno Spataro, Vincent Miele, and Simon Chamaillé-Jammes. 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, 2023. 1

[18] J. Marcus Rowcliffe, Roland Kays, Bart Kranstauber, Chris Carbone, and Patrick A. Jansen. Quantifying levels of animal activity using camera trap data. *Methods in Ecology and Evolution*, 5(11):1170–1179, 2014. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/2041-210X.12278. 1

[19] Frank Schindler and Volker Steinhage. Identification of animals and recognition of their actions in wildlife videos using deep learning techniques. *Ecological Informatics*, 61: 101215, 2021. 1

[20] Alexandra Swanson, Margaret Kosmala, Chris Lintott, Robert Simpson, Arfon Smith, and Craig Packer. Snapshot Serengeti, high-frequency annotated camera trap images of 40 mammalian species in an African savanna. *Scientific Data*, 2(1):150026, 2015. Publisher: Nature Publishing Group. 2

[21] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need, 2017. arXiv:1706.03762 [cs]. 2

[22] Carmen Vazquez, J. Marcus Rowcliffe, Kamiel Spoelstra, and Patrick A. Jansen. Comparing diel activity patterns of wildlife across latitudes and seasons: Time transformations using day length. *Methods in Ecology and Evolution*, 10(12):2057–2066, 2019. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/2041-210X.13290. 3

[23] Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, Jiazheng Xu, Bin Xu, Juanzi Li, Yuxiao Dong, Ming Ding, and Jie Tang. CogVLM: Visual Expert for Pretrained Language Models, 2024. arXiv:2311.03079 [cs]. 2, 4

[24] Robin C. Whytock, Jędrzej Świeżewski, Joeri A. Zwerts, Tadeusz Bara-Słupski, Aurélie Flore Koumba Pambo, Marek Rogala, Laila Bahaa-el din, Kelly Boekee, Stephanie Brittain, Anabelle W. Cardoso, Philipp Henschel, David Lehmann, Brice Momboua, Cisquet Kiebou Opepa, Christopher Orbell, Ross T. Pitman, Hugh S. Robinson, and Katharine A. Abernethy. Robust ecological analysis of camera trap data labelled by a machine learning model. *Methods in Ecology and Evolution*, 12(6):1080–1092, 2021. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/2041-210X.13576. 1

[25] Marco Willi, Ross T. Pitman, Anabelle W. Cardoso, Christina Locke, Alexandra Swanson, Amy Boyer, Marten Veldthuis, and Lucy Fortson. Identifying animal species in camera trap images using deep learning and citizen science. *Methods in Ecology and Evolution*, 10(1):80–91, 2019. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/2041-210X.13099. 1

[26] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid Loss for Language Image Pre-Training, 2023. arXiv:2303.15343 [cs]. 2, 4