Comparing and fine-tuning animal images classification models for polish species

Kacper Müller
AGH University of Science and Technology
Email: kacpermuller@student.agh.edu.pl

***Abstract*—Camera traps photos help biologists in wildlife conservation, but classification by-hand is a time-consuming task. There are several available Deep Learning models that automate this process. However, the results may not be satisfying for our case, as output species of one model may not be the same as species present in the area where our photos were taken. In order to get the best results, we can fine-tune the models. In this paper I fine-tuned and compared the models for polish species. As a result, many species are predicted with more than 90% accuracy. Thus, the models might be potentially used for partially automating the animals classification task.**

I. Introduction

The monitoring of wildlife helps determine the presence of endangered species, invasions of species in regions unusual to them, or monitoring animal diseases, especially the ones that can be transmitted to humans [1]. Camera traps have helped in the way that large amounts of data (photos) are now available. The problem is that the process of deriving information from the data is a time-consuming task because every picture needs to be classified as some species. Scientists have tried to find a solution to this problem in the field of Artificial Intelligence (AI) [2] [3] [4]. Although it must be noted that relying solely on images when trying to monitor biodiversity could pose some risks, especially when using images from citizen science platforms. The images could be faked, AI-generated, badly labeled or the location could be incorrect [5].

Couple of models have been created for animals images classification. The problem is the models have strict output labels. Using such model we would like it to output only the species that we are expecting in our environment, e.g. we do not want images taken in Poland to be classified as a coyote, instead of a wolf. Fine tuning a model could be the solution, as we can adjust it's output to match our needs. [6]

The standard classification pipeline can be seen in the figure 1. The detector detects animal on the image and crops the image to focus on the animal. The cropped photo is then classified by the classifier, predicting the animal's species [7].

My contribution is a comparison of the available models in order to find the best one to use for polish species, and a creation of fine tuned models that would give better results. The paper also describes how such a model could be created so that future researchers, maybe in different regions, could also tailor the models to their requirements. The code and app using the fine-tuned model is openly available [8].

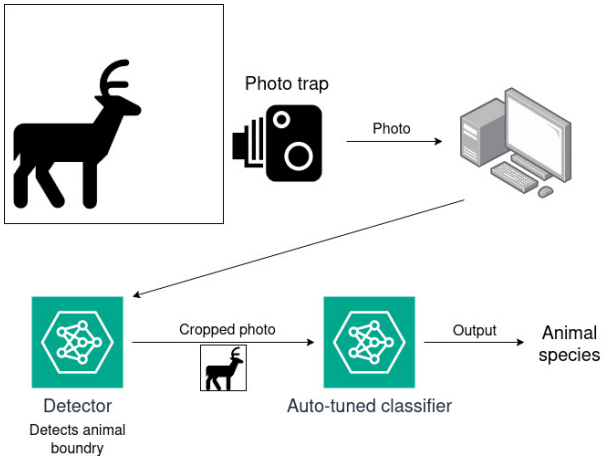

Fig. 1. Diagram of the system.

II. State of the art solutions

In animal detection models using the YOLO architecture are among the most widely used and effective. In particular, the implementation using the YOLOv5x architecture called Megadetector [9] shows promising results - one research suggests 94.6% accuracy [2]. In addition, the creators of the DeepFaune model created a detector based on the YOLOv8s architecture [3].

Ready to use models that have proven high accuracy in animal classification are DeepFaune (VIT), trained on european species [3], and Google's Speciesnet (CNN), trained on species from around the world [10] [4].

When it comes to fine-tuning the CNN and VIT models, there is a wide range of methods available. A method that is not complex and is giving promising results is head fine-tuning. It is described in section 3.5, as it is the chosen way of fine-tuning models in this article. It has been used successfully in studies where relatively small models were used to classify fish species [11] [12], as well as in a "Deepfaune new england: A trail camera species classification model for northeastern north america" (2024) article [6], very similar to this one.

Another method, called two stage classification, involves using one model, called the "global model" to classify species more generally as a family (for example, mink, marten, stoat and polecat images would be classified as "mustelidae") and then a different model, called "expert model", would distinguish the specific species ("mustelid" would be reclassified

as mink, marten, etc.). The drawback is that it requires more computational power, as it is needed to train and use multiple models, instead of one. [1]

There is also an emerging approach: Vision-Language Models (VLMs). There are several studies that demonstrate their strong accuracy and generalization performance [13] [**?**]. The main problem is that these models are big and require a lot of computational power. This makes them unavailable to many users who do not have powerful enough computers. This is why the article focuses on smaller models: CNN or VIT.

III. Methods

A. Data

Data for the verification of the models was provided by Izabela Fedyń. The dataset consisted of 33,301 labeled camera-trap photos collected at 900 locations between 2022 and 2023. The locations covered diverse forest ecosystems in Poland, including Beskid Niski, Bieszczady Mountains, Kotlina Sandomierska, Wyżyna Kielecka and the Białowieża Forest. The captured animals (number of images in brackets): red deer (12501), roe deer (7608), human (2606), fox (2144), bird (1461), wildboar (1348), beaver (1271), bison (855), marten (558), undefined (536), empty (502), wolf (392), squirrel (317), moose (277), badger (151), bear (140), vehicle (94), fallow deer (92), rodent (92), hare (81), wildcat (70), raccoon dog (54), dog (49), otter (28), lynx (22), cat (17), mink (14), polecat (9), other (8), stoat (3), weasel (1).

The red deer and roe deer were thinned: 30% of roe deers and 20% of red deers images were randomly selected to be classified. Around 15 images were corrupted.

After looking at images that were classified by models as empty, but were not labeled as empty, a lot of them seemed incorrectly labeled. Pictures of a certain animal were taken in series, often in couple of seconds intervals, so the labeler could have "interpolated" the label to several images, even if the animal was not necessarily in the picture. An example of such an image can be seen in figure 2. Images taken a couple of seconds earlier and a couple of seconds later had a deer in them, so the image in between was also labeled as one. As a result, such images were not taken into account when calculating the statistics of classification.

Fig. 2. Example of an image not labeled as "empty".

Training data was taken from the Chernobyl Exclusion Zone camera trap photographs data [14], the iNaturalist platform [15], and the WildCapture dataset [16].

Species from the Chernobyl data: badger, red deer. Other species in this data set were not of sufficient quality.

Species from the WildCapture dataset: badger, dog, fallow deer, fox, hare, marten, polecat, red deer, roe deer, wild boar, wildcat, wolf.

Species from the iNaturalist data: badger (*Meles meles*, 989), bear (*Ursus arctos*, 974), beaver (*Castor fiber*, *Myocastor coypus*, 1015), bird (all species occuring in Poland, 999), bison (*Bos bonasus*, 996), cat (*Felis catus*, 1000), dog (*Canis familiaris*, 688), fallow deer (*Dama dama*, 671), fox (*Vulpes vulpes*, 1403), hare (*Lepus europaeus, Lepus timidus, Oryctolagus cuniculus*, 687), lynx (*Lynx lynx*, 142), marten (*Martes foina, Martes martes*, 1301), mink (*Mustela lutreola* and *Neogale vison* also included, since only 19 european minks photos were available, 579), moose (*Alces alces*, 550), otter (*Lutra lutra*, 244), polecat (*Mustela putorius*, 793), raccoon (*Procyon lotor*, 748), raccoon dog (*Nyctereutes*, 942), red deer (*Cervus elaphus*, 757), roe deer (*Capreolus capreolus*, 728), squirrel (*Sciurus vulgaris*, 1497), stoat (*Mustela erminea*, 866), weasel (*Mustela nivalis*, 1442), wild boar (*Sus scrofa*, 858), wildcat (*Felis silvestris*, 548), wolf (*Canis lupus*, 605).

After downloading a certain picture, the detector model was run on it. If the model detected the animal and the cropped picture was not smaller than 115 x 115 pixels, it was saved [8]. Some of them contained very blurry images of animals, some were empty (the detector is not perfect [2]), so every image was inspected and the invalid ones were deleted.

The photos from the iNaturalist platform were selected as follows. The latin names of species were searched. If there were many observations, the location was adjusted to Poland and, if necessary, it's nearest neighbors. In each case, it was aimed to retrieve around 2000-3000 images, because many images did not pass the checks: either they were pictures of animals' tracks, bones, fur, or the cropped image was too small. It was especially noticeable in the case of beavers, as most of the images depicted trees cut by them.

B. Detector

Images from Chernobyl's and WildCapture datasets were cropped, but iNaturalists' images had to be cropped. In order to do that, the Megadetector [9] was used. The Deepfaune detector was faster, but it struggled more to detect animals and works on smaller inputs, so the output images were also smaller.

The validation images were also not cropped, so a Megadetector was run on them as well.

C. Models

The models used in the research were DeepFaune [3] and Speciesnet [17]. The former uses a VIT architecture ("vit_large_patch14_dinov2.lvd142m" backbone), whereas the latter is a CNN with "EfficientNetV2-M" architecture.

The main problem of comparing the models is that they use different labels. DeepFaune uses labels similar to the ones available in the verification dataset, since it aims at European animals. Speciesnet, on the other hand, has over 2000 species and was much harder to unify. In order to be able to compare them, outputs had to be mapped, standardized. It is the simplest way of fine-tuning the model.

D. Mapping results

Mapping results involves just renaming the outputs of a certain model to the desired ones. When using a model in Europe, it could classify some animal as ”American bison” species. We know that in the location where an image was taken such species does not occur, and so we can map it to the similar looking species like ”European bison”.

In the results of Deepfaune model ”golden jackal” was mapped to the ”wolf” category, as well as ”nutria” and ”marmot” to the ”beaver” one. ”reindeer” was mapped to ”red deer” and other species, not included in the provided verification data, were mapped to ”other”.

The Speciesnet model was much more problematic, since it is trained to cover over 2000 species. All mappings can be found in the code [8]. Examples of mappings: all the bird species were mapped to ’bird’, all hare species were mapped to ’hare’, ’grey fox’, ’hoary fox’, ’puma’ and ’red fox’ were mapped to ’fox’. A lot of single species were mapped to the ’other’ category, unusual examples might include ’blue monkey’, ’golden snub-nosed monkey’, or ’western gray kangaroo’.

E. Changing the head

Using this method, we freeze the whole model, except the classification head - the last linear layer, or layers, which map outputs from convolution network.

The idea is that the big vision models, trained on millions of photos, are already really good at detecting certain features. In the animal case, neurons at the end of CNN might indicate that an ear of certain shape or distinct fur is present in the image. Then the classification head, maps that information onto the output species label. That means we don’t have to retrain whole model, which takes a lot of time, computational power, and electrical energy, but just the classification head, which might be a small linear neural network.

The method was successfully used in several animal images classification researches [11] [12] [6].

IV. Results and Discussion

A. Models results

Accuracies for all models and species can be seen in figure 3. The confusion matrices can be found in figures at the bottom of the article. Many species have accuracy of over 90% and in most cases it increased with the fine tuning of the classifier head.

The original DeepFaune model does not classify different mustelidae species, thus marten, mink, polecat and stoat are zeros. Although SpeciesNet does differentiate them, it got only martens right. But in the fine tuned models we can see some improvement. It is a hard task though, because mustelidae species are similar to each other. They might differ in some details, like a white spot on the muzzle, which might not even be visible in the photos. For example, we can see in figure 6 that minks are classified as martens and polecats.

Models also struggle with beavers. Fine tuning the DeepFaune model improved this by 10%, but it might still be improved by better data. Fine tuning the SpeciesNet model significantly improved bison and raccoon dog classification. The original model probably classified bisons as a species that was not mapped to the ”bison” label, but ”other”, maybe ”cattle”, or something similar.

A big loss in accuracy can be seen in the bear species. It was perhaps due to the fact that the available bear images were really low resolution. It could probably be increased when retrained using higher quality images. However, a great increase can be observed in wildcat detections. Although it was probably at the price of cat detections. This too is an example of really similar looking species.

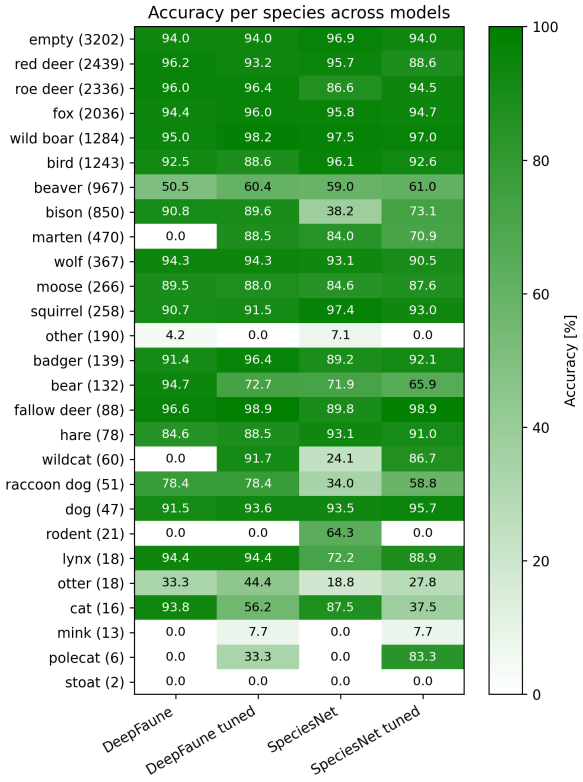

Fig. 3. Accuracies of models per species.

B. Limitations

The main limitation in the research was the data. For some species, such as red deer or fox, a lot of good quality images can be found. Unfortunately, for many species, like bears, lynx, polecat quality images are hard to find.

Another limitation is that the studied models are big. It takes a lot of time to train and assess them, so comprehensive

studying, e.g. number of epochs to learn, learning rate, whether to use squarely cropped or stretched images, is difficult on a consumer GPU. The methods of fine tuning are also limited by that, because retraining whole model would be almost impossible. Finally, the SOTA model architecture VLM could not be tested, because of that as well.

V. Conclusion

The examined models can successfully classify majority of animals photos. Unfortunately, they cannot fully automate the annotation task yet, as some species are still hard to distinguish, especially the similar-looking ones. However, this research was limited by data and computational power, so it is highly possible that better results can be achieved with the same method, just with a higher quality of data.

Which model of the four would give the best results depends on what we are trying to achieve. If mustelidae and felinae species are not important to distinguish, the mapped DeepFaune model will probably be the best. However, if we do want to distinguish similar looking species, we could try the classifier head fine-tuning method.

There are also approaches that were not examined in this research, but could give promising results. They are described in the section 2 - state of the art solutions.

A review of current state-of-the-art solutions for animal photos classification should be created. All the reviews I found were outdated. The camera trap image classification field has developed greatly in the past two years, and new approaches are constantly emerging.

As I am not very experienced yet, I encourage researchers to verify all of the results and point out any mistakes.

Acknowledgment

I would like to thank Andres Vejar Perez and Krzysztof Rusek for suggesting ways of fine tuning the models, as well as Izabela Fedyń for the data and help with everything species related, and Agnieszka Kleszcz for model suggestions.

References

[1] M. Mulero-Pázmány, S. Hurtado, C. Barba-González, M. L. Antequera-Gómez, F. Díaz-Ruiz, R. Real, I. Navas-Delgado, and J. F. Aldana-Montes, "Addressing significant challenges for animal detection in camera trap images: a novel deep learning-based approach," *Scientific Reports*, vol. 15, no. 1, p. 16191, May 2025. [Online]. Available: https://doi.org/10.1038/s41598-025-90249-z

[2] S. Leorna and T. Brinkman, "Human vs. machine: Detecting wildlife in camera trap images," *Ecological Informatics*, vol. 72, p. 101876, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1574954122003260

[3] N. Rigoudy, G. Dussert, A. Benyoub, A. Besnard, C. Birck, J. Boyer, Y. Bollet, Y. Bunz, G. Caussimont, E. Chetouane, J. C. Carriburu, P. Cornette, A. Delestrade, N. De Backer, L. Dispan, M. Le Barh, J. Duhayer, J.-F. Elder, J.-B. Fanjul, J. Fonderflick, N. Froustey, M. Garel, W. Gaudry, A. Gérard, O. Gimenez, A. Hemery, A. Hemon, J.-M. Jullien, D. Knitter, I. Malafosse, M. Marginean, L. Ménard, A. Ouvrier, G. Pariset, V. Prunet, J. Rabault, M. Randon, Y. Raulet, A. Régnier, R. Ribière, J.-C. Ricci, S. Ruette, Y. Schneylin, J. Sentilles, N. Siefert, B. Smith, G. Terpereau, P. Touchet, W. Thuiller, A. Uzal, V. Vautrain, R. Vimal, J. Weber, B. Spataro, V. Miele, and S. Chamaillé-Jammes, "The deepfaune initiative: a collaborative effort towards the automatic identification of european fauna in camera trap images," *European Journal of Wildlife Research*, vol. 69, no. 6, p. 113, Oct 2023. [Online]. Available: https://doi.org/10.1007/s10344-023-01742-7

[4] A. Hehmeyer, "Using the power of ai to identify and track species," *World Wildlife Fund*, 2025. [Online]. Available: https://www.worldwildlife.org/news/stories/using-the-power-of-ai-to-identify-and-track-species/

[5] D. S. Campos, R. F. D. Oliveira, L. D. O. Vieira, P. H. N. D. Bragança, J. L. S. Nunes, E. C. Guimarães, and F. P. Ottoni, "Revisiting the debate: documenting biodiversity in the age of digital and artificially generated images," *Web Ecology*, vol. 23, no. 2, pp. 135–144, 2023. [Online]. Available: https://we.copernicus.org/articles/23/135/2023/

[6] J. T. K. G. A. F. Clarfeld, L. and T. Donovan, "Deepfaune new england: A trail camera species classification model for northeastern north america." 2024. [Online]. Available: https://code.usgs.gov/vtcfwru/deepfaune-new-england

[7] S. Beery, D. Morris, and S. Yang, "Efficient Pipeline for Camera Trap Image Review." [Online]. Available: http://github.com/agentmorris/MegaDetector

[8] K. Müller, "Code for "comparing and fine-tuning animals images classification models for polish species"," 2025, python code, Jupyter notebooks. [Online]. Available: https://github.com/PLKplkPLK/mgr

[9] S. Beery, "The megadetector: Large-scale deployment of computer vision for conservation and biodiversity monitoring," *AI for Social Impact*. [Online]. Available: https://ai4sibook.org/wp-content/uploads/2022/08/MegaDetector.pdf

[10] H. K. D. M. S. B. T. B. J. A. Tomer Gadot, Stefan Istrate, "To crop or not to crop: Comparing whole-image and cropped classification on a large dataset of camera trap images," *IET Computer Vision*, vol. 18, 2024. [Online]. Available: https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/cvi2.12318

[11] R. Jongjaraunsuk, W. Taparhudee, S. Sirisuay, M. Kaewnern, V. Dulyapurk, and S. Janekitkarn, "Transfer learning model application for rastrelliger brachysoma and r. kanagurta image classification using smartphone-captured images," *Fishes*, vol. 9, no. 3, 2024. [Online]. Available: https://www.mdpi.com/2410-3888/9/3/103

[12] J. Jareño, G. Bárcena-González, J. Castro-Gutiérrez, R. Cabrera-Castro, and P. L. Galindo, "Automatic labeling of fish species using deep learning across different classification strategies," *Frontiers in Computer Science*, vol. Volume 6 - 2024, 2024. [Online]. Available: https://www.frontiersin.org/journals/computer-science/articles/10.3389/fcomp.2024.1326452

[13] H. Markoff and J. Galaktionovs, "Hierarchical re-classification: Combining animal classification models with vision transformers," 2025. [Online]. Available: https://arxiv.org/abs/2510.14594

[14] C. Barnett, S. Gashchak, N. Beresford, and M. Wood, "Wildlife camera trap photographs from the chornobyl exclusion zone, ukraine (november 2020 - march 2021) following extensive wildfires," 2022. [Online]. Available: https://doi.org/10.5285/a657ffc3-8f62-458f-bcb7-30e116807174

[15] iNaturalist community, "Observations of [species list] from poland and near neighbors observed between 2022-2025," 2025. [Online]. Available: https://www.inaturalist.org

[16] L. Cultrera, L. Seidenari, and A. Del Bimbo, "Leveraging visual attention for out-of-distribution detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 4447–4456.

[17] T. Gadot, Istrate, H. Kim, D. Morris, S. Beery, T. Birch, and J. Ahumada, "To crop or not to crop: Comparing whole-image and cropped classification on a large dataset of camera trap images," Nov. 2024.

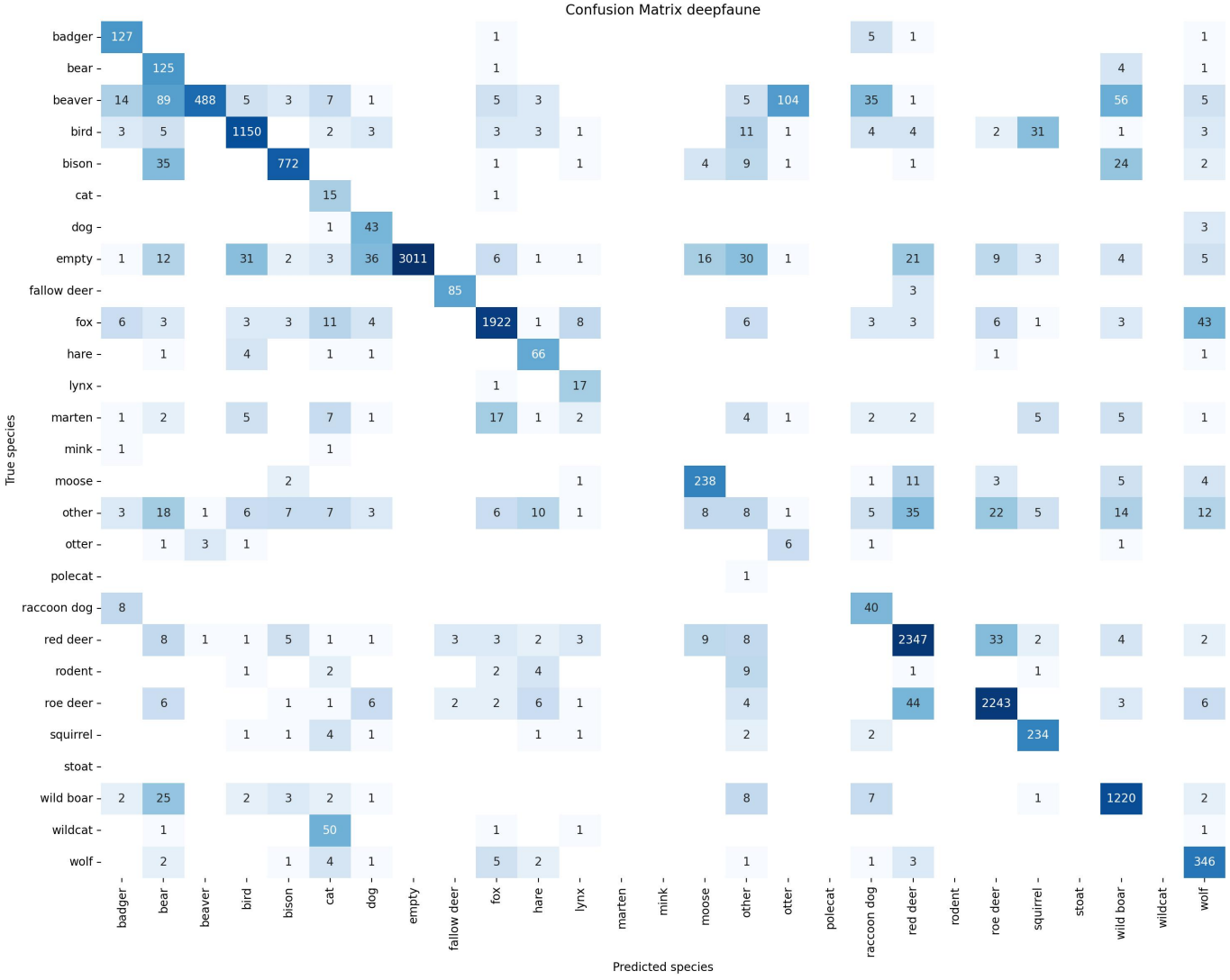

Fig. 4. Deepfaune results confusion matrix.

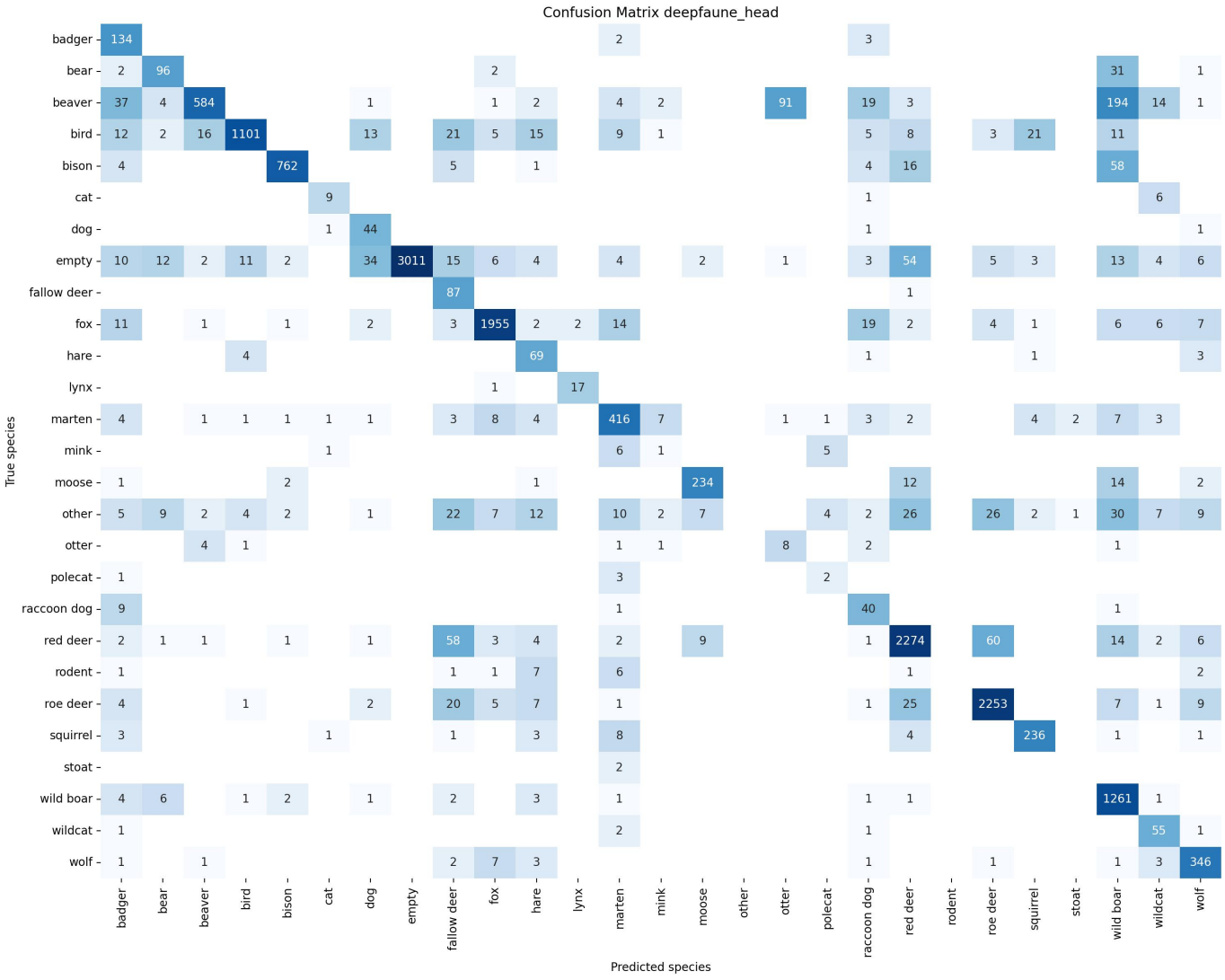

Fig. 5. Deepfaune fine-tuned results confusion matrix.

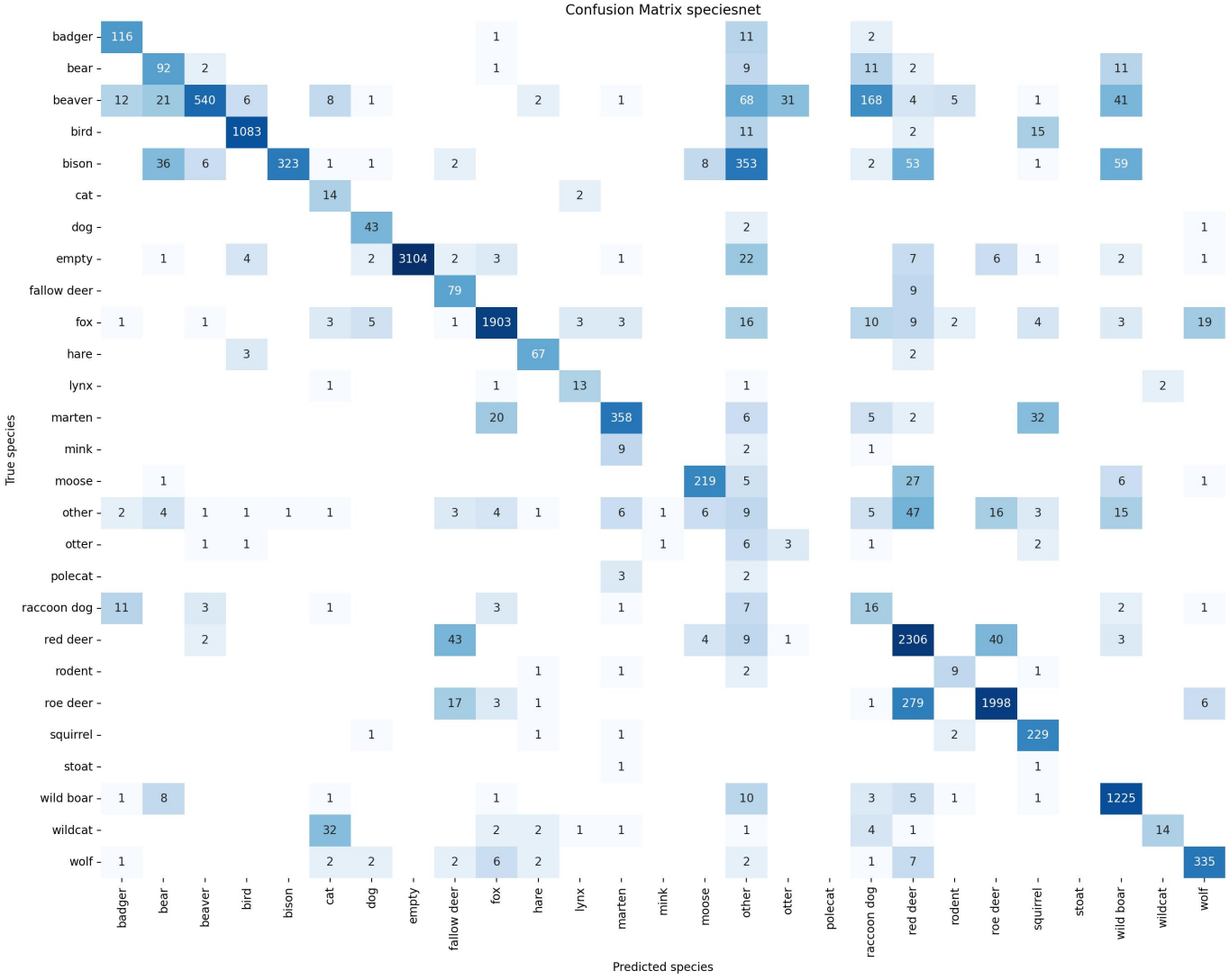

Fig. 6. Speciesnet results confusion matrix.

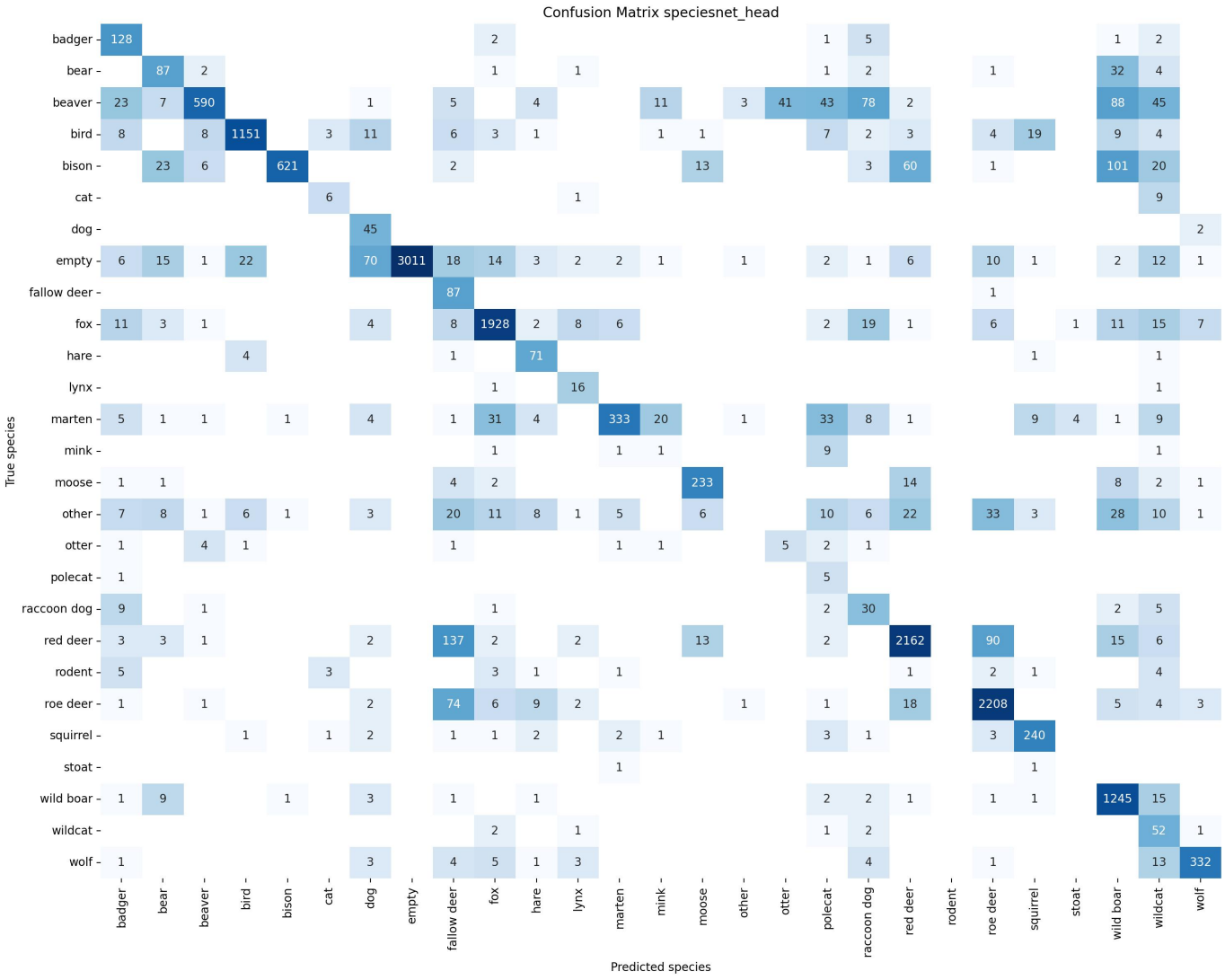

Fig. 7. Speciesnet fine-tuned results confusion matrix.