Received: 15 December 2023 | Revised: 18 May 2024 | Accepted: 12 June 2024

IET Computer Vision

DOI: 10.1049/cvi2.12294

ORIGINAL RESEARCH

Recognition of European mammals and birds in camera trap images using deep neural networks

Daniel Schneider[1] | Kim Lindner[2] | Markus Vogelbacher[1] | Hicham Bellafkir[1] | Nina Farwig[2] | Bernd Freisleben[1]

[1]Department of Mathematics & Computer Science, University of Marburg, Marburg, Germany

[2]Department of Biology, University of Marburg, Marburg, Germany

Correspondence

Daniel Schneider.
Email: schneider@informatik.uni-marburg.de

Funding information

Deutsche Forschungsgemeinschaft, Grant/Award Number: 210487104; Hessisches Ministerium für Wissenschaft und Kunst

Abstract

Most machine learning methods for animal recognition in camera trap images are limited to mammal identification and group birds into a single class. Machine learning methods for visually discriminating birds, in turn, cannot discriminate between mammals and are not designed for camera trap images. The authors present deep neural network models to recognise both mammals and bird species in camera trap images. They train neural network models for species classification as well as for predicting the animal taxonomy, that is, genus, family, order, group, and class names. Different neural network architectures, including ResNet, EfficientNetV2, Vision Transformer, Swin Transformer, and ConvNeXt, are compared for these tasks. Furthermore, the authors investigate approaches to overcome various challenges associated with camera trap image analysis. The authors' best species classification models achieve a mean average precision (mAP) of 97.91% on a validation data set and mAPs of 90.39% and 82.77% on test data sets recorded in forests in Germany and Poland, respectively. Their best taxonomic classification models reach a validation mAP of 97.18% and mAPs of 94.23% and 79.92% on the two test data sets, respectively.

KEYWORDS

computer vision, convolutional neural nets, image classification, image recognition, neural nets

1 | INTRODUCTION

Due to the ongoing worldwide decline of biodiversity over the last centuries [1, 2], there is an urgent need to comprehensively monitor ecosystems and initiate conservation measures if necessary. For this purpose, automatic recorders can be deployed directly in the field to autonomously collect large amounts of data over long time spans and at large spatial scales with little to no human interference [3].

Camera traps are commonly used to monitor mammal populations in a non-intrusive way. First introduced in 1956, they have significantly contributed to the field of wildlife ecology in recent decades [4, 5]. Camera traps are heat- or motion-activated cameras placed in the wild to automatically record images and/or videos of animals. To allow continuous data recordings, cameras are usually equipped with an infrared lens for night-time images and a customary lens for day-time images. After deployment in the target habitat, the cameras record data for several weeks without disturbing the animals by the presence of humans. The manual analysis of the autonomously recorded huge amounts of data requires expertise and is therefore time-consuming and expensive. Automated approaches to identify different animal species are desirable while also ensuring that the results are less biased by observers [6]. In recent years, machine learning methods, and particularly deep neural network models, have been used to analyse large amounts of data in the field of ecology—apparently for the first time in 2014 [7].

Field data obtained through visual recordings have largely been used to monitor mammals, while automated bird population surveys mainly use microphones that continually record audio [8]. Bird species occurring in the recording area can then

be identified using automated bird call recognition methods [9]. However, not all bird species frequently vocalise, and vocal activity patterns change over the year [10]. Monitoring approaches solely focusing on audio recordings will therefore likely fail to sufficiently cover migrating or overwintering bird species. The use of image recognition has since been extended to identify aerial images of species assemblies, often collected through unmanned aerial vehicles [11]. While the use of aerial imagery is a powerful tool for monitoring species communities in more open habitats, it is less suitable for forest-dwelling species surrounded by heterogeneous vegetation. Here, the extension of camera trapping into the field of avian monitoring is a promising approach.

However, most animal species recognition models are limited to mammal identification only and thus group all birds into a single class. Deep neural network models for visually identifying bird species, in turn, are usually not designed for processing camera trap images but high quality bird photographs and cannot discriminate between mammals. In this article, we present deep neural network models to recognise the highly desirable combination of both mammal and bird species in camera trap images. This allows us to identify non-vocal bird species, which is not possible with methods based solely on audio recordings.

Existing deep neural network models (e.g. [12–14]) are usually limited to specific regions and small sets of species from the respective regions for which sufficiently large numbers of training images are available. For Africa and North America, several large publicly available data sets exist, but for Europe mainly smaller data sets with only a few species are provided.

These data sets are mostly labelled with the exact species names. However, since it is often difficult to accurately identify species under poor visibility conditions, many data sets also contain images that are only labelled at a higher taxonomic level (e.g. genus or order) in which the animals are clearly separable. Furthermore, some data sets do not differentiate between similar-looking subspecies or label species using English names, which in some cases does not allow a clear assignment to the scientific species names. Such data is therefore only of limited use for training an animal species classifier. To be able to nevertheless use the available data, we have developed taxonomic classification models that predict not only the species but also the taxonomy of the corresponding animal, that is, genus, family, order, group, and class names. We include data that is not labelled up to the species level for training and set the missing taxonomic labels to a predefined ignore label. Predicting the taxonomic hierarchy can assist researchers in analysing challenging images where accurate species classification is difficult by allowing automatic determination of the higher taxonomy levels. The taxonomic predictions can also contribute to a better understanding of how the neural networks are making their predictions and where errors occur. They can also aid in detecting cases where the model was uncertain during automatic species identification by looking for inconsistent taxonomic predictions. In these cases, either an attempt can then be made to correct them automatically or they can be forwarded to a human annotator for review.

We have limited the training of our deep neural networks to temperate forest species, but our animal selection includes a number of species that are not yet included in any available animal recognition model. In particular, our neural networks recognise 25 mammal and 63 bird species known to occur in Central European forests with a focus on two field study sites in the Marburg Open Forest (MOF) in Hesse, Germany and the Białowieża National Park (BNP) in Podlaskie Voivodeship, Poland. These include some species that are very difficult to distinguish visually, such as various marten species and closely related bird species. Since there is no data set available that covers all species relevant to our study, and collecting and annotating larger amounts of images of these species ourselves would be very time-consuming, we gathered and combined data from several publicly available data sets to train our models.

In this work, we apply and compare different neural network architectures to our mammal and bird recognition task, including ResNet [15], EfficientNetV2 [16], Vision Transformer (ViT) [17], Swin Transformer [18], and ConvNeXt [19]. We also investigate measures to overcome various challenges associated with camera trap image analysis, such as limited amounts of available data, overfitting to locations, low image quality, incorrectly labelled data, and species imbalance.

Our best deep neural network for species classification achieves a mean average precision (mAP) of 97.91% on a validation data set left out from our training data and mAPs of 90.39% and 82.77% on our test data sets recorded in forests in Germany and Poland. Our best taxonomic classification model reaches an overall validation mAP of 97.18% and mAPs of 94.23% and 79.92%, on the two test data sets, respectively.

To summarise, our main contributions are as follows:

- We present a deep learning model for recognising European mammals and, for the first time, birds in camera trap images at the same time.
- To the best of our knowledge, we are the first to develop taxonomic classification models for mammals and birds that predict not only the animal species but the taxonomic hierarchy, which can be advantageous for images that are difficult to analyse.
- We compare several state-of-the-art neural network architectures such as ResNet [15], EfficientNetV2 [16], ViT [17], Swin Transformer [18], and ConvNeXt [19] operating on the task of camera trap image analysis.
- We make our best trained models publicly available at https://github.com/umr-ds/Mammal-Bird-Camera-Trap-Recognition to enable other researchers to build on our work and apply our models to their own data or develop them further.
- We publish our MOF and BNP test data sets consisting of roughly 2500 and 15,000 labelled camera trap images, respectively, in the same repository.

The rest of this article is organised as follows. Section 2 discusses related work. Section 3 presents our recognition workflow, data sets, neural network models, challenges, and evaluation metrics. Section 4 presents comparisons of our models and experimental results. Section 5 discusses the proposed approach. Section 6 concludes this article and outlines areas for future work.

2 | RELATED WORK

2.1 | Deep neural networks

In (supervised) deep learning, neural network models are trained to recognise desired patterns using large amounts of labelled data [20]. In the area of image processing, convolutional neural networks (CNNs) [21] have achieved great successes. They learn filter weights during training that react to certain features in the input. Prominent examples are AlexNet (the first major breakthrough of CNNs) [22] and ResNet (introduction of skip connections, which improves the training of deeper networks) [15, 23]. A highly optimised type of CNN is EfficientNet [24]. It is based on neural architecture search to investigate how different ways of scaling a baseline CNN architecture affects prediction quality and resource requirements. In a follow-up paper, the authors presented updated EfficientNetV2 model configurations that further improve the trade-off between performance and resource requirements [16].

The Transformer model architecture [25] does not use any convolutions but instead solely relies on attention mechanisms that help the network to focus on the most relevant parts of the input. This architecture was initially proposed in the field of natural language processing. Dosovitskiy et al. [17] applied transformers to image analysis with the introduction of the ViT. Since then, several works have emerged in this field which further optimise the underlying principle of the ViT and overcome some of its limitations, for instance, the Swin Transformer [18]. Liu et al. [19] developed a CNN without attention mechanisms, called ConvNeXt, by adapting the block structure of a ResNet architecture to that of Swin Transformers and adopting other minor design adjustments from newer model architectures. Through these improvements, they were able to outperform previous CNN and transformer architectures and achieve state-of-the-art results.

In our work, we apply the ResNet-152 [15], EfficientNetV2-M [16], ViT-Base [17], Swin-Base [18], and ConvNeXt-Base [19] model architectures to the task of recognising mammals and birds in camera trap images and compare their performance.

2.2 | Hierarchical image classification

The topic of hierarchical classification in image processing has been addressed in several works in the literature. Srivastava et al. [26] extended a CNN with a prior that learns a tree-like class hierarchy and applies it to the class weights. This helped the authors to achieve improved performance on rare classes. Yan et al. [27] introduced Hierarchical Deep CNNs, which consist of a component for the coarse classification of more easily discriminable classes and multiple components for the fine classification of the more challenging classes. Wehrmann et al. [28] used a neural network with several local output heads for different hierarchy levels and a global total output. The loss is calculated for each of these outputs and summed up to a total loss. Koo et al. [29] expanded this idea by presenting a combination of a CNN and a recurrent neural network (RNN) in which a hierarchical image representation is learnt using the CNN and the features of different convolutional layers are passed to the RNN, which predicts the path in a tree-like hierarchy. Practical applications of hierarchical models for image processing can be found, for example, in Turkoglu et al. [30], who performed hierarchical recognition of grain varieties on satellite images, and Elhamod et al. [31], who present a framework for recognising species and genus of fish. Cramer et al. [32] also applied taxonomy hierarchies to the training of CNN models for the acoustic classification of birds and thereby achieved state-of-the-art results.

Recently, Wang et al. [33] presented a CNN-based approach for acoustic bird identification trained with hierarchical labels. They used an Xception model trunk to first identify high-level features. Subsequently, they used the rest of the model and a hierarchical attention mechanism to process this feature information at different levels and make a hierarchical prediction. Inconsistent predictions are then corrected using a path correction strategy.

To train our taxonomic classification networks, we do not use a special neural network architecture designed for hierarchical labels. Instead, we use a single-output backbone model that we split up into multiple output heads at the end, one for each hierarchy level, which are all trained simultaneously. This approach is basically similar to the idea of Wang et al., but we neither split up the backbone network nor use attention layers to keep the network architecture as simple as possible.

2.3 | Animal recognition

2.3.1 | Animal classification

Automated animal classification apparently dates back to 2013, where Yu et al. [34] performed sparse coding spatial pyramid matching to extract relevant features from previously manually cropped images. Using these features, they trained a linear support vector machine classifier on a (by today's standards) small data set of 7000 camera trap images with 18 animal species. Chen et al. [7] used deep neural networks for animal species classification in 2014. They performed automatic image segmentation using a graph-cut algorithm to separate the areas showing animals from the background. For classification, they used a small CNN trained on 14,000 images containing 20 species.

In the following years, artificial neural networks and especially CNNs became the state of the art for most image processing tasks. In most cases, models are pre-trained on large data sets like ImageNet [35] with millions of images and then fine-tuned on smaller camera trap data sets, a process called transfer learning. With the publication of the Snapshot Serengeti data set in 2015, which consists of 3.2 million images of 48 animal species [36], a data basis for training more complex animal recognition models was created.

In 2017, Gomez et al. compared the animal classification performance of CNN architectures of different sizes on a 26-class subset of the Snapshot Serengeti data set. Deeper models like ResNet-101 performed better than smaller ones, reaching a maximum accuracy of 88.9% on a class-balanced subset consisting only of images with animals in the foreground [37].

In 2018, Norouzzadeh et al. [12] trained 9 network architectures to not only classify animal species but simultaneously count animals and determine other attributes such as behaviour, a process known as multitask learning. They used two networks for this purpose: the first one detected images containing animals, and the second one subsequently analysed these images. For training, the authors used the Snapshot Serengeti data set with all 48 animal species. They obtained the best results for the animal species classification with a ResNet-152 that achieved an accuracy of 93.8% on the test data.

In 2019, Tabak et al. [13] introduced the North American Camera Trap Images (NACTI) data set that consists of over 3 million images of 27 species taken in North America and Canada. The authors trained a ResNet-18 on this data set and achieved an accuracy of 97.6%. They also performed out-of-sample validation by applying their trained model to images from locations that were not present in the training set. Here, the model achieved an accuracy of 81.8%.

In the same year, Willi et al. [38] analysed how transfer learning can be used to train CNNs for animal species classification on much smaller data sets as they are realistic in practice. The authors compared the performance of ResNet-18 models that had been pre-trained on Snapshot Serengeti and then transferred to one of four significantly smaller data sets to those that were trained only on the small data sets. They found that in almost all cases, the models trained using transfer learning performed better, and in particular for small data sets, transfer learning led to a very strong improvement.

The interest in analysing camera trap images has continued to grow since then, supported by competitions such as the iWildCam Challenge [39], an annual contest since 2018 that focuses on model generalisation to new environments. In addition, companies launched initiatives to strengthen research in deep learning for ecology, for example, AI for Earth by Microsoft or Wildlife Insights by Google.

One recent trend is the attempt to develop preferably small models that can also run on less powerful (embedded) edge devices and thus carry out animal species recognition directly in the field. For example, Islam et al. [40] trained a CNN model to recognise small reptiles like frogs, snakes, and lizards found in Texas and deployed their model on a NVIDIA Jetson Nano edge device that is connected to the cameras. Zualkernan et al. [41] presented a system that analyses camera trap images directly on edge devices and transmits the classification results to a mobile app. The authors compared several network architectures geared towards mobile applications and tested their deployment on various edge devices. Jia et al. [42] performed neural architecture search on camera trap image data sets to find a lightweight CNN architecture for animal classification that performs on par to other networks but can run on edge devices.

The automated analysis of camera trap images from a fixed scope works best when the models are trained on images that are as similar as possible to the images to recognise later, ideally recorded at the same exact location. Auer et al. [43] proposed an active learning system to specialise their models on single camera traps with a small number of training images annotated by experts. The authors performed their experiments on a non-public data set from the Bavarian Forest National Park.

Recently, transformer architectures were used for species classification. Kyathanahally et al. [44] employed Data-efficient image Transformers (DeiTs) for ecological monitoring. The authors compared ensembles of CNNs and DeiTs on several animal data sets and showed that the latter achieved similar performance with significantly less hyperparameter optimisation efforts. Zheng et al. [45] presented a transformer-based model that uses cross-attention blocks for the re-identification of land animal individuals on camera trap images. Their model achieved state-of-the-art results on three public animal re-identification data sets. Agilandeeswari et al. [46] used a contrastive self-supervised learning approach based on the Swin transformer architecture for the detection and classification of animals. The authors used contrastive learning to harness the large amount of unlabelled images available for training. Their models achieved state-of-the-art results on several image recognition benchmark data sets, including 97.2% accuracy on Snapshot Serengeti.

In 2023, Fabian et al. [47] presented the WildMatch system that uses Multimodal Large Language Models for the zero-shot detection of wildlife species on camera trap images. The authors trained their models on a combination of human-annotated and automatically generated image descriptions (using descriptions of the species' appearance from Wikipedia). For the predictions, they also used a taxonomy hierarchy as an external knowledge source with which the descriptions were matched. The good classification performance they achieved comes at the cost of high computational requirements.

What all these approaches have in common is that they are limited to a comparatively small number of species that usually stem from a specific geographical region. In our models, we limit ourselves to Central European temperate forests, a region that has not been the focus of many larger studies before.

2.3.2 | Animal detection

The animal recognition approaches presented so far perform animal species classification at the image level without determining where the animals are located in the image. In some cases, however, it is desirable to carry out a localisation, for example, to count the number of animals or to track the position of the animals over a sequence of images. In 2018, Schneider et al. [48] trained object detection models to localise animals in camera trap images. For this purpose, they labelled a subset of the Snapshot Serengeti data set with bounding boxes delimiting the positions of the animals.

In 2020, Carl et al. [49] applied a Faster-RCNN object detection model, trained on a data set of 600 everyday classes, to the detection of 10 European mammal species. The model detected correct bounding boxes in 94% of the cases and often managed to predict a correct higher taxonomic rank as the classification. This shows that general-purpose models can suffice for camera trap image analysis in some cases. Shepley et al. [50] trained models for pig detection on camera trap images in the same year, using camera trap data sets as well as images from FlickR. They investigated how well models trained with data from one region could be applied to other regions.

Microsoft developed and made publicly available an animal detection model called MegaDetector [51] as part of its AI for Earth programme. The model was trained on a large number of camera trap images—some of which are not publicly available—and recognises objects of the classes animal, person or vehicle, but does not further distinguish between individual animal species. The latest version 5, released in 2022, uses the YOLOv5[1] object detection architecture.

In 2021, Norouzzadeh et al. [52] presented a system in which animal detection and animal species classification run separately. Using the MegaDetector, they localised animals in the images and cropped the images to the relevant area while sorting out empty images. They trained a simple classification model using active learning, a process in which human experts are presented with a selection of images for labelling that promises the greatest benefit for further training. The authors showed that this can greatly reduce the amount of human labelling work without sacrificing significant classification quality.

Choiński et al. [14] applied the lightweight YOLOv5 object detection architecture to the analysis of camera trap images. They used a small data set of 12 animal species for training. According to the authors, the use of a lightweight architecture opens up the possibility of deploying a trained model on edge devices directly at the cameras to perform real-time analysis of the images. Recently, Simões et al. [53] applied object detection models to videos recorded by camera traps. They extended the MegaDetector from pure detection to the classification of 17 animal species and developed methods to count the detected individuals.

In our work, we follow the idea of Norouzzadeh et al. of using the MegaDetector to detect the animals in the images and then identify them using our own classification model.

2.3.3 | Bird species recognition

The majority of work on the automated monitoring of bird populations is based on microphones and performs automatic recognition of bird species based on audio data. This method has proven to be successful in practice because microphones require little power and are therefore easy to deploy. To train deep learning models that analyse bird calls, large databases like Xeno-Canto,[2] where users from all over the world upload their recordings, can be used.

A well-known approach for automated bird species recognition is BirdNET [54]. It is based on a CNN model trained on a large audio data set using extensive data preprocessing, augmentation, and mixup to achieve state-of-the-art performance. Mühling et al. [55] proposed a task-specific neural network created by neural architecture search, which won the BirdCLEF 2020 challenge [56]. It operates on raw audio data and contains multiple auxiliary heads and recurrent layers. In 2022, Höchst et al. [57] published a system called Bird@Edge, where the analysis of bird calls is performed on edge devices directly in a forest and only the results are transmitted to a server, supporting real-time monitoring of a particular area.

There are far fewer publications on the visual recognition of birds than on auditory recognition. Many of these approaches are limited to the mere recognition and counting of birds from a greater distance or only make a rough genus determination [58–60]. The approaches that identify species are usually limited to a few bird species from a restricted geographic region. For example, Huang et al. [61] developed a mobile app in which images of birds can be identified using a CNN. However, their recognition model is limited to 27 bird species native to Taiwan. Raj et al. [62] used a CNN to recognise 60 species found in the Asian sub-continent. Ferreira et al. [63] trained a CNN model to recognise bird individuals from which they had previously collected training images using an automated method. Recently, Chalmers et al. [64] developed a bird species recognition model specifically for camera trap images. They trained a Faster-RCNN model for the real-time classification of 10 bird species. Their training data was obtained from iNaturalist.

To the best of our knowledge, there are no models for the visual recognition of bird species occurring in European forests, and only few models exist for the analysis of mammals in this region. Neural network models that can recognise both mammals and bird species in camera traps do not exist at all in the literature. Models for predicting the taxonomy of animals in camera traps have not been studied in previous works. Finally, there are no studies that compare more recent CNN

[1] https://github.com/ultralytics/yolov5.

[2] https://xeno-canto.org.

models and transformer architectures with regard to the task of animal classification.

3 | MATERIALS & METHODS

3.1 | Recognition workflow

Our camera trap image analysis workflow performs two steps: animal detection and animal classification, as shown in Figure 1.

First, we perform object detection using Microsoft's MegaDetector [51] to find the areas in the images showing animals. This also allows us to sort out the images where no animals are visible, that is, where the camera was falsely triggered. The MegaDetector model was trained on a very large number of camera trap images from various sources and therefore has a very good detection rate for a wide range of animal species. However, its high sensitivity for animal detection even in difficult environments also repeatedly leads to false detections of, for example, tree trunks or rocks. Such false detections should later be classified as 'empty' by our classification model and sorted out in this way. We use MegaDetector v5a, released in 2022.[3] It employs a YOLOv5 model that was fine-tuned on a large number of public and non-public camera trap data sets and can identify objects of the classes 'animal', 'human', and 'vehicle'. For each input image, MegaDetector returns a list of detected objects with bounding box coordinates, detected class, and detection confidence score.

We then perform species classification on all image areas where MegaDetector predicted the class 'animal'. We do this by cropping each image to the area of the predicted bounding box and resizing it to the input size of the classification model of 224 × 224 pixels, which corresponds to the input size of the pre-trained models used. We compare different neural network architectures for animal classification, including ResNet [15], EfficientNetV2 [16], ViT [17], Swin Transformer [18], and ConvNeXt [19].

Our models are trained to recognise 25 mammal and 63 bird species that we consider in our studies. This includes several, sometimes closely related species that are hard to distinguish in camera images, especially under difficult viewing conditions, such as House cat (*Felis catus*) and Wild cat (*Felis silvestris*), European hare (*Lepus europaeus*) and European rabbit (*Oryctolagus cuniculus*), Roe deer (*Capreolus capreolus*) and Red deer (*Cervus elaphus*), European pine marten (*Martes martes*) and Beech marten (*Martes foina*), Willow warbler (*Phylloscopus trochilus*) and Common chiffchaff (*Phylloscopus collybita*) and thrushes such as Song thrush (*Turdus philomelos*) and Mistle thrush (*Turdus viscivorus*). Our models also include raptors such as the Common buzzard (*Buteo buteo*) as well as bird species that in Germany are found only during migration, such as the Redwing (*Turdus iliacus*), which are easily missed in acoustic monitoring approaches. A full list of all species is provided in Appendix B. We also added an 'empty' class, which brings the total number of species labels to 89, which is much higher than in most other works on automated camera trap image analysis.

3.2 | Taxonomic prediction

In addition to the animal species classification models, we also trained models to make predictions at multiple taxonomic

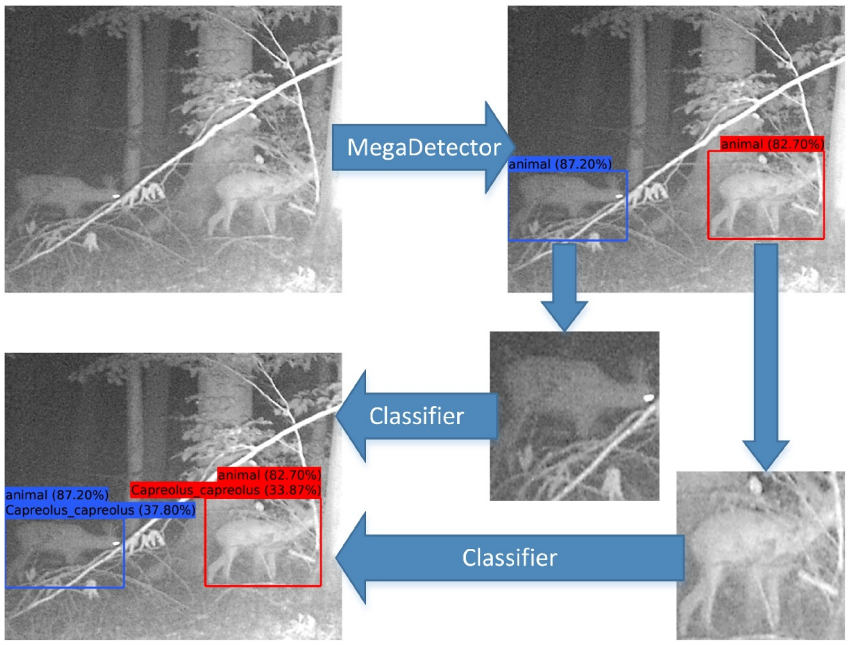

FIGURE 1 Our two-step animal recognition workflow. First, we detect image regions that contain animals. For these, we then carry out animal species classification.

[3] https://github.com/agentmorris/MegaDetector.

levels. These models have six output heads for taxonomic class (Mammalia, Aves or Empty), group (differentiation by animal size or habitat, e.g. small mammal, large mammal, waterbird, woodpecker, or raptor), order, family, genus, and species. These models can assist researchers in analysing more challenging images where accurate species classification is not possible. In cases where the model cannot make a clear species prediction, meaning that all labels have low confidence values or several labels reach similarly high values, the predictions of the higher taxonomy levels can be used as a fallback. To answer ecological questions, this is often better than relying on an uncertain, possibly incorrect species prediction.

The output heads of the taxonomic models are trained together using images labelled with the respective taxonomic hierarchy. A loss value is calculated independently for each taxonomic level and added up to a weighted total loss. To obtain taxonomic hierarchy labels for training and to resolve ambiguous annotations in the data sets, we used an API[4] provided by the Global Biodiversity Information Facility. For each label of a data set, we performed a name lookup query that searches for the best matching taxonomic hierarchy for the specified label. We extracted the fields required for our considered taxonomic levels from the returned taxonomy and saved them in a mapping dictionary. Using this dictionary, we mapped the image labels of a data set to the most precise label in the taxonomy. Finally, we assigned these label designations to the bounding box results of the MegaDetector and grouped the results in individual metadata files for each label.

Training models with taxonomic hierarchy labels allows us to use images for training that are not fully annotated at the species level. This includes data that is labelled at a higher taxonomic level (e.g. genus or order) because the animals could not be distinguished from each other more precisely, but also images from data sets that do not distinguish between similar-looking subspecies such as domestic cats or wildcats, or where the species are annotated with English names, which in some cases does not allow a clear assignment to the scientific species names. When training the models for taxonomic classification, we load the remaining taxonomic hierarchy for each image based on the saved most precise label so that a ground truth value can be assigned to the model output of each taxonomic level. Starting from the saved label taxonomy level, we reconstruct all higher taxonomic levels. For taxonomic levels that lie below the saved label, we set an ignore label (−1), which excludes the label from the loss calculation for the corresponding taxonomic level.

For each input image, the trained taxonomic models provide a probability distribution over all possible labels for each hierarchy level in which each label is assigned a confidence value that the image contains the corresponding animal type. We derive the animal names from the indices of the highest confidence value for each output head during prediction. In a post-processing step, we then check whether the predicted taxonomy is coherent. If this is not the case, we either compute the taxonomy with the highest overall confidence that is coherent, which automatically corrects minor errors in the predictions, or we mark these images as uncertain to be presented to a human expert for a second assessment.

[4] https://techdocs.gbif.org/en/openapi/.

3.3 | Data sets

Our neural network models should not be limited to one study site but generalise well across different locations. Furthermore, we focus on a selection of European mammal and bird species that do not occur in this composition in any data set available online. Therefore, we combined data from a variety of data sets to train our models. There are some larger freely available data sets with labelled camera trap images. However, most of them were recorded in Africa or North America and accordingly show the corresponding native species. For European species, the available data is much more limited. Some of the species we aim to recognise also occur in North America, so we can draw on data sets recorded there. We used images from the data sets Caltech Camera Traps [65], ENA24-detection [66], Idaho Camera Traps,[5] Missouri Camera Traps [67] and NACTI [13]. Additionally, we used images from the WCS Camera Traps data set, a collection of images from 12 countries created by the Wildlife Conservation Society.[6]

We also used two data sets recorded in Germany, which best matched our mammal species selection. One was taken from the Long-Term Population Trends of Disease-Transmitting Rodents research project (hereafter referred to as Rodent) [68], the other one is the Tierschnappschuss data set.[7]

To add more bird images to our training data, we used two bird-only data sets, namely a bird classification competition data set from Kaggle[8] and the NABirds data set containing bird species from North America.[9] None of these data sets is perfectly suited for training a model to recognise birds on camera trap images because they mostly contain bird photographs showing the birds in great visibility, and both cover only very few of the bird species we intended to recognise.

We supplemented the images from the various data sets with crawled images to fill out the remaining gaps and to increase the image diversity. We downloaded images of our desired species from the eMammal camera trap data management system,[10] crawled photographs of live animals with verified captions from iNaturalist[11] and finally used Google image search to collect more images of 7 underrepresented mammal species.

Some of the data sets came with manually annotated bounding boxes, but in most cases, the data was only labelled at the image or image sequence level. In this case, we applied the MegaDetector to obtain bounding boxes for the images and sorted out empty images. A few data sets contain images with

[5] https://lila.science/datasets/idaho-camera-traps.
[6] https://lila.science/datasets/wcscameratraps.
[7] https://emammal.si.edu/tierschnappschuss.
[8] https://www.kaggle.com/datasets/gpiosenka/100-bird-species.
[9] https://dl.allaboutbirds.org/nabirds.
[10] https://emammal.si.edu.
[11] https://inaturalist.org.

no animals, which were labelled as empty. We used the bounding boxes detected on these images as the training data for our 'empty' class. We divided the images of each data set into training and validation images. For this purpose, we grouped the images of each data set by label and used the first 15% of the images of each label as validation data (Val) and the rest as training data. We repeated this splitting process 3 times with different random seeds to ensure the quality of our trained models does not depend on the training/validation data split. A total of 1,273,379 species classification bounding boxes were split into 1,226,158 training and 47,221 validation boxes. For images labelled only on higher taxonomic levels, we applied the same splitting method at each of our considered taxonomic levels, which gave us a total of 6,377,885 taxonomic classification bounding boxes split into 6,283,826 training and 94,059 validation boxes.

For evaluation, we used two data sets recorded in the target areas of our studies. We deployed camera traps in the MOF[12] in Hesse, Germany in the first half of 2021, and we used camera trap images recorded in BNP in Podlaskie Voivodeship, Poland in spring to early summer 2013. We again applied the MegaDetecor to the recorded images to locate the animals. Subsequently, biologists in our team manually classified the animals visible in the bounding boxes so we could compare the predictions of our models with these labels. We publish download links to both annotated test data sets in our github repository. Table 1 gives an overview of all data sets we used in this work. See Appendix A for a more detailed overview of the number of bounding boxes per label and the split between training and validation data.

3.4 | Neural network models

We trained and compared neural network models of different architectural backbones, which we embedded in a model consisting of the following layers: The input layer contains $224 \times 224 \times 3$ neurons, which corresponds to the size of the RGB images that are fed into the network. The input is first processed by an augmentation layer, which generates one augmented version of each image by applying up to 10 consecutive augmentation operations from a predefined selection to the input image. The augmented images are then fed into the backbone model (ResNet-152, EfficientNetV2-M, ViT-Base, Swin-Base or ConvNeXt-Base), which is initialised with weights pre-trained on the ImageNet data set. We use a global average pooling layer to aggregate the feature maps output by the backbone model to one feature vector and add a dropout layer with a dropout rate of 0.6 as a regularisation method to reduce overfitting. This layer is followed by the output classification layers for the individual taxonomic levels (class, group, order, family, genus, and species). In the case of species prediction only, there is just one output layer. The classification layers use the softmax activation function to

TABLE 1 Overview of all data sets used in our experiments.

		Species	Images	Species classification		Taxonomic classification	
Data set name	Recording locations	Total	Total	Labels	Boxes	Labels	Boxes
Caltech Camera Traps [65]	Southwestern USA	21	~240,000	3	10,915	16	62,126
ENA24-detection [66]	Eastern North America	23	~10,000	5	2232	17	9623
Idaho Camera Traps	USA	62	~1.5 M	2	75,699	31	347,783
Missouri Camera Traps [67]	USA	21	~25,000	8	11,517	16	33,528
North American Camera Traps [13]	USA	28	~3.7 M	9	584,436	32	4,810,962
WCS Camera Traps	Worldwide (12 countries)	675	~1.4 M	7	10,820	62	361,240
Rodent [68]	Germany	41	~14,000	25	14,327	39	18,025
Tierschnappschuss	Southern Germany	41	~170,000	18	140,635	29	163,648
Kaggle Birds	Worldwide (Internet searches)	525	~90,000	6	1104	47	96,703
North American Birds	North America	400[a]	~48,000	2	429	43	49,840
eMammal	Worldwide (Crawled subset)			16	13,485	22	15,125
InatCrawl	Worldwide (Crawled subset)			88	406,366	90	407,868
WebCrawl	Worldwide (Crawled subset)			7	1414	7	1414
Marburg Open Forest (MOF)	Germany			19	2420	27	2729
Białowieża National Park (BNP)	Poland			20	4808	36	16,829

[a]The data set contains 555 categories, but some of them are of the same species.

[12]https://www.uni-marburg.de/de/fb19/fachbereich/infrastruktur/mof.

output a probability distribution over the confidence values of all possible labels for the respective taxonomic level. Figure 2 illustrates the model structure.

We trained our species classification models for 100 epochs and our taxonomic classification models for 120 epochs, each using the AdamW optimiser with a weight decay of 0.05 and the categorical cross-entropy loss function. For the taxonomic classification models, the total loss is calculated from the weighted sum of the losses of each output head. We either set the loss weights to fixed values or apply a *loss weight shifting strategy* that updates the weights during training as shown in Table 2. At the beginning, the highest taxonomic levels are weighted significantly higher. Every 15 epochs, the highest weighting is shifted down one taxonomic level until the species have the highest weight. This strategy first minimises the error for the higher taxonomic levels and then later for the lower levels. In this way, the model first learns to make a rough distinction between classes or families and then becomes more accurate in distinguishing the individual species, mimicking the way a human would approach animal recognition. Using this strategy led to a performance increase of the taxonomic models.

During our experiments, we optimised various hyperparameters to obtain better models. We investigated different learning rates and types of learning rate adjustment during training. For most models, *plateau reduction* proved to be a good choice. We set the learning rate to a fixed starting value (usually 5×10^{-4}) and reduced it by a factor of 0.2 if the validation loss did not decrease for five epochs. Alternatively, we used a *cosine decay* learning rate schedule that first increases the learning rate linearly from 0 to the starting value (5×10^{-4}) over 15 warm-up epochs and then gradually reduces it to 10^{-10} over the remaining training epochs following a cosine distribution. We experimentally determined which learning rates deliver the best results for which backbone architecture. Considering the variety of models and possible configurations, we made incremental changes until we found a parameter configuration that yielded satisfactory results.

As a further hyperparameter, we examined various input sizes. The pre-trained transformer-based models are fixed to square input sizes with 224 or 384 neurons; so in order to achieve good comparability, we examined these sizes as well as the in-between value of 300 for the ConvNext models. The models with an input size of 224×224 neurons performed best, which is why we carry out all comparisons with models of this input size.

Table 3 provides an overview of the model architectures we have trained and the hyperparameters selected for them.

3.5 | Challenges

Training a deep neural network model for recognising animals in camera trap images presents several challenges [6]. We

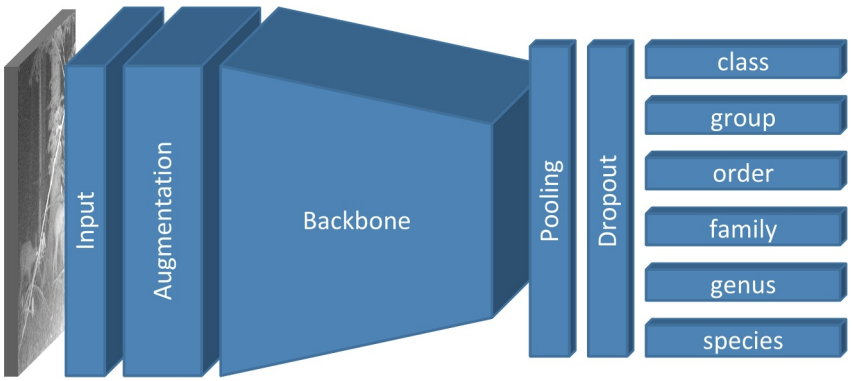

FIGURE 2 The structure of our animal classification models. We use several different architectures as backbones and allow up to six output layers for the considered taxonomic levels.

Strategy	Epochs	Class	Group	Order	Family	Genus	Species
Fixed weights	1–120	0.2	0.1	0.1	0.1	0.1	0.4
Weight shifting	1–15	0.9	0.1	0.001	0.001	0.001	0.001
	16–30	0.2	0.7	0.1	0.001	0.001	0.001
	31–45	0.1	0.2	0.6	0.1	0.001	0.001
	46–60	0.05	0.1	0.2	0.6	0.1	0.001
	61–75	0.01	0.05	0.1	0.2	0.6	0.1
	76–90	0.01	0.01	0.05	0.1	0.2	0.7
	91–120	0.01	0.01	0.01	0.05	0.1	0.9

TABLE 2 The loss weight strategies used during the training of the taxonomic models. We either set the loss weights to fixed values or adjust the weights every 15 epochs during training to first minimise the error for the higher taxonomic levels and later for the lower levels.

TABLE 3 Overview of the classification model architectures used in our experiments and their respective hyperparameters.

Model architecture	Number of parameters	Input size	Learning rate
ResNet-152 [15]	58.3 M	224 × 224	Plateau reduction from 5e-4
EfficientNetV2-M [16]	53.1 M	224 × 224	Plateau reduction from 5e-4
ViT-Base [17]	85.7 M	224 × 224	Plateau reduction from 8e-4
Swin-Base [18]	87.1 M	224 × 224	Cosine decay from 5e-4
ConvNeXt-Base [19]	87.6 M	224 × 224	Plateau reduction from 5e-4
ConvNeXt-Base (taxonomic classification)	87.8 M	224 × 224	Plateau reduction from 5e-4

investigated measures to overcome these challenges as best as possible.

3.5.1 | Amount of data

A large amount of labelled data is required to ensure that our models can reliably recognise the animal species we are looking for despite different appearances (e.g. different fur patterns, juvenile animals, day and night images). If the amount of data is too small, the model is prone to overfitting on the training data, meaning that the network adapts to the training data set and does not generalise well to other data. We addressed this problem by merging images from publicly available camera trap data sets and images crawled from animal databases into one large data set (see Section 3.3). We also used extensive data augmentation to create variations in the images during training, increasing the diversity of the training data set. The operations performed are random contrast, brightness and saturation changes, rotation, horizontal flipping, zooming, shearing, cropping, masking, and adding Gaussian noise.

3.5.2 | Adaptation to locations

Due to the static placement of camera traps, the cameras always show the same field of view. If a network is trained on many images from a limited number of locations, this can negatively affect the ability of the models to generalise to other locations. The neural networks might adapt not only to the animal species but also to similar image backgrounds and, as a result, perform significantly worse on images from previously unseen camera locations. One way to reduce this issue is to crop the training images to the areas where the animals are seen. For this reason, we used the MegaDetector to determine bounding boxes for all data sets considered. When training the classification model, we used these bounding boxes to crop the training images to the areas that show the animals. Since our model uses quadratic input images, we expanded the bounding boxes to square boxes with a side length equal to the larger side of the original rectangular boxes. In this way, the aspect ratio is preserved when cropping the images. However, since parts of the background are still visible despite cropping, it is important to use images from as many different locations as possible for training to prevent location adaptation.

3.5.3 | Data quality

Images from camera traps might differ significantly in quality. In many cases, the animals are not well captured in the image but appear cropped, obscured behind objects, or blurred because they were photographed in motion. Depending on the time of day and lighting conditions, the animals are also visible in varying quality. In particular, the infrared night shots are sometimes strongly overexposed or underexposed. This distinguishes camera traps from animal photographs, such as those found on iNaturalist, in which the animals are captured in good quality. However, training exclusively with such photos does not prepare the model well for the more challenging camera trap images. We addressed this problem by using a mix of camera trap data sets and higher quality photos. We also discarded images where the MegaDetector returned bounding boxes of very small size or with a confidence value below 0.6 to avoid very poor training examples.

3.5.4 | Incompletely labelled data

Many publicly available data sets are not fully labelled at the species level, for example, some images are only labelled with order or family, or the animals are labelled with their English common names that cannot be clearly assigned to a Latin species name, for example, rabbit (might refer to *L. europaeus* or *O. cuniculus*) or cat (might refer to *F. catus* or *F. silvestris*). To address these inaccuracies, we first tried label smoothing, but since this did not improve the performance, we decided against using this technique for the final models. To make insufficiently labelled images useable for training, we used the taxonomic hierarchy described in Section 3.2. This allows us to strongly increase the amount of available training data and make it more diverse.

3.5.5 | Species imbalance

Some animal species are significantly more abundant in the available data sets than others. While our data set contains over 300,000 samples for Red deer (*C. elaphus*), there are only 24 examples for Great snipe (*Gallinago media*). This imbalance may cause the network to focus on the more frequently occurring animal species and neglect the rare species. However, these rare species are usually of greater interest to ecologists, and hence, it is important to identify them correctly. To achieve this, we used a data generator that randomly samples an equal number of images for each species in each training epoch to ensure that the few images of the species with less available data are repeated more often. When using taxonomic hierarchy labels, we ensured that the number of images per label is somewhat balanced both at the highest level (taxonomic class) and at the lowest level (species). We did this by drawing samples across taxonomic levels. For example, if we select samples for the class mammal, these are chosen from all images that stem from Mammal in the taxonomic hierarchy. By using the taxonomic hierarchy labels, we can also use images of related species not found at our study sites for training, which can help the models learn more general representations that also support the recognition of the desired species.

3.6 | Evaluation

We conducted several experiments to evaluate the quality of our approach. We restricted ourselves to evaluating our trained classification models and use the MegaDetector as published by Microsoft. In our evaluation, we used a validation data set (Val) consisting of images held out from our training data, and our MOF and BNP data sets, which were not used for training. We trained three instances of each model using different training/validation splits of our data sets and calculated the following metrics for the quantitative evaluation of our models. In our results, we always provide the mean and standard deviation of the metrics for the related model instances.

To determine the overall quality of our models, we used the Accuracy metric, defined as follows:

$$Acc_k = \frac{|correct_k|}{|imgs|}, \tag{1}$$

where $imgs = \{i_1, i_2, \ldots i_N\}$ is a list of analysed images and $correkt_k \subseteq imgs$ are the images where the correct label is found in the top k predictions. We consider not only the Top 1 accuracy but also the Top 3 as well as Top 5 accuracy metrics, which indicate whether the correct classification is among the predicted classes with the highest 3 or 5 confidence values, respectively. One problem with these metrics is that they become biased for data sets with unbalanced label counts. In the case of a strong imbalance, as it occurs in our data sets, the correctness of the rarely occurring classes is barely reflected at all.

To assess more precisely how well our models perform for each class $c \in C$, where C is the list of all classes to be recognised, we additionally considered the class-wise accuracy and the F1 score for each class. The F1 score is defined as

$$\begin{aligned} F1_c &= \frac{2 \times Precision_c \times Recall_c}{Precision_c + Recall_c} \text{ with} \\ Precision_c &= \frac{TP_c}{TP_c + FP_c} \text{ and} \\ Recall_c &= \frac{TP_c}{TP_c + FN_c}, \end{aligned} \tag{2}$$

where TP_c, FP_c and FN_c are the amounts of true positive, false positive and false negative predictions for the class c, respectively. To evaluate the quality of the overall model, we calculated the mean class-wise F1 score (mF1) over all classes from those values.

$$mF1 = \frac{1}{|C|} \sum_{c=1}^{C} F1_c \tag{3}$$

We also used mAP as an additional quality measure. The mAP score is the most commonly used quality measure for retrieval results and approximates the area under the recall-precision curve. The task of animal species recognition can be considered as a retrieval problem for each species where the annotated images represent the relevant documents. We calculate the AP for each class $c \in C$, where C is the list of all classes to be recognised, as follows:

$$\begin{aligned} AP_c &= \frac{1}{|relevant_c|} \sum_{k=1}^{|imgs|} prec@k * rel@k \\ \text{with} \quad rec@k &= \frac{|relevant_c \cap retrieved_k|}{|retrieved_k|} \\ \text{and} \quad rel@k &= \begin{cases} 1 & \text{if } i_k \in relevant_c \\ 0 & \text{otherwise} \end{cases}, \end{aligned} \tag{4}$$

where $imgs = \{i_1, i_2, \ldots i_N\}$ is a list of analysed images ranked by the prediction score for the class c. $relevant_c \subseteq imgs$ denotes the relevant images for the class c, that is, the images containing an animal of class c and $retrieved_k = \{i_1, i_2, \ldots, i_k\}$ with $k \leq N$ are the images up to the rank k. $prec@k$ denotes the precision@k score, which is the ratio of retrieved relevant images over the retrieved images, and $rel@k$ is a relevance function which equals 1 if the image at rank k is relevant and 0 otherwise. Generally speaking, AP is the average of the precision values at each relevant image. To evaluate the overall performance, we calculate the mAP score by summing up and averaging the AP scores of each class $c \in C$:

$$mAP = \frac{1}{|C|} \sum_{c=1}^{C} AP_c \quad (5)$$

In the Appendix, we also present the predictions in the form of a confusion matrix, where the groundtruth labels are plotted in the rows and the predictions of the model in the columns. The colour of each entry indicates how many images of each label were assigned to which prediction. With a perfect classifier, all values would be located on the diagonal.

4 | RESULTS

4.1 | Neural network architectures

First, we compare the species prediction results of the different model architectures we trained. Table 4 shows the metrics calculated over all classes as percentage values.

Overall, the ConvNeXt models perform best, reaching mAPs of up to 97.91% on the validation data and 90.39% and 82.88% on the MOF and BNP test data sets, respectively. The Swin Transformer and EfficientNetV2 models both perform slightly worse on the validation data (Val mAP −2.01% and −2.85%, respectively) but decrease in performance on the MOF (mAP −3.45% and −2.08%, respectively) and even stronger on the BNP (mAP −8.22% and −14.07%, respectively) test data sets. The ResNet models perform worse but still achieve acceptable performance (Val mAP −7.85%, MOF mAP −9.25%, and BNP mAP −26.24%). ViT, on the other hand, performs even worse and hardly works at all on the test data sets.

All models perform better on the MOF test data set than on the BNP data (mAP difference of 7.62% [ConvNeXt] up to 24.61% [ResNet]), which we attribute to the fact that the images of the latter have an overall poorer image quality. In addition, the images were taken from lower camera angles than in most of the training data sets and focus on smaller animals such as rodents rather than deer, leading to a larger distributional shift between training and test images.

It is also noticeable that most models achieve very high values in the Top 5 Accuracy, while the Top 1 Accuracy is much lower, especially on the test data. This shows that the models often only slightly miss the correct prediction.

4.2 | Class-wise evaluation

In the following comparisons, we focus on our best-performing ConvNeXt models. In Figure 3, we show the F1 score for each species. We also plot the amount of available training data for the corresponding species. The edible dormouse species (*Glis glis*), for which only 426 training boxes existed, was recognised worst on the test data (mean F1 score of 0.0%). The domestic dog (*Canis familiris*, mean F1 score of 32.20%) and Empty (mean F1 score of 33.02%) classes, for which a lot of training data existed, also performed rather poorly in comparison, which can be attributed to challenging images in the test data set. On the other hand, the Eurasian woodcock class (*Scolopax rusticola*), for which only 167 training boxes existed, was recognised very reliably (mean F1

TABLE 4 The results of the different model architectures. We report the mean metric values and the standard deviations in parentheses as percentage values.

Data set	Metric	ResNet-152	EfficientNetV2-M	ViT-Base	Swin-Base	ConvNeXt-Base
Val	Top 1 accuracy	89.23 (±0.07)	93.86 (±0.36)	59.95 (±3.96)	93.67 (±0.76)	95.45 (±0.32)
	Top 3 accuracy	96.53 (±0.07)	98.35 (±0.11)	77.91 (±3.20)	98.30 (±0.29)	98.90 (±0.13)
	Top 5 accuracy	97.95 (±0.06)	99.03 (±0.08)	84.26 (±2.69)	99.01 (±0.17)	99.37 (±0.08)
	Mean F1 score	79.91 (±0.25)	88.66 (±0.73)	32.67 (±4.36)	89.69 (±2.04)	93.85 (±1.69)
	Mean average precision	90.06 (±0.15)	95.06 (±0.27)	40.10 (±6.37)	95.90 (±1.58)	97.91 (±1.24)
MOF	Top 1 accuracy	74.77 (±1.41)	84.71 (±0.64)	16.64 (±0.68)	80.56 (±0.65)	86.38 (±0.61)
	Top 3 accuracy	88.50 (±1.09)	93.67 (±0.26)	34.42 (±2.13)	92.51 (±0.52)	94.29 (±0.35)
	Top 5 accuracy	92.78 (±0.92)	96.37 (±0.30)	47.03 (±3.06)	96.04 (±0.35)	96.94 (±0.24)
	Mean F1 score	73.24 (±1.07)	81.54 (±0.50)	20.13 (±1.54)	78.81 (±0.88)	82.43 (±1.12)
	Mean average precision	81.14 (±0.58)	88.31 (±0.36)	27.92 (±1.88)	86.94 (±0.82)	90.39 (±0.68)
BNP	Top 1 accuracy	69.05 (±0.27)	76.21 (±0.52)	34.27 (±0.87)	73.01 (±3.00)	77.37 (±2.41)
	Top 3 accuracy	80.12 (±0.58)	85.50 (±0.20)	42.45 (±0.96)	84.07 (±1.61)	87.98 (±1.76)
	Top 5 accuracy	84.10 (±0.48)	88.63 (±0.28)	47.36 (±0.82)	87.65 (±1.27)	91.04 (±1.48)
	Mean F1 score	33.08 (±1.12)	45.71 (±3.25)	4.01 (±1.51)	46.82 (±5.84)	62.03 (±1.52)
	Mean average precision	56.53 (±1.55)	68.70 (±3.44)	11.09 (±2.17)	74.55 (±4.05)	82.77 (±1.07)

Abbreviations: BNP, Białowieża National Park; MOF, Marburg Open Forest.

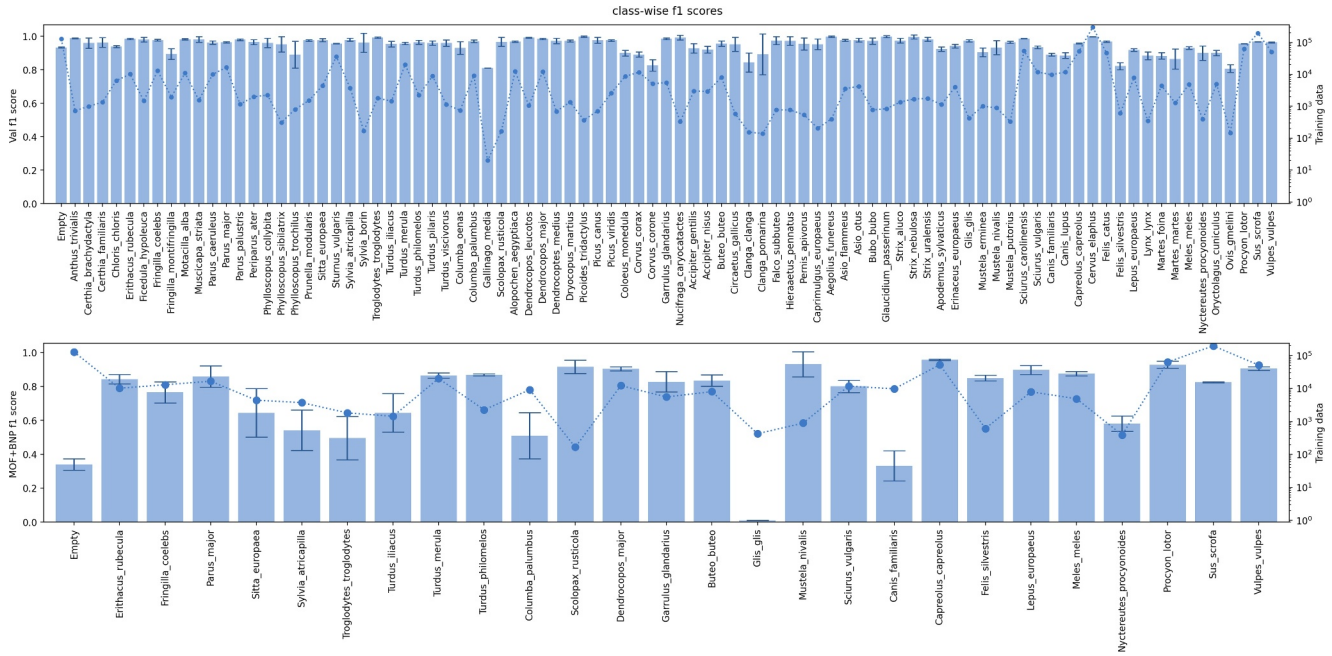

FIGURE 3 The F1 score calculated for each animal species. We show the mean values and the standard deviations as a bar chart. The dotted line indicates the number of available training data for each species.

score of 90.61%). We calculated correlation scores to investigate whether a statistical relationship exists between classification performance and the amount of training data used for each class. For the validation data, the Pearson correlation coefficient (PCC) is 0.142 and the Spearman's rank correlation coefficient (SCC) is 0.087; for the test data, the PCC is 0.025 and SCC is 0.180. Thus, no significant correlation is visible.

The confusion matrix (see Appendix F) shows that in many error cases, the models confuse related animal species that are difficult to distinguish visually, such as Carrion crow (*Corvus corone*) and Common raven (*Corvus corax*), Willow warbler (*P. trochilus*) and Common chiffchaff (*P. collybita*), Greater spotted eagle (*Clanga clanga*) and Lesser spotted eagle (*Clanga pomarina*), House cat (*F. catus*) and Wild cat (*F. silvestris*), European rabbit (*O. cuniculus*) and European hare (*L. europaeus*) as well as Stoat (*Mustela erminea*) and Least weasel (*Mustela nivalis*). In these cases, however, the correct species is listed in the top 5 results most of the time. The models therefore manage to recognise most classes with a high degree of reliability, but the performance of a class does not depend directly on the amount of available training data.

4.3 | Mammal and bird recognition

Our presented approach can recognise both mammals and bird species using a single model. For comparison, we also trained models to perform the recognition in two steps. We used a first model that was trained to distinguish between birds, mammals and empty images and further classified the non-empty images with either a model trained solely on mammal species or one trained solely on bird species recognition. We compared the results of this two-step classification approach to the one-step approach presented before in Table 5. Here, we also evaluated how well our models performed for recognising mammal species and bird species to investigate whether there is a difference in bird and mammal recognition performance. All compared models are based on the ConvNeXt-Base architecture.

The neural network models perform very well in recognising both birds and mammals. On the validation data, there is no significant difference in classification performance for birds and mammals. On the MOF test data, the models perform slightly more reliably for mammals and in the BNP test data slightly better for birds. One possible reason for this is that more birds were recorded in the latter, and due to the lower positioning of the camera, the birds were clearly visible in most cases, whereas in the MOF data set, the birds often appear further away from the camera and are often much harder to distinguish. However, since only a small proportion of the bird species to be recognised are represented in the two test data sets and some of the bird images in the validation data are high-quality photographs in which the birds are much easier to recognise than on camera trap images, a final judgement on the bird recognition quality of our models is difficult. On the other hand, larger mammals often appeared cropped in the BNP images and the data set contains some species that are difficult to distinguish, which is why the mammal performance of the models on this data set decreases. Again, some of the studied species are rarely or not at all present in the test data sets, but the validation data consists almost exclusively of camera trap images, which provides a good estimate of the overall performance for mammal recognition.

TABLE 5 A comparison between our one-stage approach and a two-stage approach. Here, we present the metrics for all species and for birds and mammals only. We report the mean metric values and the standard deviations in parentheses as percentage values.

		One model			Separate models		
Data set	**Metric**	**All**	**Birds**	**Mammals**	**All**	**Birds**	**Mammals**
Val	Top 1 accuracy	95.45 (±0.32)	95.14 (±0.94)	96.49 (±0.24)	85.29 (±0.07)	93.89 (±0.28)	97.59 (±0.18)
	Mean F1 score	93.85 (±1.69)	95.41 (±1.90)	93.21 (±1.40)	90.67 (±0.04)	92.35 (±0.21)	88.51 (±0.50)
	Mean average precision	97.91 (±1.24)	98.13 (±1.23)	98.11 (±1.06)	94.07 (±0.08)	95.78 (±0.21)	91.79 (±0.22)
MOF	Top 1 accuracy	86.38 (±0.61)	86.28 (±2.27)	87.25 (±0.91)	88.08 (±0.41)	93.03 (±1.39)	93.77 (±0.55)
	Mean F1 score	82.43 (±1.12)	79.84 (±1.31)	87.92 (±1.01)	83.61 (±3.55)	83.42 (±7.49)	86.21 (±0.70)
	Mean average precision	90.39 (±0.68)	90.96 (±1.97)	94.52 (±0.51)	83.78 (±3.70)	85.51 (±7.40)	86.29 (±0.84)
BNP	Top 1 accuracy	77.37 (±2.41)	78.04 (±2.61)	67.66 (±2.94)	85.22 (±0.42)	85.71 (±0.48)	78.07 (±2.32)
	Mean F1 score	62.03 (±1.52)	70.84 (±3.31)	56.07 (±3.91)	61.02 (±1.39)	65.58 (±2.53)	54.20 (±1.07)
	Mean average precision	82.77 (±1.07)	90.52 (±1.09)	81.94 (±1.06)	80.45 (±0.75)	90.90 (±0.57)	66.87 (±2.44)

Abbreviations: BNP, Białowieża National Park; MOF, Marburg Open Forest.

The models that identify bird and mammal species in a single step perform slightly better than the models of the two-step approach on both the validation (Val mAP +3.84%) and the test data sets (MOF mAP +6.61%, BNP mAP +2.32%). This shows that the combined classification strategy helps the models to learn better representations. In addition, with the two-step approach, errors in the first step (distinguishing between mammals and birds) directly lead to an incorrect overall result. In such cases, not even the Top 5 prediction is helpful, since the species recognition models can only predict species of their respective class and cannot make a meaningful prediction for images of the other class. The two-stage approach also requires significantly more computing resources, since three models must be loaded simultaneously. It is therefore not advantageous to introduce an additional model that initially distinguishes between birds and mammals, but instead to process all images of a data set with a single model. This makes it possible to identify both mammals and bird species at the same time.

4.4 | Measure comparison

We have taken various measures to overcome the challenges described in Section 3.5, namely using data augmentation, cropping the images to the areas where animals were detected, filtering out images with low detection confidence, sampling the training images in each epoch to an approximately equal number of images of each species, and applying label smoothing to counteract inaccurately labelled data. We now investigate the influence of these measures on the overall performance. For our baseline model, we applied data augmentation, cropping, filtering and sampling and compare it to models where we omit one of the measures each time or add label smoothing. Again, all compared models use the ConvNeXt-Base architecture.

As Table 6 shows, omitting any of the measures used in training our baseline model leads to a decrease in model performance. Not using sampling during training has a negative effect on validation performance (Val mAP −4.19%), but a slight improvement is visible on the test data (MOF mAP +0.58%, BNP +0.43%). We attribute this to the fact that sampling primarily improves the recognition of rare species, which in turn do not occur in our test data. The fact that a difference is visible on the validation data shows that our sampling method helps to reduce the imbalance between the classes. Not filtering out detections with small confidences and small bounding boxes leads to a minor performance reduction on the MOF data set (mAP −0.41)%, but a visible reduction on the BNP data (mAP −4.00%). This decrease is due to the fact that in some cases, training is performed on falsely detected areas that do not contain any animals at all. A performance decrease on both test data sets is also achieved by omitting data augmentation (MOF mAP −1.44%, BNP mAP −5.38%), which shows that although the training data set we used is quite large, increasing its variability via augmentation is still beneficial. Not cropping the images to the areas found by the detection model causes a considerable drop in performance (MOF mAP −29.58%, BNP mAP −35.39%). For comparability, we here used only the images for training where the MegaDetector had detected an animal with the minimum confidence. The results show that training only on the image crops showing the animals is highly beneficial. Finally, we investigated adding label smoothing as a measure against falsely labelled data. We found that this resulted in a decrease in model performance on the validation data (Val mAP −2.18%), made almost no difference on the MOF data set (MOF mAP −0.16%), and only led to a slight improvement on the BNP data set (BNP maP +1.81%). Therefore, we decided not to use this measure in our final models.

During our hyperparameter optimisation, we examined models with different input sizes. For this purpose, we trained models of the ConvNext-Base architecture with square input images that were scaled to a side length of 224, 300, and 384 pixels. The number of neurons in the input layer was adjusted accordingly. We found that the models with an input

TABLE 6 The ConvNext model performance with and without various training measures. We report the mean metric values and the standard deviations in parentheses as percentage values.

Data set	Metric	Baseline	No augmentation	No detection	No filtering	No sampling	Label smoothing
Val	Top 1 accuracy	95.45 (±0.32)	94.97 (±0.05)	93.99 (±0.06)	91.27 (±0.16)	96.28 (±0.13)	95.80 (±0.06)
	Mean F1 score	93.85 (±1.69)	91.24 (±0.13)	88.04 (±0.26)	85.80 (±0.58)	89.46 (±0.37)	92.79 (±0.12)
	Mean average precision	97.91 (±1.24)	95.24 (±0.22)	93.71 (±0.15)	92.87 (±0.26)	93.72 (±0.25)	95.73 (±0.02)
MOF	Top 1 accuracy	86.38 (±0.61)	84.80 (±0.60)	66.31 (±1.98)	81.63 (±0.59)	88.99 (±1.01)	87.30 (±0.75)
	Mean F1 score	82.43 (±1.12)	80.85 (±0.76)	47.31 (±2.51)	81.98 (±2.49)	82.26 (±1.41)	84.30 (±0.59)
	Mean average precision	90.39 (±0.68)	88.95 (±0.64)	60.81 (±3.35)	89.98 (±2.05)	90.97 (±0.88)	90.23 (±0.51)
BNP	Top 1 accuracy	77.37 (±2.41)	68.06 (±0.74)	40.04 (±2.97)	74.39 (±0.92)	67.43 (±2.58)	77.76 (±1.00)
	Mean F1 score	62.03 (±1.52)	60.24 (±2.71)	32.26 (±2.25)	61.31 (±1.88)	66.32 (±4.13)	64.70 (±3.27)
	Mean average precision	82.77 (±1.07)	77.59 (±1.34)	47.38 (±4.66)	78.77 (±1.88)	83.20 (±2.45)	84.58 (±0.63)

Abbreviations: BNP, Białowieża National Park; MOF, Marburg Open Forest.

size of 224 pixels performed best on the test data sets, closely followed by the models with an input size of 300 pixels and with a considerable gap to the models with an input size of 384 pixels. We also investigated different strategies to reduce the learning rate and found that a plateau reduction when the validation loss is stagnant works best, with a cosine decay performing only slightly worse. A more detailed comparison of the performance of the individual models can be found in Appendix C.

4.5 | Day/night performance

To investigate how well our models can recognise the desired species on images taken during the day and at night, we divided the validation and test data into day and night images. Since the capture time was not always present in the image metadata, we made the distinction based on whether the images were colour or black and white, which we determined based on the similarity of the mean values of each colour channel. The black and white images are most likely infrared images recorded at night. Although this method does not guarantee 100% accuracy, it is sufficient to perform a general analysis of the day and night performance.

In Figure 4, we show the F1 scores for each species on the day and night images side by side. We also indicate the respective number of available day and night training images. For all bird species, we had far more images available for training that were taken during the day, which is also due to the fact that many of them are not camera trap images and most of the birds studied are not nocturnal. For the mammals, the ratio of daytime and nighttime images is much more balanced. In Appendix D, we show the calculated overall metrics separated for day and night images. On the validation data, our models perform considerably better on the day images (mAP 97.93%) than on the night images (mAP 91.64%, difference of 6.29%). This trend is also evident for the images of the BNP test data set (mAP difference of 8.43%). On the MOF test data, however, the models perform better on the night images (mAP difference of −5.22%), which surprised us. We assume that this is due to the fact that many roe deer (*C. capreolus*), raccoons (*Procyon lotor*), wild boars (*Sus scrofa*) and red foxes (*Vulpes vulpes*) appear in this data set at night, which our models can recognise very reliably.

An examination of the class-wise metrics shows that our models can even recognise bird species on night images for which hardly any night training images were available, albeit less reliably than on day images. For example, the Eurasian blackcap (*Sylvia atricapilla*), for which 3722 training day images and only 4 training night images were available, achieved an F1 score of 74.75% on the night test images. This shows that the models manage to generalise to a certain degree between day and night images. There are also species, such as wild boar, for which our models perform better on the night images than on the day images (F1 score difference of 32.57%). This is probably due to the fact that the MOF test data set contains some very challenging daytime images that show herds with many individuals that are difficult to distinguish visually.

We again calculated correlation coefficients to check whether there is a statistical relationship between the species recognition performance on day and night images and the number of corresponding training images. On the validation data, we found a PCC of 0.014 and SCC of 0.162 for the day images and a PCC of 0.200 and SCC of 0.008 for the night images. For the test data, we found a PCC of −0.560 and SCC of −0.007 for the day images and a PCC of −0.088 and SCC of 0.435 for the night images, neither of which indicates a strong correlation. This evaluation shows that the models perform better on the day images, but also reliably recognise many species on night images, even when hardly any night images were available for training.

4.6 | Taxonomic classification

We now examine how our models trained for taxonomic classification perform. We compare our best species

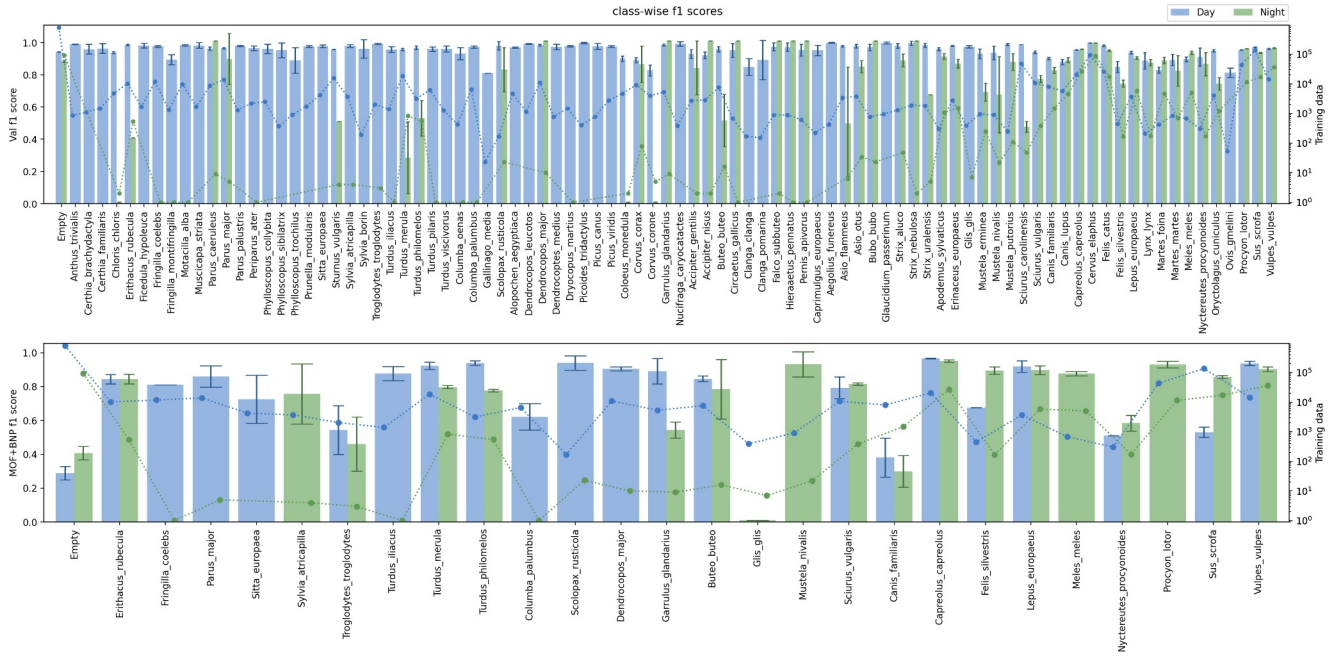

FIGURE 4 The F1 score calculated for each animal species. We show the mean values and the standard deviations as a bar chart for the day images in blue and the night images in green, respectively. The dotted lines indicate the number of available day or night training data, respectively.

classification models with taxonomic models where an additional fully connected output layer for each other taxonomic level was added after a shared pooling and dropout layer. We trained taxonomic classification models with fixed loss weights for each output head and compare it to our *loss weight shifting strategy* described in Section 3.4, which adjusts the loss weights during training. Apart from the loss weights, the taxonomic models were trained with the same settings as the species classification models.

One disadvantage of training the taxonomic models is the significantly increased time required, since the much larger data sets lead to around four times the training time than for the species models. However, the inference time only increases slightly, since the model size only increases by the additional output heads. In Table 7, we show the metrics for the species output head as well as the metrics for all output heads combined. We also calculated separate metrics for all six output heads in Appendix E.

The taxonomic models can make robust predictions for all taxonomic levels on the validation and test data sets. However, for pure species classification, the taxonomic models perform considerably worse on all data sets than our models trained only for species classification (Val mAP −5.67%, MOF mAP −7.96%, and BNP mAP −3.28%). This indicates that the large amount of additional training data alone does not lead to the learning of better species representations. However, by introducing our loss weight shifting strategy, the performance of the taxonomic models increases on all data sets (Val mAP +3.98%, MOF mAP +14.05%, and BNP mAP +1.00%) and even reaches a new best value of an mAP of 96.48% on the MOF data set. On the other data sets, the performance approaches that of the species classification models but does not quite reach it (Val mAP difference of 1.69%, BNP mAP difference of 2.28%).

We further investigated to what extent incorrect predictions of the model were carried through the entire taxonomic hierarchy and found that errors at the species level were very often accompanied by errors at a higher taxonomic level. However, the higher parts of the predicted taxonomy were often correct. Errors at the taxonomic class level, however, tend to run through the entire taxonomic hierarchy. This shows that the outputs have not learnt independent representations but are correlated with each other. There are also cases where only the prediction at a single taxonomic level is incorrect. Therefore, we developed a post-processing method to find inconsistencies in a taxonomy predicted by the model. This method iterates over the predicted taxonomic hierarchy labels from bottom to top and checks for each label whether the predicted label of the next higher level can be a taxonomic parent of the current label. If there is no match, the prediction is marked as inconsistent. In this case, we propose two possible approaches. We can calculate the taxonomy with the highest overall confidence, which is coherent, and thereby try to correct minor errors automatically. We do this by sorting the predictions across levels in the descending order of confidence. Then, we always replace one prediction with the next lowest confidence prediction of the same level and test whether this results in a coherent taxonomy. Finally, we return the coherent taxonomy with the highest overall probability. However, in cases where more than one prediction is incorrect, this strategy can also result in the overall prediction being coherent but with greater deviation from the desired solution. Therefore, our alternative solution is not to resolve images with an inconsistent taxonomic prediction but to mark them as uncertain in

order to submit them to a human expert for evaluation. When examining large amounts of data, a large saving in working time is still achieved, since a much smaller part of the data has to be analysed by a human.

5 | DISCUSSION

The presented neural network models achieve very good results on our two test data sets and show that they can be used in practice without prior adaptation to the test environment. Most errors are made on images that are challenging also for human experts due to the poor visibility of the animals.

Figure 5 shows some examples of images that were challenging for our models. From left to right, four images each are shown for which our best models provided a correct Top 1, Top 5, Top 10 classification or no correct Top 10 classification at all. Many of the instances that are difficult to identify for humans are correctly identified by our model. The false classifications shown can mostly be attributed to the poor visibility of the animals. In many cases, they appear blurred, cropped, or partially obscured, so that relevant distinct features are barely visible or not visible at all. This leads to confusion between similar looking classes or the model predicting 'empty'. However, there are also images where the animals can actually be seen well enough for recognition, for example, the squirrel in

TABLE 7 Species level and overall taxonomy recognition results of the taxonomic classification models in comparison to the species classification models. We report the mean metric values and the standard deviations in parentheses as percentage values.

		Species level			All levels		
Data set	**Metric**	**Species model**	**Taxonomic model**	**Weight shift**	**Species model**	**Taxonomic model**	**Weight shift**
Val	Top 1 accuracy	95.45 (±0.32)	47.62 (±33.23)	94.89 (±0.39)	97.76 (±1.87)	44.42 (±29.39)	87.53 (±0.71)
	Mean F1 score	93.85 (±1.69)	74.07 (±12.55)	91.09 (±0.84)	93.85 (±1.69)	79.86 (±8.68)	92.13 (±0.72)
	Mean average precision	97.91 (±1.24)	92.24 (±2.83)	96.22 (±0.36)	97.91 (±1.24)	94.52 (±1.78)	97.18 (±0.26)
MOF	Top 1 accuracy	86.38 (±0.61)	81.28 (±4.73)	87.19 (±1.48)	87.35 (±0.16)	77.54 (±4.49)	83.36 (±1.79)
	Mean F1 score	82.43 (±1.12)	76.90 (±5.80)	86.67 (±3.67)	82.43 (±1.12)	78.58 (±3.15)	86.44 (±1.09)
	Mean average precision	90.39 (±0.68)	82.43 (±5.99)	96.48 (±0.39)	90.39 (±0.68)	85.96 (±3.35)	94.23 (±0.32)
BNP	Top 1 accuracy	77.37 (±2.41)	68.46 (±4.17)	75.26 (±1.84)	86.90 (±6.02)	68.51 (±7.83)	72.35 (±2.05)
	Mean F1 score	62.03 (±1.52)	58.21 (±3.73)	53.56 (±2.36)	62.03 (±1.52)	62.25 (±1.68)	61.48 (±1.14)
	Mean average precision	82.77 (±1.07)	79.49 (±3.30)	80.49 (±2.20)	82.77 (±1.07)	83.25 (±1.51)	79.92 (±1.36)

Abbreviations: BNP, Białowieża National Park; MOF, Marburg Open Forest.

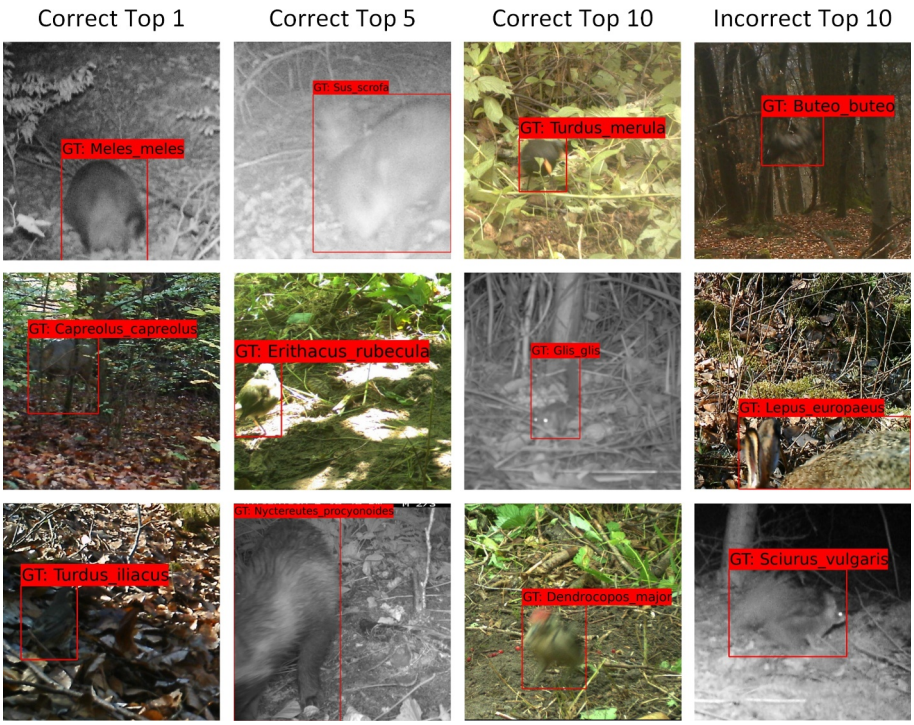

FIGURE 5 A selection of images from our test data sets on which our models had increasing difficulties predicting the correct species.

the bottom right image. This shows that the models do not work perfectly in all situations and need to be better adapted to such cases, for example, with more training data from the same domain.

The combined classification of mammals and birds with only a single model performs superior to the classification with a two-step approach with models trained only on mammals or birds. This eliminates the necessity of having to use another model to distinguish between birds and mammals. The classification of birds on camera trap images, which we carried out for the first time in this way, also works well despite the small amount of training data from a matching domain. Even though bird recognition in audio recordings is a proven method that works very well in most cases, our approach offers a viable alternative that allows researchers to monitor mammals and birds simultaneously with camera traps. Especially for abundance studies, our method is better suited than audio-based approaches, since multiple individuals can be better distinguished in images than in soundscape recordings. Furthermore, by detecting non-vocal bird species, our method bridges a gap that is left open by monitoring exclusively with microphones.

By using the taxonomic hierarchy labels and ignore labels, we were able to use a much larger pool of training data from existing datasets. We hoped that the models would also be able to learn features about the species we were looking for in our study from the images of related or inaccurately labelled species. The results show that the outputs of different taxonomic levels of a model show correlations, but this does not lead to better generalising models. For pure species classification, the performance of the taxonomic classification models falls slightly behind that of the models trained only for species classification on less data; only on the MOF data set do our taxonomic models outperform the species classification models. We suspect that due to the complexity of the data and the existing domain shift between different data sets, even larger amounts of training data might not bring any major improvement.

The taxonomic predictions can, however, be used to help understanding how the network achieved its predictions, for example, how the outputs correlate and what images challenge the model. We plan to conduct further research in the area of explainable artificial intelligence to develop new methods that can be derived from the prediction hierarchy. The method presented for detecting inconsistencies in the taxonomy also helps to find images where the model's predictions are uncertain. This information is particularly useful when analysing large amounts of data, where a complete check of all predictions by experts would be very time-consuming. By finding inconsistencies, a smaller selection of the most important images to be checked can be made in a simple way. This method could also be used in an active learning workflow to determine the images that should be labelled next in order to maximise the benefit for further training.

Our comparison of various current neural network architectures shows that modern CNN architectures such as ConvNeXt and EfficientNetV2 are well suited for the challenging analysis of camera trap images. ConvNeXt also outperforms a Swin Transformer architecture with the same number of parameters. The poor performance of the ViT, which has been known to perform better in other tasks, was surprising to us. Our attempts to find a hyperparameter combination that would improve training were not successful in this case.

In summary, our proposed approach provides a new method for the visual recognition of birds that can be very valuable for ecological research. We make our models and test data sets available for download to enable other researchers to apply or further improve the models. One possible example would be the fine-tuning of our taxonomic classification models to animal species from a different study region that are related to those on which our models were trained.

6 | CONCLUSION

In this article, we presented a deep learning approach for analysing camera trap images that can distinguish not only between mammals but also between bird species, providing a viable method for researchers to monitor mammals and birds simultaneously with camera traps. We developed neural network models for species classification as well as for predicting the animal taxonomy, that is, genus, family, order, group, and class names. The latter can assist researchers in analysing challenging images where accurate species classification is difficult by allowing an automatic determination of the higher taxonomic level. In addition, the outputs of the model can make it easier to find cases where the model was uncertain regarding the automatic species identification by searching for inconsistent taxonomic predictions.

We first localised the animals using Microsoft's MegaDetector and then predicted the species and taxonomy using our trained classification models. We compared the performance of several recent neural network architectures and found that the models of the ConvNeXt architecture performed best. Our species classification models achieved a mAP of 97.91% on a validation set left out from our training data and mAPs of 90.39% and 82.77%, respectively, on our a test data sets recorded in Germany and Poland. Our best taxonomic classification models reached an overall validation mAP of 97.18% and mAPs of 94.23% and 79.92% on the two test data sets, respectively. Most of the models' errors can be attributed to animals appearing cropped, obscured or blurred in the images, which makes recognition difficult even for human experts.

There are several areas of future work. We plan to explore ways to reduce the domain shift between high quality animal photos and camera trap images in order to make the models robust for species for which hardly any camera trap images exist for training. We also intend to collect more camera trap images of birds to enable a more thorough evaluation of our models for bird species recognition. We want to investigate to what extent the prediction of a taxonomic hierarchy can be scaled to a larger set of species and how fine-tuning of an existing neural network is possible. Furthermore, we will search for better strategies to balance the amount of data per species during training to achieve performance improvements for both frequent and rare classes. Finally, we plan to explore ways to

improve the training of our taxonomic classification models so that they can learn better hierarchical representations of the data, for example, by incorporating the taxonomic hierarchy more strongly during loss calculation.

AUTHOR CONTRIBUTIONS

Daniel Schneider: Conceptualisation; data curation; methodology; software; validation; visualisation; writing – original draft; writing – review & editing. **Kim Lindner**: Conceptualisation; data curation; methodology; resources; writing – original draft; writing – review & editing. **Markus Vogelbacher**: Data curation; writing – review & editing. **Hicham Bellafkir**: Data curation; writing – review & editing. **Nina Farwig**: Funding acquisition; supervision; writing – review & editing. **Bernd Freisleben**: Funding acquisition; supervision; writing – original draft; writing – review & editing.

ACKNOWLEDGEMENTS

This work is funded by the Hessian State Ministry for Higher Education, Research and the Arts (HMWK) (LOEWE Natur 4.0, LOEWE emergenCITY and hessian.AI Connectom AI4Birds, AI4BirdsDemo), and the German Research Foundation (DFG, Project 210487104—SFB 1053 MAKI). We thank the administration of the BNP, the forestry administrations of Białowieża, Hajnówka and Browsk, and Polish authorities (Ministry of Environment, GDOS and RDOS) for the permissions to work in Białowieża Forest.

Open access funding enabled and organized by Projekt DEAL.

CONFLICT OF INTEREST STATEMENT

The authors declare no potential conflict of interests.

DATA AVAILABILITY STATEMENT

We make our best trained models publicly available at https://github.com/umr-ds/Mammal-Bird-Camera-Trap-Recognition to enable other researchers to build on our work and apply our models to their own data or develop them further. We publish our MOF and BNP test data sets consisting of roughly 2500 and 15,000 camera trap images, respectively, in the same repository.

FINANCIAL DISCLOSURE

None reported.

ORCID

Daniel Schneider https://orcid.org/0000-0002-2798-4973

REFERENCES

1. Hooper, D.U., et al.: A global synthesis reveals biodiversity loss as a major driver of ecosystem change. Nature 486(7401), 105–108 (2012). https://doi.org/10.1038/nature11118
2. Díaz, S.M., et al.: Summary for policymakers of the global assessment report on biodiversity and ecosystem services. Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services (IPBES) (2019)
3. Wägele, J.W., et al.: Towards a multisensor station for automated biodiversity monitoring. Basic Appl. Ecol. 59, 105–138 (2022). https://doi.org/10.1016/j.baae.2022.01.003
4. Silveira, L., Jácomo, A.T.A., Diniz-Filho, J.A.F.: Camera trap, line transect census and track surveys: a comparative evaluation. Biol. Conserv. 114(3), 351–355 (2003). https://doi.org/10.1016/s0006-3207(03)00063-6
5. O'Connell, A.F., Nichols, J.D., Karanth, K.U.: Camera Traps in Animal Ecology: Methods and Analyses, vol. 271. Springer, New York (2011)
6. Schneider, S., et al.: Three critical factors affecting automated image species recognition performance for camera traps. Ecol. Evol. 4, 10 (2020)
7. Chen, G., et al.: Deep convolutional neural network based species recognition for wild animal monitoring. In: IEEE International Conference on Image Processing (ICIP), pp. 858–862 (2014)
8. Shonfield, J., Bayne, E.: Autonomous recording units in avian ecological research: current use and future applications. Avian Conserv. Ecol. 12(1), art14 (2017). https://doi.org/10.5751/ace-00974-120114
9. Priyadarshani, N., Marsland, S., Castro, I.: Automated birdsong recognition in complex acoustic environments: a review. J. Avian Biol. 49(5) (2018). https://doi.org/10.1111/jav.01447
10. Ross, SRPJ, et al.: Passive acoustic monitoring provides a fresh perspective on fundamental ecological questions. Funct. Ecol. 37(4), 959–975 (2023). https://doi.org/10.1111/1365-2435.14275
11. Chen, A., et al.: Using computer vision, image analysis and UAVs for the automatic recognition and counting of common cranes (*Grus grus*). J. Environ. Manag. 328, 116948 (2023). https://doi.org/10.1016/j.jenvman.2022.116948
12. Norouzzadeh, M.S., et al.: Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning. Proc. Natl. Acad. Sci. USA 6(25), 115 (2018). https://doi.org/10.1073/pnas.1719367115
13. 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). https://doi.org/10.1111/2041-210x.13120
14. Choiński, M., et al.: A first step towards automated species recognition from camera trap images of mammals using AI in a European temperate forest. In: Saeed, K., Dvorský, J. (eds.) Computer Information Systems and Industrial Management, pp. 299–310. Springer, Cham (2021)
15. He, K., et al.: Identity mappings in deep residual networks. In: European Conference on Computer Vision (ECCV), pp. 630–645. Springer. Springer International Publishing (2016). http://link.springer.com/10.1007/978-3-319-46493-0_38
16. Tan, M., Le, Q.: EfficientNetV2: smaller models and faster training. In: International Conference on Machine Learning (ICML), pp. 10096–10106. PMLR (2021)
17. Dosovitskiy, A., et al.: An image is worth 16x16 words: transformers for image recognition at scale. In: 9th International Conference on Learning Representations (ICLR), Austria. OpenReview.net (2021)
18. Liu, Z., et al.: Swin transformer: hierarchical vision transformer using shifted windows. In: IEEE/CVF International Conference on Computer Vision, pp. 10012–10022 (2021)
19. Liu, Z., et al.: A ConvNet for the 2020s. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 11976–11986 (2022)
20. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature 521(7553), 436–444 (2015). https://doi.org/10.1038/nature14539
21. Fukushima, K.: Neocognitron: a self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Biol. Cybern. 36(4), 193–202 (1980). https://doi.org/10.1007/bf00344251
22. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. Adv. Neural Inf. Process. Syst. 25, 1097–1105 (2012)
23. He, K., et al.: Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778 (2016)
24. Tan, M., Le, Q.: EfficientNet: rethinking model scaling for convolutional neural networks. In: 36th International Conference on Machine Learning (ICML), vol. 97, pp. 6105–6114. PMLR (2019)
25. Vaswani, A., et al.: Attention is all you need. Adv. Neural Inf. Process. Syst. 30, 5999–6009 (2017)
26. Srivastava, N., Salakhutdinov, R.R.: Discriminative transfer learning with tree-based priors. Adv. Neural Inf. Process. Syst. 26 (2013)
27. Yan, Z., et al.: HD-CNN: hierarchical deep convolutional neural networks for large scale visual recognition. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2740–2748 (2015)

28. Wehrmann, J., Cerri, R., Barros, R.: Hierarchical multi-label classification networks. In: International Conference on Machine Learning, pp. 5075–5084. PMLR (2018)
29. Koo, J., Klabjan, D., Utke, J.: Combined convolutional and recurrent neural networks for hierarchical classification of images. In: 2020 IEEE International Conference on Big Data (Big Data), pp. 1354–1361. IEEE (2020)
30. Turkoglu, M.O., et al.: Crop mapping from image time series: deep learning with multi-scale label hierarchies. Rem. Sens. Environ. 10, 264 (2021)
31. Elhamod, M., et al.: Hierarchy-guided neural network for species classification. Methods Ecol. Evol. 13(3), 642–652 (2022). https://doi.org/10.1111/2041-210x.13768
32. Cramer, A.L., et al.: Chirping up the right tree: incorporating biological taxonomies into deep bioacoustic classifiers. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 901–905. IEEE (2020)
33. Wang, Q., et al.: Hierarchical-taxonomy-aware and attentional convolutional neural networks for acoustic identification of bird species: a phylogenetic perspective. Ecol. Inf. 5, 80 (2024)
34. Yu, X., et al.: Automated identification of animal species in camera trap images. EURASIP J. Image Video Process. 2013(1), 52 (2013). https://doi.org/10.1186/1687-5281-2013-52
35. Russakovsky, O., et al.: ImageNet large scale visual recognition challenge. Int. J. Comput. Vis. 115(3), 211–252 (2015). https://doi.org/10.1007/s11263-015-0816-y
36. Swanson, A., et al.: Snapshot Serengeti, high-frequency annotated camera trap images of 40 mammalian species in an African savanna. Sci. Data 6(1), 2 (2015). https://doi.org/10.1038/sdata.2015.26
37. Villa, A.G., Salazar, A., Vargas, F.: Towards automatic wild animal monitoring: identification of animal species in camera-trap images using very deep convolutional neural networks. Ecol. Inf. 41, 24–32 (2017). https://doi.org/10.1016/j.ecoinf.2017.07.004
38. Willi, M., et al.: Identifying animal species in camera trap images using deep learning and citizen science. Methods Ecol. Evol. 10, 80–91 (2019). https://doi.org/10.1111/2041-210x.13099
39. Beery, S., et al.: The iWildCam 2018 challenge dataset. arXiv preprint arXiv:190405986 (2019)
40. Islam, S.B., Valles, D.: Identification of wild species in Texas from camera-trap images using deep neural network for conservation monitoring. In: 2020 10th Annual Computing and Communication Workshop and Conference, CCWC 2020, pp. 537–542 (2020)
41. Zualkernan, I., et al.: An IoT system using deep learning to classify camera trap images on the edge. Computers 1, 11 (2022)
42. Jia, L., Tian, Y., Zhang, J.: Domain-Aware neural architecture search for classifying animals in camera trap images. Animals 2, 12 (2022)
43. Auer, D., et al.: Minimizing the Annotation Effort for Detecting Wildlife in Camera Trap Images with Active Learning. Gesellschaft für Informatik, Bonn (2021). INFORMATIK
44. Kyathanahally, S.P., et al.: Ensembles of data-efficient vision transformers as a new paradigm for automated classification in ecology. Sci. Rep. 12(1), 12 (2022). https://doi.org/10.1038/s41598-022-21910-0
45. Zheng, Z., et al.: Wild terrestrial animal re-identification based on an improved locally aware transformer with a cross-attention mechanism. Animals 12(24), 12 (2022). https://doi.org/10.3390/ani12243503
46. Agilandeeswari, L., Meena, S.D.: SWIN transformer based contrastive self-supervised learning for animal detection and classification. Multimed. Tool. Appl. 82(7), 10445–10470 (2023). https://doi.org/10.1007/s11042-022-13629-x
47. Fabian, Z., et al.: Multimodal foundation models for zero-shot animal species recognition in camera trap images. http://arxiv.org/abs/2311.01064 (2023)
48. Schneider, S., Taylor, G.W., Kremer, S.C.: Deep learning object detection methods for ecological camera trap data. In: 15th Conference on Computer and Robot Vision (CRV), pp. 321–328 (2018)
49. Carl, C., et al.: Automated detection of European wild mammal species in camera trap images with an existing and pre-trained computer vision model. Eur. J. Wildl. Res. 66(4), 1–7 (2020). https://doi.org/10.1007/s10344-020-01404-y
50. Shepley, A., et al.: Location invariant animal recognition using mixed source datasets and deep learning. bioRxiv (2020)
51. Beery, S., et al.: Efficient pipeline for automating species ID in new camera trap projects. Biodiversity Inf. Sci. Stand. 3 (2019). https://doi.org/10.3897/biss.3.37222
52. 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 (2021). https://doi.org/10.1111/2041-210x.13504
53. Simões, F., Bouveyron, C., Precioso, F.: DeepWILD: Wildlife Identification, Localisation and estimation on camera trap videos using Deep learning. Ecol. Inf. 75, 102095 (2023). https://doi.org/10.1016/j.ecoinf.2023.102095
54. Kahl, S., et al.: BirdNET: a deep learning solution for avian diversity monitoring. Ecol. Inf. 61, 101236 (2021). https://doi.org/10.1016/j.ecoinf.2021.101236
55. Mühling, M., et al.: Bird species recognition via neural architecture search. In: Conference and Labs of the Evaluation Forum (CLEF), vol. 2696. Thessaloniki. CEUR-WS.org (2020)
56. Kahl, S., et al.: Overview of BirdCLEF 2020: bird sound recognition in complex acoustic environments. In: Conference and Labs of the Evaluation Forum (CLEF), vol. 2696. Thessaloniki. CEUR-WS.org (2020)
57. Höchst, J., et al.: Bird@Edge: bird species recognition at the edge. In: 10th International Conference on Networked Systems (NETYS), Virtual, May 17–19, pp. 69–86. Springer (2022)
58. Hong, S.J., et al.: Application of deep-learning methods to bird detection using unmanned aerial vehicle imagery. Sensors 4, 19 (2019)
59. Akçay, H.G., et al.: Automated bird counting with deep learning for regional bird distribution mapping. Animals 10, 1–24 (2020). https://doi.org/10.3390/ani10071207
60. Gradolewski, D., et al.: Comprehensive bird preservation at wind farms. Sensors 21, 1–35 (2021). https://doi.org/10.3390/s21010267
61. Huang, Y.P., Basanta, H.: Bird image retrieval and recognition using a deep learning platform. IEEE Access 7, 66980–66989 (2019). https://doi.org/10.1109/access.2019.2918274
62. Raj, S., et al.: Image based bird species identification using convolutional neural network. Int. J. Eng. Res. Technol. 9(06) (2020). https://doi.org/10.17577/ijertv9is060279
63. Ferreira, A.C., et al.: Deep learning-based methods for individual recognition in small birds. Methods Ecol. Evol. 11, 1072–1085 (2020). https://doi.org/10.1111/2041-210x.13436
64. Chalmers, C., et al.: Removing human bottlenecks in bird classification using camera trap images and deep learning. Rem. Sens. 5, 15 (2023)
65. Beery, S., Horn, G.V., Perona, P.: Recognition in terra incognita. In: 15th European Conference on Computer Vision (ECCV), pp. 472–489. Munich (2018)
66. Yousif, H., Kays, R., He, Z.: Dynamic programming selection of object proposals for sequence-level animal species classification in the wild. IEEE Trans. Circ. Syst. Video Technol. 20 (2019)
67. Zhang, Z., et al.: Animal detection from highly cluttered natural scenes using spatiotemporal object region proposals and patch verification. IEEE Trans. Multimed. 18(10), 2079–2092 (2016). https://doi.org/10.1109/tmm.2016.2594138
68. Imholt, C., et al.: Langfristige Populationsentwicklung krankheitsübertragender Nagetiere: Interaktion von Klimawandel, Landnutzung und Biodiversität: Abschlussbericht. Climate Change. Umweltbundesamt Deutschland (2021)

How to cite this article: Schneider, D., et al.: Recognition of European mammals and birds in camera trap images using deep neural networks. IET Comput. Vis. 18(8), 1162–1192 (2024). https://doi.org/10.1049/cvi2.12294

APPENDIX A

DATA SETS

The following Table A1 shows the number of training and validation bounding boxes we used from each data set (on the left side for training our species classification networks and on the right side for training our taxonomic classification networks).

APPENDIX B

SPECIES LABELS

The following Table B1 shows the labels we used for each taxonomic level (class, group, order, family, genus, and species) separated in birds and mammals. It also shows the number of training and validation bounding boxes we used for each label.

APPENDIX C

HYPERPARAMETER COMPARISON

We examined models with different input sizes. The following Table C1 shows a comparison of models of the ConvNext Base architecture with a square input of 224, 300 or 384 neurons side length that were trained with learning rate plateau reduction. The last model also has an input size of 224 × 224 neurons but is trained with a cosine learning rate decay.

APPENDIX D

DAY/NIGHT PERFORMANCE

Table D1 shows how the ConvNext models perform on images recorded by day or by night (infrared), respectively.

APPENDIX E

TAXONOMIC CLASSIFICATION

Table E1 shows the separate metrics for all six output heads of our taxonomic classification ConvNext models in comparison to the best species classification models.

APPENDIX F

CONFUSION MATRICES

The following images show the species confusion matrices for our best ConvNeXt models. We averaged the prediction counts of the models trained with different splits. Figures F1 and F2 show the matrices for the models trained only on species classification, Figures F3 and F4 show the matrices for the taxonomic classification models.

TABLE A1 The number of training and validation bounding boxes from each data set.

	Used for species classification				Used for taxonomic classification			
Data set name	**Labels**	**Boxes total**	**Boxes train**	**Boxes validation**	**Labels**	**Boxes total**	**Boxes train**	**Boxes validation**
Caltech Camera Traps	3	10,915	9882	1033	16	62,126	56,259	5867
ENA24-detection	5	2232	1788	444	17	9623	7705	1918
Idaho Camera Traps	4	75,699	74,699	1000	31	347,783	343,054	4729
Missouri Camera Traps	8	11,517	9768	1749	16	33,528	28,999	4529
North American Camera Trap Images	9	584,436	581,157	3279	32	4,810,962	4,799,749	11,213
Rodent	25	14,327	12,645	1682	39	18,025	15,634	2391
Tierschnappschuss	18	140,635	135,675	4960	29	163,648	156,479	7169
WCS Camera Traps	7	10,820	9805	1015	62	361,240	350,661	10,579
Kaggle Birds	6	1104	886	218	47	96,703	89,091	7612
North American Birds	2	429	344	85	43	49,840	44,168	5672
eMammal	16	13,485	11,140	2345	22	15,125	12,455	2670
InatCrawl	88	406,366	377,234	29,132	90	407,868	378,437	29,431
WebCrawl	7	1414	1135	279	7	1414	1135	279
Sum	198	1,273,379	1,226,158	47,221	451	6,377,885	6,283,826	94,059
Marburg Open Forest (MOF)	19	2420	0	0	27	2731	0	0
Białowieża National Park (BNP)	20	4810	0	0	36	16,831	0	0

TABLE B1 The number of bounding boxes for each label of the taxonomic levels used.

Hierarchy level	Bird names	Boxes train	Boxes validation	Mammal names	Boxes train	Boxes validation
Class	Aves	214,182	3967	Mammalia	102	25
Group	Bird_raptor	532	131	Mammal_large	56,764	1369
	Bird_small	9010	1229	Mammal_small	19,952	1595
Order	Accipitriformes	869	215	Artiodactyla	103,792	1060
	Anseriformes	186	45	Carnivora	52,195	1990
	Caprimulgiformes	1432	357	Eulipotyphla	289	70
	Charadriiformes	7942	1006	Rodentia	28,434	2075
	Passeriformes	43,951	1056			
	Piciformes	1343	335			
	Strigiformes	393	97			
Family	Accipitridae	3059	762	Bovidae	3,753,855	1511
	Anatidae	9722	1017	Canidae	14,552	1347
	Caprimulgidae	389	96	Cervidae	363,883	2262
	Certhiidae	312	77	Erinaceidae	16	3
	Columbidae	3298	823	Felidae	29,978	1466
	Corvidae	2430	607	Gliridae	0	0
	Falconidae	426	104	Leporidae	14,579	1957
	Fringillidae	4260	911	Muridae	1587	396
	Motacillidae	428	107	Mustelidae	1768	439
	Muscicapidae	1137	283	Procyonidae	3286	820
	Paridae	1033	257	Sciuridae	39,682	808
	Picidae	1962	489	Suidae	1354	338
	Scolopacidae	2688	671			
	Sittidae	164	40			
	Strigidae	254	61			
	Sturnidae	899	223			
	Sylviidae	97	24			
	Troglodytidae	559	139			
	Turdidae	1748	436			
Genus	Accipiter	386	95	Apodemus	1130	281
	Aegolius	80	20	Canis	30,682	1674
	Anthus	253	63	Cervus	143,094	1487
	Asio	132	33	Erinaceus	2	0
	Bubo	302	75	Felis	4085	547
	Buteo	1175	293	Lepus	15,324	661
	Caprimulgus	1	0	Lynx	32,021	1303
	Certhia	257	63	Martes	1958	488
	Columba	657	163	Mustela	82	20
	Corvus	2962	739	Ovis	14,718	1884
	Dendrocopos	127	31	Procyon	239	59
	Dryocopus	96	24	Sciurus	3262	787

(Continues)

TABLE B1 (Continued)

Hierarchy level	Bird names	Boxes train	Boxes validation	Mammal names	Boxes train	Boxes validation
	Erithacus	142	35	Sus	210	52
	Falco	1084	270			
	Ficedula	2	0			
	Gallinago	94	23			
	Glaucidium	86	21			
	Motacilla	8	0			
	Nucifraga	222	54			
	Parus	78	19			
	Pernis	2	0			
	Picoides	318	78			
	Scolopax	46	11			
	Sitta	459	113			
	Strix	99	24			
	Troglodytes	213	52			
	Turdus	807	200			
Species	Accipiter_gentilis	3018	538	Apodemus_sylvaticus	1135	283
	Accipiter_nisus	2907	500	Canis_familiaris	9760	2170
	Aegolius_funereus	398	99	Canis_lupus	11,832	1614
	Alopochen_aegyptiaca	12,427	539	Capreolus_capreolus	52,890	1741
	Anthus_trivialis	727	181	Cervus_elaphus	299,534	1505
	Asio_flammeus	3549	500	Erinaceus_europaeus	4014	788
	Asio_otus	4252	540	Felis_catus	47,156	1096
	Bubo_bubo	766	191	Felis_silvestris	611	149
	Buteo_buteo	8034	512	Glis_glis	426	106
	Caprimulgus_europaeus	207	51	Lepus_europaeus	7925	1699
	Certhia_brachydactyla	992	248	Lynx_lynx	352	86
	Certhia_familiaris	1365	341	Martes_foina	4401	697
	Chloris_chloris	6363	500	Martes_martes	1273	314
	Circaetus_gallicus	581	145	Meles_meles	4893	876
	Clanga_clanga	155	38	Mustela_erminea	1007	251
	Clanga_pomarina	140	34	Mustela_nivalis	893	220
	Coloeus_monedula	8637	500	Mustela_putorius	332	82
	Columba_oenas	746	186	Nyctereutes_procyonoides	390	96
	Columba_palumbus	9176	500	Oryctolagus_cuniculus	5022	1000
	Corvus_corax	11,700	1025	Ovis_gmelini	148	37
	Corvus_corone	5076	502	Procyon_lotor	63,470	1869
	Dendrocopos_leucotos	1051	262	Sciurus_carolinensis	53,743	1063
	Dendrocopos_major	12,210	508	Sciurus_vulgaris	11,811	1128
	Dendrocoptes_medius	693	173	Sus_scrofa	196,089	2601
	Dryocopus_martius	1354	337	Vulpes_vulpes	51,067	1683
	Erithacus_rubecula	10,178	507	Empty	128,055	1678

TABLE B1 (Continued)

Hierarchy level	Bird names	Boxes train	Boxes validation	Mammal names	Boxes train	Boxes validation
	Falco_subbuteo	772	193			
	Ficedula_hypoleuca	1476	369			
	Fringilla_coelebs	13,011	500			
	Fringilla_montifringilla	1962	490			
	Gallinago_media	20	4			
	Garrulus_glandarius	5538	604			
	Glaucidium_passerinum	839	209			
	Hieraaetus_pennatus	767	191			
	Motacilla_alba	11,045	500			
	Muscicapa_striata	1560	389			
	Nucifraga_caryocatactes	332	82			
	Parus_caeruleus	9966	500			
	Parus_major	16,605	538			
	Parus_palustris	1159	289			
	Periparus_ater	1998	499			
	Pernis_apivorus	540	134			
	Phylloscopus_collybita	2237	500			
	Phylloscopus_sibilatrix	309	77			
	Phylloscopus_trochilus	801	200			
	Picoides_tridactylus	369	92			
	Picus_canus	712	177			
	Picus_viridis	2624	513			
	Prunella_modularis	1516	378			
	Scolopax_rusticola	167	41			
	Sitta_europaea	4418	502			
	Strix_aluco	1355	338			
	Strix_nebulosa	1695	422			
	Strix_uralensis	1760	439			
	Sturnus_vulgaris	36,200	593			
	Sylvia_atricapilla	3737	500			
	Sylvia_borin	168	42			
	Troglodytes_troglodytes	1820	455			
	Turdus_iliacus	1442	360			
	Turdus_merula	20,072	529			
	Turdus_philomelos	2266	500			
	Turdus_pilaris	8836	500			
	Turdus_viscivorus	1133	283			

TABLE C1 Results of the ConvNext Base models with different hyperparameter configurations. We report the mean metric values and the standard deviations in parentheses as percentage values.

Data set	Metric	Input size: 224 Plateau reduction	Input size: 300 Plateau reduction	Input size: 384 Plateau reduction	Input size: 224 Cosine decay
Val	Top 1 accuracy	95.45 (±0.32)	95.87 (±0.16)	23.70 (±0.07)	95.66 (±0.28)
	Mean F1 score	93.85 (±1.69)	95.46 (±0.30)	61.85 (±1.38)	92.60 (±0.41)
	Mean average precision	97.91 (±1.24)	98.94 (±0.09)	90.22 (±0.36)	96.53 (±0.13)
MOF	Top 1 accuracy	86.38 (±0.61)	84.97 (±0.97)	76.81 (±0.29)	87.59 (±0.51)
	Mean F1 score	82.43 (±1.12)	82.16 (±1.67)	70.08 (±0.66)	83.40 (±0.43)
	Mean average precision	90.39 (±0.68)	88.75 (±1.36)	76.57 (±0.36)	90.46 (±0.87)
BNP	Top 1 accuracy	77.37 (±2.41)	77.16 (±0.95)	79.40 (±0.18)	76.61 (±0.46)
	Mean F1 score	62.03 (±1.52)	61.99 (±2.32)	52.82 (±1.24)	63.37 (±1.53)
	Mean average precision	82.77 (±1.07)	81.87 (±0.78)	78.98 (±1.69)	82.10 (±1.51)

TABLE D1 Results of the ConvNext Base models on the daytime and nighttime images. We report the mean metric values and the standard deviations in parentheses as percentage values.

Data set	Metric	All	Day	Night
Val	Top 1 accuracy	95.45 (±0.32)	95.56 (±0.35)	95.08 (±0.25)
	Mean F1 score	93.85 (±1.69)	94.15 (±1.64)	79.89 (±4.06)
	Mean average precision	97.91 (±1.24)	97.93 (±1.22)	91.64 (±1.96)
MOF	Top 1 accuracy	86.38 (±0.61)	86.81 (±0.58)	86.23 (±0.76)
	Mean F1 score	82.43 (±1.12)	79.06 (±1.61)	86.83 (±0.84)
	Mean average precision	90.39 (±0.68)	88.62 (±0.65)	93.84 (±0.61)
BNP	Top 1 accuracy	77.37 (±2.41)	83.10 (±3.58)	71.49 (±1.37)
	Mean F1 score	62.03 (±1.52)	70.76 (±5.60)	60.61 (±3.54)
	Mean average precision	82.77 (±1.07)	87.43 (±2.96)	79.00 (±0.28)

TABLE E1 Results of all output heads of the taxonomic classification models in comparison to the species classification models. We report the mean metric values and the standard deviations in parentheses as percentage values.

Data set	Taxonomic level	Metric	Species model	Taxonomic model	Weight shift model
Val	Species	Top 1 accuracy	95.45 (±0.32)	47.62 (±33.23)	94.89 (±0.39)
		Mean F1 score	93.85 (±1.69)	74.07 (±12.55)	91.09 (±0.84)
		Mean average precision	97.91 (±1.24)	92.24 (±2.83)	96.22 (±0.36)
	Genus	Top 1 accuracy		78.92 (±11.39)	95.33 (±0.48)
		Mean F1 score		81.15 (±9.24)	94.22 (±0.83)
		Mean average precision		95.50 (±1.63)	97.88 (±0.27)
	Family	Top 1 accuracy		83.72 (±2.51)	89.14 (±0.78)
		Mean F1 score		84.39 (±5.44)	92.95 (±0.71)
		Mean average precision		95.79 (±1.19)	97.94 (±0.25)
	Order	Top 1 accuracy		89.68 (±0.16)	91.23 (±0.46)
		Mean F1 score		91.92 (±0.58)	92.27 (±0.37)
		Mean average precision		98.46 (±0.30)	98.28 (±0.10)
	Group	Top 1 accuracy		90.95 (±0.20)	92.31 (±0.23)
		Mean F1 score		89.65 (±0.27)	90.21 (±0.15)
		Mean average precision		97.57 (±0.42)	97.44 (±0.09)
	Class	Top 1 accuracy		90.96 (±0.15)	92.14 (±0.16)
		Mean F1 score		75.62 (±0.10)	76.99 (±0.22)
		Mean average precision		95.89 (±0.31)	95.62 (±0.07)
	All	Top 1 accuracy	97.76 (±1.87)	44.42 (±29.39)	87.53 (±0.71)
		Mean F1 score	93.85 (±1.69)	79.86 (±8.68)	92.13 (±0.72)
		Mean average precision	97.91 (±1.24)	94.52 (±1.78)	97.18 (±0.26)
MOF	Species	Top 1 accuracy	86.38 (±0.61)	81.28 (±4.73)	87.19 (±1.48)
		Mean F1 score	82.43 (±1.12)	76.90 (±5.80)	86.67 (±3.67)
		Mean average precision	90.39 (±0.68)	82.43 (±5.99)	96.48 (±0.39)
	Genus	Top 1 accuracy		80.73 (±4.08)	85.70 (±1.64)
		Mean F1 score		75.04 (±3.81)	84.84 (±1.09)
		Mean average precision		84.69 (±4.19)	93.72 (±0.37)
	Family	Top 1 accuracy		84.72 (±3.82)	89.28 (±1.15)
		Mean F1 score		78.24 (±3.68)	89.60 (±0.69)
		Mean average precision		86.78 (±4.47)	96.16 (±0.47)
	Order	Top 1 accuracy		91.22 (±0.34)	92.20 (±0.49)
		Mean F1 score		80.40 (±0.45)	81.60 (±0.16)
		Mean average precision		87.72 (±3.66)	88.61 (±0.65)
	Group	Top 1 accuracy		94.50 (±0.22)	95.41 (±0.15)
		Mean F1 score		87.09 (±1.35)	87.61 (±1.93)
		Mean average precision		92.55 (±1.08)	92.79 (±0.42)
	Class	Top 1 accuracy		95.79 (±0.14)	96.08 (±0.14)
		Mean F1 score		87.10 (±0.42)	87.95 (±0.23)
		Mean average precision		90.40 (±0.72)	91.94 (±0.64)
	All	Top 1 accuracy	87.35 (±0.16)	77.54 (±4.49)	83.36 (±1.79)
		Mean F1 score	82.43 (±1.12)	78.58 (±3.15)	86.44 (±1.09)

(Continues)

TABLE E1 (Continued)

Data set	Taxonomic level	Metric	Species model	Taxonomic model	Weight shift model
		Mean average precision	90.39 (±0.68)	85.96 (±3.35)	94.23 (±0.32)
BNP	Species	Top 1 accuracy	77.37 (±2.41)	68.46 (±4.17)	75.26 (±1.84)
		Mean F1 score	62.03 (±1.52)	58.21 (±3.73)	53.56 (±2.36)
		Mean average precision	82.77 (±1.07)	79.49 (±3.30)	80.49 (±2.20)
	Genus	Top 1 accuracy		77.96 (±2.57)	78.22 (±1.71)
		Mean F1 score		58.34 (±2.37)	57.06 (±1.53)
		Mean average precision		82.13 (±3.11)	79.67 (±1.25)
	Family	Top 1 accuracy		81.49 (±1.12)	81.48 (±1.34)
		Mean F1 score		61.32 (±0.54)	63.35 (±1.65)
		Mean average precision		79.43 (±1.35)	76.57 (±1.49)
	Order	Top 1 accuracy		88.18 (±0.09)	89.71 (±0.57)
		Mean F1 score		75.85 (±0.27)	77.06 (±0.82)
		Mean average precision		94.97 (±0.48)	83.33 (±1.10)
	Group	Top 1 accuracy		91.40 (±0.21)	92.51 (±0.49)
		Mean F1 score		65.73 (±0.26)	68.18 (±1.21)
		Mean average precision		91.14 (±1.76)	78.62 (±1.87)
	Class	Top 1 accuracy		94.53 (±0.44)	95.17 (±0.28)
		Mean F1 score		96.19 (±0.17)	96.67 (±0.16)
		Mean average precision		99.61 (±0.03)	99.70 (±0.02)
	All	Top 1 accuracy	86.90 (±6.02)	68.51 (±7.83)	72.35 (±2.05)
		Mean F1 score	62.03 (±1.52)	62.25 (±1.68)	61.48 (±1.14)
		Mean average precision	82.77 (±1.07)	83.25 (±1.51)	79.92 (±1.36)

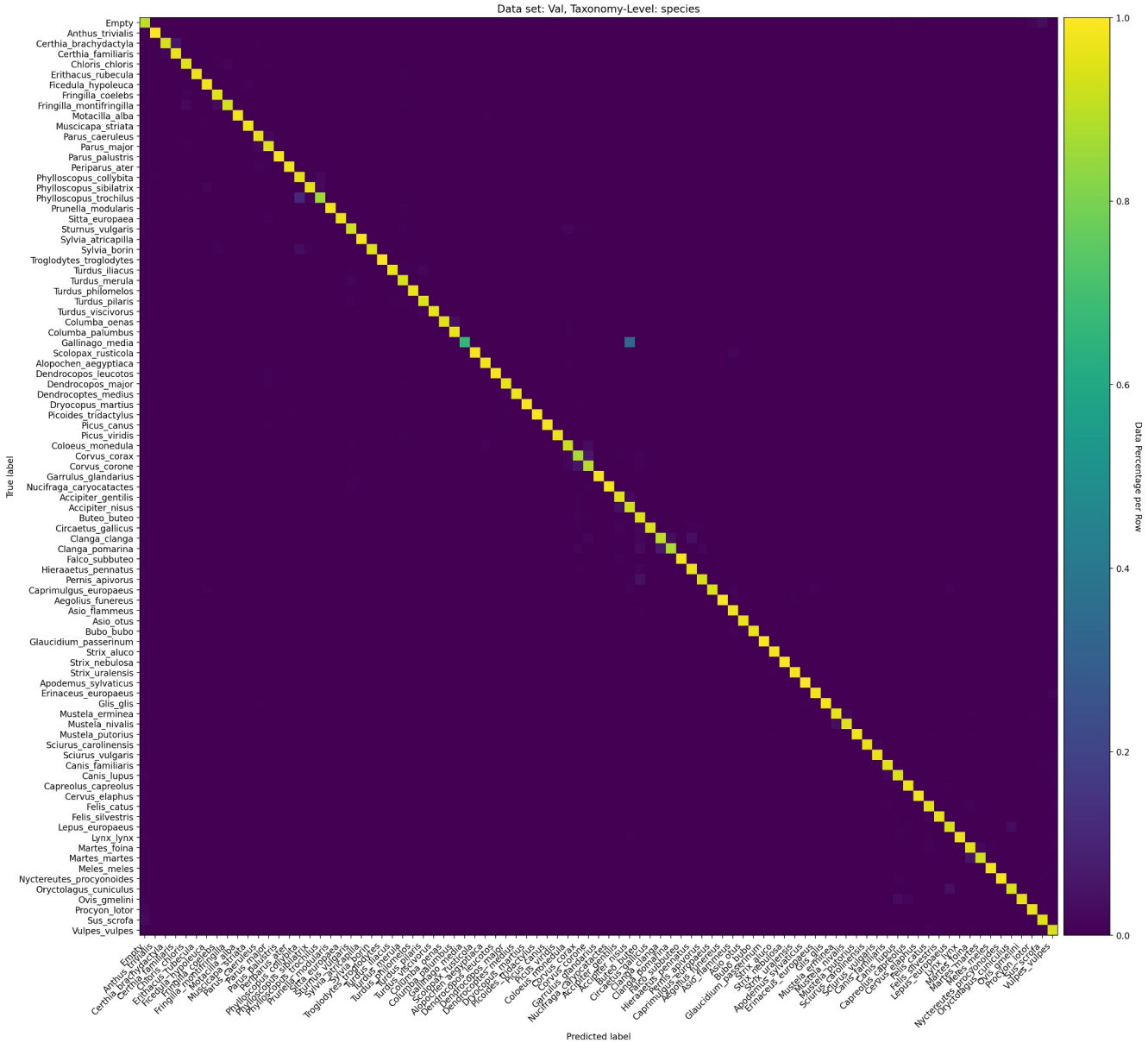

FIGURE F1 Averaged confusion matrix of the species classification ConvNeXt models on the validation data set.

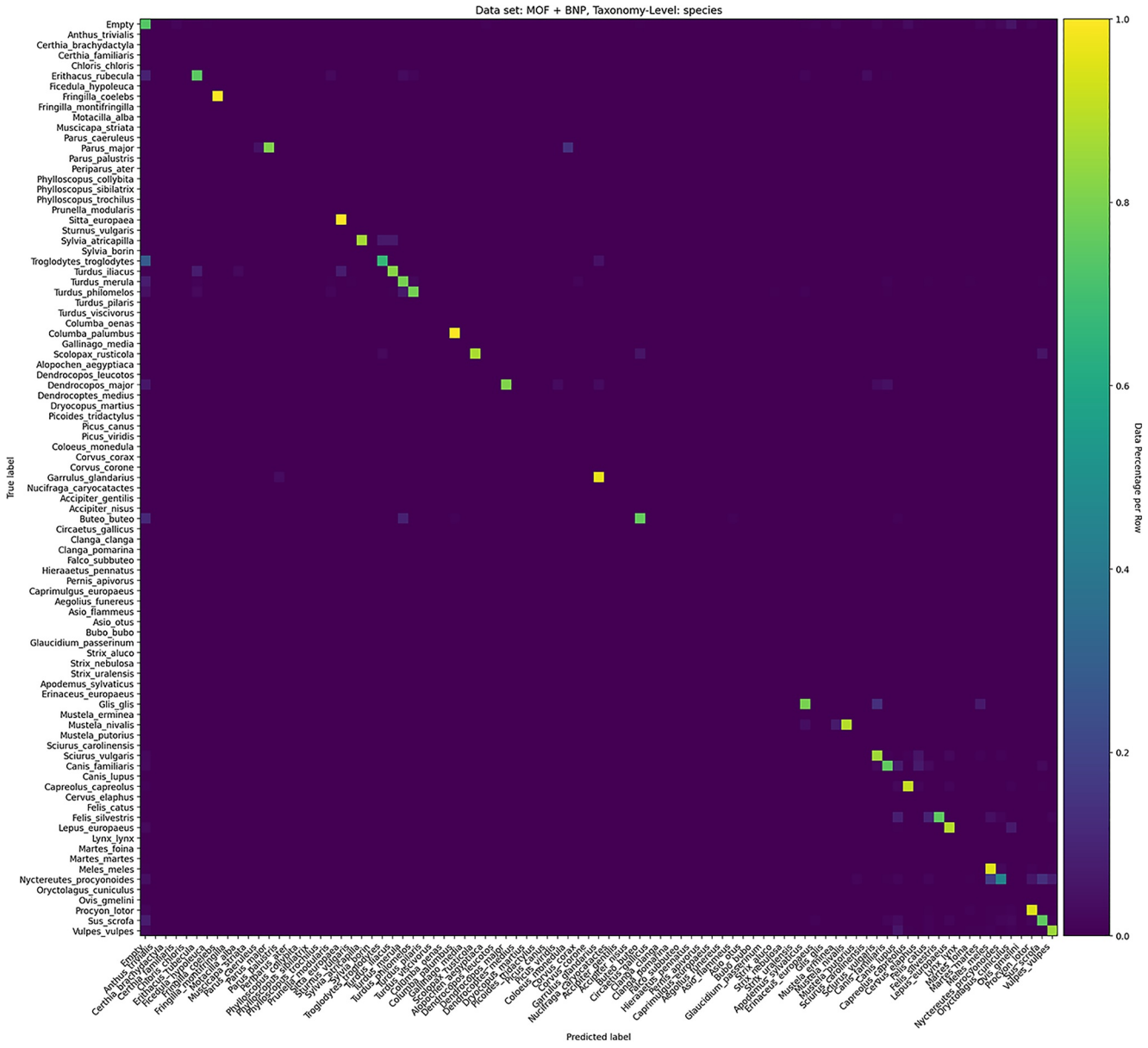

FIGURE F2 Averaged confusion matrix of the species classification ConvNeXt models on the test data sets.

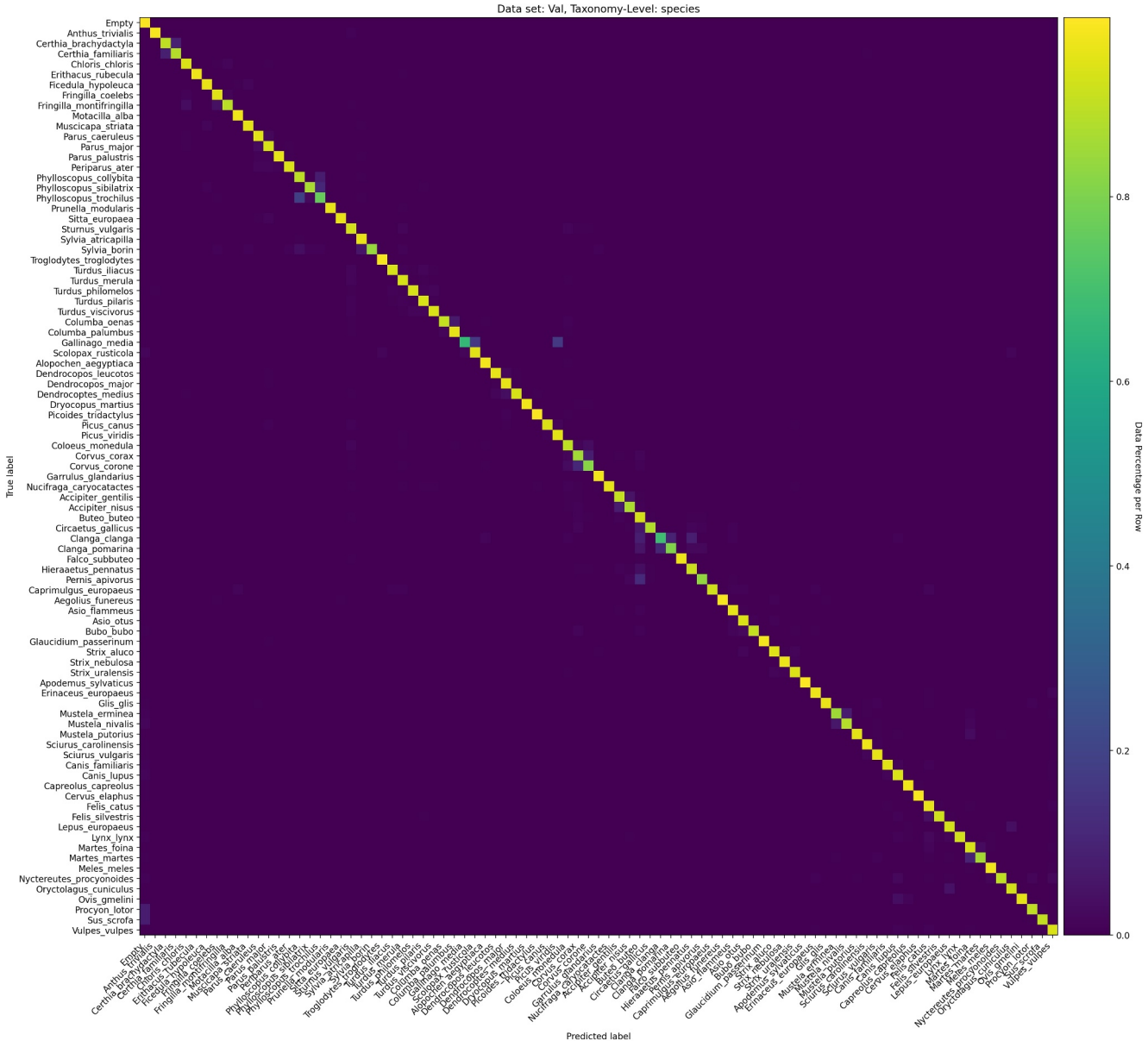

FIGURE F3 Averaged confusion matrix of the taxonomic classification ConvNeXt models on the validation data set.

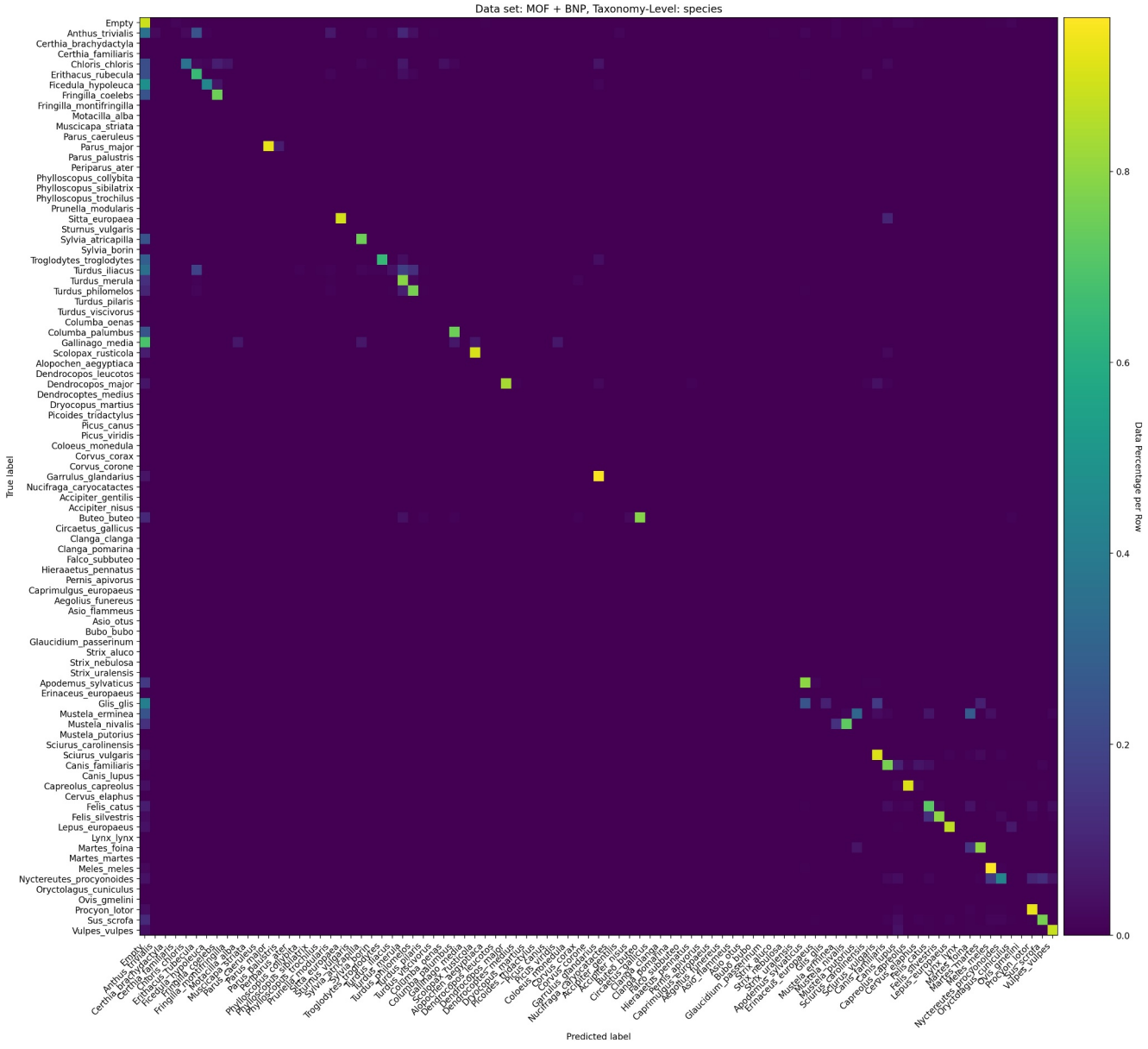

FIGURE F4 Averaged confusion matrix of the taxonomic classification ConvNeXt models on the test data sets.