Recognizing European mammals and birds in camera trap images using convolutional neural networks

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

[1] *Department of Mathematics & Computer Science, University of Marburg, Hans-Meerwein-Straße 6, D-35043 Marburg*
[2] *Department of Biology, University of Marburg, Karl-von-Frisch-Straße 8, D-35043 Marburg*
** E-mail: daniel.schneider@uni-marburg.de*

Abstract: A common way to study animal populations in the wild in an unobtrusive manner is using heat- or motion-activated cameras placed in natural habitats to automatically record images and/or videos. Manual analysis of the potentially large amounts of visual data obtained in this way is a time-consuming process, so automation through machine learning models trained on images and/or videos is desirable. Most visual animal recognition models are limited to mammal identification and group birds into a single class. Machine learning models for visually discriminating birds, in turn, cannot discriminate mammals and are also usually not designed for camera trap images. In this paper, we present convolutional neural network models based on the EfficientNetV2 and ConvNext architectures to recognize both mammals and bird species in camera trap images. Our ConvNextBase model achieves a mean average precision of 96.89% on our validation data set and a mean average precision of 93.88% on a test camera trap data set recorded in a forest in Hesse, Germany. This opens up a new way of automated bird monitoring besides the widely used method of bird call identification through audio recordings, which is limited to vocal bird species.

1 Introduction

Due to the sharp and ongoing worldwide decline of biodiversity over the last centuries [10, 18], there is an urgent need for a comprehensive monitoring of ecosystems such that early conservation measures can be taken if necessary. Automatic recorders can be deployed directly in the field to autonomously collect large amounts of data over long time spans with little to no human interference.

To monitor bird populations, microphones that continually record audio data are increasingly used [41]. Bird species occurring in the recording area can then be identified using automated bird call recognition methods[34]. However, not all bird species frequently vocalize, and vocal activity differs during the year [36]. Thus, monitoring approaches relying on audio recordings alone insufficiently cover seasonal variations in bird phenology, leaving out mass events such as migration or dispersal and the overwintering of populations.

To monitor mammal populations, camera traps are used, which were first introduced in 1956 [14] and have contributed greatly to wildlife ecology in recent decades [33, 42]. 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, the cameras are usually equipped with an infrared lens for nighttime images and a customary lens for daytime images. However, since the cameras are not only triggered by animals, but also by environmental influences, e.g., sunlight or motion caused by wind, some pictures are seemingly taken with no animals present.

Manual analysis of the automatically recorded huge amounts of data requires expertise and is therefore time-consuming and expensive. Thus, automation is desirable, which also ensures that the results are less biased by observers [39]. In recent years, machine learning methods, and particularly deep neural network models, have been used to analyze large amounts of data in the field of ecology.

Since then, several deep learning approaches for analyzing field microphone recordings and camera trap images have been explored and have yielded promising results (see section 2). However, most animal species recognition models are limited to mammal identification and group birds into a single class. Deep neural network models for visually identifying bird species, in turn, cannot discriminate mammals and are also usually not designed for processing camera trap images, but high quality bird photographs. Training deep neural networks that generalize well even under difficult circumstances requires large data sets of annotated images showing a sufficiently large number of all species to be recognized. Due to the huge amount of animal species occurring worldwide, it is hardly possible to cover all of them. Therefore, available deep neural network models (e.g., [8, 31, 45]) are usually limited to specific regions and small sets of species from each region for which sufficiently large sets of training images are available.

In this paper, we present convolutional neural network models that recognize the highly desirable combination of both mammal and bird species in camera trap images. In particular, our neural networks recognize 25 mammal and 63 bird species known to occur in Central European forests with a focus on our field study site in the Marburg Open Forest in Hesse, Germany. These include some species that are very difficult to distinguish visually, such as various marten species and closely related bird species. Our selection includes a range of species that is not covered in any available animal recognition model so far. Our main contributions are:

- We present a deep learning approach for recognizing European mammals and, for the first time, birds in camera trap images.
- To the best of our knowledge, we are the first to apply the EfficientNetV2 [47] and the ConvNext [29] neural network architectures to the task of camera trap image analysis.
- We make our trained models publicly available at `https://github.com/umr-ds/Marburg-Camera-Traps` to enable other researchers to build on our work.
- We publish our Marburg Open Forest test data set consisting of around 2500 camera trap images recorded in Hesse, Germany, in the same repository.

2 Related Work

2.1 Deep Neural Networks

In (supervised) deep learning, neural network models are trained to recognize desired patterns using large amounts of labeled data [27]. In the area of image processing, convolutional neural networks

(CNNs) [12] 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) [26] and ResNet (introduction of skip connections, which improves the training of deeper networks) [15].

A highly optimized type of CNN is EfficientNet [46]. 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 resources [47].

The Transformer model [48] does not use any convolutions, but instead relies on attention mechanisms that help the network to focus on the most relevant parts of the input. This architecture was initially used primarily in the field of natural language processing. Dosovitiskiy et al. [9] applied transformers to image analysis with the introduction of the VisionTransformer (ViT). Since then, several works have emerged in this field, which further optimize the underlying principle of the VisionTransformer and overcome some of its limitations, for instance the SwinTransformer [28].

Liu et al. [29] developed a CNN without attention mechanisms, called ConvNext, by adapting the block structure of a ResNet architecture to that of SwinTransformers and adopting other minor design adjustments from newer model architectures. Through these improvements, they have been able to outperform previous CNN and Transformer architectures and achieve new state of the art results.

In our work, we use the EfficientNetV2M [47] and ConvNextBase [29] models for recognizing animals in camera trap images.

2.2 *Automated Animal Classification*

The first work on automated animal classification dates back to 2013, where Yu et al. [51] performed sparse coding spatial pyramid matching (ScSPM) to extract relevant features from previously manually cropped images. Using these features, they trained a linear support vector machine (SVM) classifier on a (by today's standards) small data set of 7000 camera trap images with 18 animal species.

Deep neural networks were first used for animal species classification by Chen et al. [7] 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 [37] with millions of images and then fine-tuned on smaller camera trap data sets, a process called transfer learning. With the publication of the SnapshotSerengeti data set in 2015, which consists of 3.2 million images of 48 animal species [44], a data basis for training more complex animal recognition models was created.

2.2.1 Animal Recognition: In 2017, Gòmez et al. compared the animal recognition performance of CNN architectures of different sizes on a a 26-class subset of the SnapshotSerengeti data set. Deeper models like ResNet101 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 [49].

In 2018, Norouzzadeh et al. [31] trained 9 network architectures to not only classify animal species, but simultaneously count animals and determine other attributes such as behavior, a process known as multitask learning. They used two networks for this purpose: the first one performed a detection of the images containing animals, and the second one subsequently performed the analysis of these images. For training, the authors used the Snapshot Serengeti data set. They obtained the best results for the animal species classification with a ResNet-152, which achieved an accuracy of 93.8% on the test data.

In 2019, Tabak et al. [45] introduced the North American Camera Trap Images (NACTI) data set, which consists of over 3 million images of 27 species taken in North America and Canada [45]. They 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%.

The interest analyzing camera trap images has continued to grow since then, supported by competitions such as the iWildCam Challenge [5], an annual contest since 2018 that focuses on model generalization to new environments. In addition, companies launched initiatives to strengthen research in deep learning for ecology, e.g., 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. [21] trained a CNN model to recognize small reptiles like frogs, snakes and lizards found in Texas and deployed their model on an NVIDIA Jetson Nano edge device that is connected to the cameras. Jia et al. [23] performed neural architecture search on camera trap image data sets to find a lightweight CNN architecture for animal recognition that performs on par to other networks, but can run on edge devices.

Automated analysis of camera trap images works best when the models have been trained on images that are as similar as possible, ideally from the same location where they will later recognize animals. Auer et al. [2] proposed an active learning system to specialize their models on single camera traps with a small number of training images annotated by experts. They performed their experiments on a non-public data set from the Bavarian Forest National Park.

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.

2.2.2 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 localization, e.g., to count the number of animals or to track the position of the animals over a sequence of images.

In 2018, Schneider et al. [38] trained object detection models to localize animals in camera trap images. For this purpose, they labeled a subset of the Snapshot Serengeti data set with bounding boxes delimiting the positions of the animals.

In 2020, Carl et al. [6] applied an 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 classification. This shows that general-purpose models can suffice for camera trap image analysis in some cases.

Shepley et al. [40] trained models for pig detection on camera trap images in the same year, using camera trap data sets as well as images from the website 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 [4] as part of its AI for Earth program. The model was trained on a large number of camera trap images - some of which are not publicly available - and recognizes 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 object detection architecture.

In 2021, Norouzzadeh et al. [32] presented a system in which animal detection and animal species classification run separately. Using the MegaDetector, they localized 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.

Recently, Simões et al. [43] 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.

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

The most well known approach for automated bird species recognition is BirdNET [25], which is a CNN model trained on a large audio data set using extensive data pre-processing, augmentation, and mixup to achieve state-of-the-art performance. Mühling et al. [30] proposed a task-specific neural network created by neural architecture search, which won the BirdCLEF 2020 challenge [24]. It operates on raw audio data and contains multiple auxiliary heads and recurrent layers. In 2022, Höchst et al. [16] 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, allowing for real-time monitoring of an area.

There are far fewer publications on 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 [1, 13, 17]. The approaches that identify species are usually limited to a few bird species from a restricted geographic region. For example, Huang et al. [19] 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.[35] used a CNN to recognize 60 species found in the Asian sub-continent. Jacob et al. [22] presented an edge application for the visual recognition of 200 bird species. Ferreira et al. [11] trained a CNN model to recognize bird individuals from which they had previously collected training images using an automated method.

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 recognize both mammals and bird species in camera traps do not exist at all in the literature.

3 Methods

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

3.1 Animal Detection

First, we perform object detection using Microsoft's MegaDetector [4] to find the areas in the images showing animals. This also allows us to sort out the images where no animals are visible, i.e., where the camera was falsely triggered. The MegaDetector model has been 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, e.g., 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 [4]. It employs a YOLOv5† model that was fine-tuned on a large number of camera trap data sets. For each input image, MegaDetector returns a list of detected objects with bounding box coordinates, detected class (animal, human, or vehicle), and detection confidence score.

3.2 Species Classification

We 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

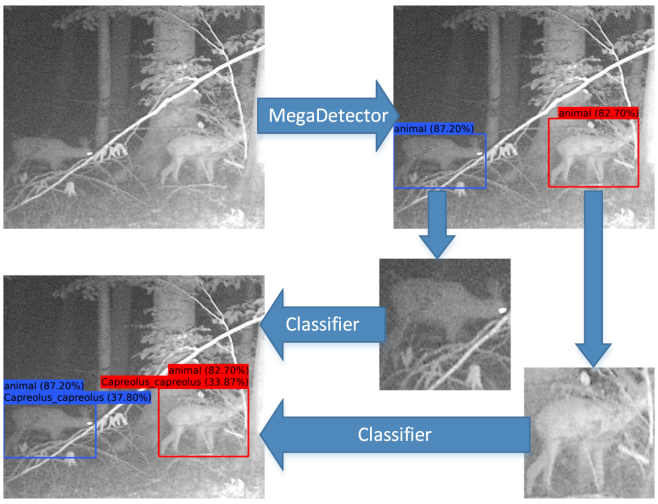

Fig. 1: Our two-step animal recognition pipeline

size of the classification model, i.e. 300 × 300 pixels. For classification, we use the EfficientNetV2M [47] (53M parameters) and the slightly larger ConvNextBase [29] (87M parameters) model.

We trained both models to recognize the 25 mammal and 63 bird species that we consider in our work. This includes many, sometimes closely related species that are hard to distinguish on camera images, especially under difficult viewing conditions, like 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 model also includes raptors like 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 our github repository. We also added an "empty" class, which brings the total number of classes to 89.

For each input image, the trained models provide a probability distribution over all possible classes, in which each class is assigned a confidence value that the image contains that class. The prediction class name is derived from the index of the highest confidence value.

3.3 Model Training

3.3.1 Challenges: Training a deep neural network model for classifying animal species in camera trap images presents several challenges [39]. We have investigated measures to overcome these as best as possible.

Amount of data: To train a neural network, a large amount of labelled data is necessary. If the amount of data is too small, this leads to overfitting on the training data, i.e., the network adapts to the training data set and does not generalize well to other data. We address 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 3.4). We also use 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, and adding Gaussian noise.

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 cause the network to adapt not only to the animal species, but also to similar image backgrounds. As a result the network performs

*https://xeno-canto.org
†https://github.com/ultralytics/yolov5

significantly worse on images from other 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 use the MegaDetector to determine bounding boxes for all data sets used. When training the classification model, we use these bounding boxes to crop the training images to the areas that show the animals. Since our model uses quadratic input images, we expand 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.

Data quality: Images from camera traps can have variable quality. In many cases, the animals are not well captured in the frame, 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 better or worse. Especially the infrared night shots are in some cases strongly overexposed or underexposed. This distinguishes camera traps from animal photographs, such as those found on websites like iNaturalist, in which an animal is captured in good visual conditions. We address this problem by adding higher quality photos to our camera trap data sets and sorting out images where the MegaDetector returns bounding boxes of very small size or with a confidence value below 0.6.

Species balance: Some animal species are significantly more abundant in the available data sets than others (in our case 24 examples minimum for Great snipe (*Gallinago media*), over 300,000 maximum for Red deer (*Cervus elaphus*)). This imbalance may cause the network to learn primarily the frequently occurring animal species and neglect the rare species. However, these rare species are usually of greater interest to ecologists, so it is important that they are correctly identified. To achieve this, we use a data generator that randomly samples an equal number of images of each species in each training epoch to ensure that the images of the rare species are repeated more often.

3.3.2 Training: Our neural network model consists of the following sequence of layers: The input layer contains $300 \times 300 \times 3$ neurons, here the input images (RGB images of size 300×300 pixels) 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 given selection to the input image. The augmented images are then fed into the backbone CNN model (EfficientNetV2M or ConvNextBase), which we initialize with weights pretrained on the ImageNet data set. We use global average pooling 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 regularization method to reduce overfitting. The final classification layer uses the softmax activation function to obtain a probability distribution over the confidence values of all possible species. We set the learning rate to $5 * 10^{-5}$ at the start of the training and reduce it by a factor of 0.2 if the validation loss has not decreased for 10 epochs. We train our models for 100 epochs. As our optimizer, we use AdamW with a weight decay of 0.05.

3.4 Data Sets

Our neural network model focuses on a selection of European mammal and bird species that do not occur in this composition in any data set available online. Therefore, we combine data from a variety of data sets. 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 respective native species. For European species, the available data is much more limited. Some of the species we aim to recognize also occur in North America, so we can draw on data sets recorded there. We use images from the data sets Caltech Camera Traps [3], ENA24-detection [50], Idaho Camera Traps*, Missouri Camera Traps [52] and North American Camera Trap Images (NACTI) [45]. Additionally, we use images from the WCS Camera Traps data set, a collection of images from 12 countries created by the Wildlife Conservation Society†.

We also use 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) [20], the other one is the Tierschnappschuss data set‡.

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

We supplemented the images from the various data sets with crawled images to fill out the remaining gaps and increase the image diversity. We downloaded images of our desired species from the eMammal camera trap data management system‖, crawled photographs of live animals with verified captions from the website iNaturalist** 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 annotated at image or image sequence level. In this case, we applied the MegaDetector to obtain bounding boxes for the images and sort out empty images. A few data sets contain images with no animals, which are labeled as empty. We use the bounding boxes detected on these images as 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 species and used the first 20% (maximum 500) of the images of each species as validation data and the rest as training data. Thus, we divided the total 1,273,379 bounding boxes into 1,226,158 training and 47,221 validation boxes.

For further evaluation, we used a data set recorded in the target area of our studies. We deployed camera traps in the Marburg Open Forest (MOF)†† in Hesse, Germany in the first half of 2021. 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. We compared our models' predictions to these labels. Table 1 gives an overview over all data sets we used in this work. See our github repository for a more detailed overview of the number of bounding boxes per species and the split between training and validation data.

4 Experimental Evaluation

We conducted extensive experiments to evaluate the quality of our approach. We restrict ourselves to evaluating our trained classification models and use the MegaDetector as published by Microsoft.

4.1 Metrics

To determine the quality of the overall model, we use the Accuracy metric, defined as:

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

where $imgs = \{i_1, i_2, ...i_N\}$ is a list of analyzed images and $correkt_k \subseteq imgs$ are the images where the correct class is found

*https://lila.science/datasets/idaho-camera-traps
†https://lila.science/datasets/wcscameratraps
‡https://emammal.si.edu/tierschnappschuss
§https://www.kaggle.com/datasets/gpiosenka/100-bird-species
¶https://dl.allaboutbirds.org/nabirds
‖https://emammal.si.edu
**https://inaturalist.org
††https://www.uni-marburg.de/de/fb19/fachbereich/infrastruktur/mof

Table 1 Overview over all data sets used in our experiments

Data Set Name	Recording Locations	Species Total	Images Total	Relevant Species	Used Boxes	Training Boxes	Validation/ Test Boxes
Caltech Camera Traps [3]	Southwestern USA	21	∼240,000	3	10,915	9,882	1,033
ENA24-detection [50]	Eastern North America	23	∼10,000	5	2,232	1,788	444
Idaho Camera Traps	USA	62	∼1.5M	2	75,699	74,699	1000
Missouri Camera Traps [52]	USA	21	∼25,000	8	11,517	9,768	1,749
North American Camera Traps [45]	USA	28	∼3.7M	9	584,436	581,157	3,279
WCS Camera Traps	Worldwide (12 countries)	675	∼1.4M	7	10,820	9,805	1,015
Rodent [20]	Germany	41	∼14,000	25	14,327	12,645	1,682
Tierschnappschuss	Southern Germany	41	∼170,000	18	140,635	135,675	4,960
Kaggle Birds	Worldwide (Internet Searches)	525	∼90,000	6	1,104	886	218
North American Birds	North America	400	∼48,000	2	429	344	85
eMammal	Worldwide (Crawled Subset)			16	13,485	11,140	2,345
InatCrawl	Worldwide (Crawled Subset)			88	406,366	377,234	29,132
WebCrawl	Worldwide (Crawled Subset)			7	1,414	1,135	279
Marburg Open Forest	Germany			19	2,420	0	2,420

in the top k predictions. As is common in the evaluation of multi-class classification models, 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 class 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. Therefore, we additionally consider the mean Class-wise Accuracy (mCA), which we calculate as the average of the Top 1 Accuracies for each class.

We also use mean average precision (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 average precision (AP) for each class $c \in C$, where C is the list of all animal species to be recognized, as follows:

$$
\begin{aligned}
AP_c &= \frac{1}{|relevant_c|} \sum_{k=1}^{|imgs|} prec@k * rel@k \\
\text{with} \quad rec@k &= \frac{|relevant_c \cap retrieved_k|}{|retrieved_k|} \qquad (2) \\
\text{and} \quad rel@k &= \begin{cases} 1 & \text{if } i_k \in relevant_c \\ 0 & \text{otherwise} \end{cases},
\end{aligned}
$$

where $imgs = \{i_1, i_2, ...i_N\}$ is a list of analyzed images ranked by the prediction score for the class c. $relevant_c \subseteq imgs$ denotes the relevant images for the class c, i.e., the images containing the animal c and $retrieved_k = \{i_1, i_2, ..., i_k\}$ 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 species.

4.2 Results

4.2.1 Model Comparison: We compare the results of our EfficientNetV2 model with those of our ConvNext model first on a validation set consisting of images withheld from our training data, and second on the Marburg Open Forest (MOF) test data set we recorded. Table 2 shows the calculated metrics.

The ConvNext model performs slightly better than the EfficientNet model on the validation data (mAP +0.47). On the MOF test data, the gap is a bit larger (mAP +1.14). It is also noticeable that both models achieve very high values in the Top 5 Accuracy, while the Top 1 Accuracy is significantly lower, especially on the test data. Both models reach almost equally high mCA values. Class-wise metrics did not show a statistically significant correlation between the recognition accuracy of a class and the number of training data available for that class. A closer analysis of the predictions 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 (*Phylloscopus trochilus*) and Common chiffchaff (*Phylloscopus collybita*), Greater spotted eagle (*Clanga clanga*) and Lesser spotted eagle (*Clanga pomarina*), House cat (*Felis catus*) and Wild cat (*Felis silvestris*), European rabbit (*Oryctolagus cuniculus*) and European hare (*Lepus europaeus*) as well as Stoat (*Mustela erminea*) and Least weasel (*Mustela nivalis*). In these cases, however, the correct species is very often listed in the top 5 results.

Table 2 Results of the different model types

Metric	EfficientNetV2M	ConvNextBase
Val Accuracy	91.84	93.03
Val Top3 Accuracy	97.42	97.78
Val Top5 Accuracy	98.37	98.62
Val mCA	99.81	99.84
Val mAP	96.42	96.89
Test Accuracy	84.42	84.24
Test Top3 Accuracy	94.07	94.44
Test Top5 Accuracy	96.65	97.20
Test mCA	99.65	99.65
Test mAP	92.74	93.88

Table 3 Results for training on different species sets

Metric	Trained on	Birds	Mammals	Both
Val mAP Birds		97.58	-	97.30
Val mAP Mammals		-	97.58	97.02
Test mAP Birds		99.23	-	98.33
Test mAP Mammals		-	96.26	95.18

4.2.2 Species Comparison: We have presented models that can recognize both mammals and bird species. For comparison, we have trained one model to classify only mammals and one to classify only birds. We compare the results of these models to a model that can recognize both types of animals in Table 3. The compared models all use the ConvNextBase architecture and are trained with all measures described in the next section.

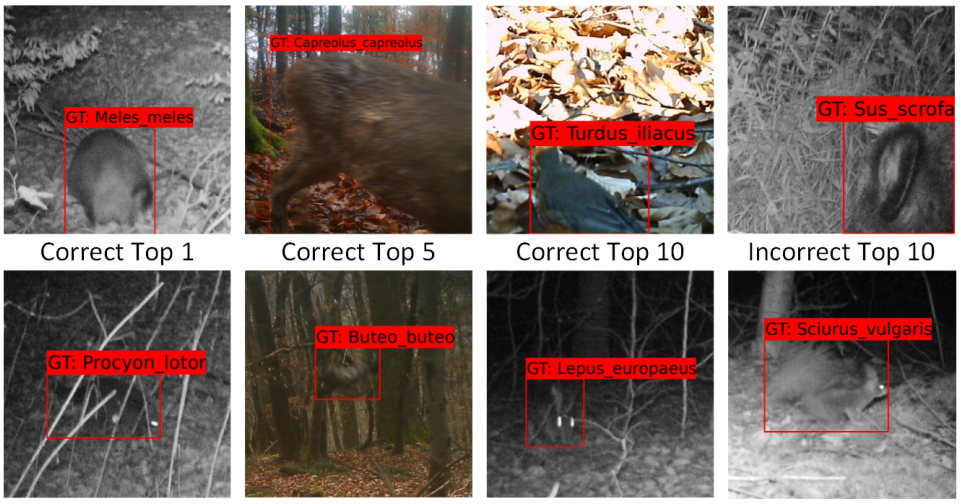

Fig. 2: A selection of images from the MOF data set on which our models had increasing difficulties predicting the correct species

Table 4 Results without different training measures

Metric	All measures	noAug	noCrop	noFilt	noSamp
Val mAP	96.89	95.71	96.08	86.13	95.67
Test mAP	93.88	93.86	75.60	83.46	95.03

The neural network models perform very well for pure bird recognition, however, only a small part of the bird species to be recognized is represented in the evaluation on the test data. The validation data largely consist of images in which the birds are very well recognizable. The recognition of the mammals works slightly worse, which is due to the fact that the camera trap images primarily used here often show the animals much less clearly. The models that identify both bird and mammal species perform slightly worse than the individual models (test mAP -0.90 to -1.08), but still reach very good results, which shows that combined recognition is successfully possible.

4.2.3 Measure Comparison: We have taken various measures to overcome the challenges described in section 3.3.1, namely using data augmentation, cropping the images to the areas where animals were detected, sorting out images with low detection confidence, and sampling the training images in each epoch to an approximately equal number of images of each species. We now investigate the influence of these measures on the overall performance. As a baseline, we trained a model where we applied all measures and compare it to models where we omit one of the measures each time. Again, all compared models use the ConvNextBase architecture.

As Table 4 shows, omitting any of these measures leads to a decrease in model performance. On the test data set a very small degredation is achieved by omitting data augmentation (noAug: test mAP -0.02), which shows that although the training data we used is quite large, increasing its variability via augmentation is still beneficial to some degree. Not filtering out detections with small confidences and small bounding boxes leeds to a much stronger reduction (noFilt: mAP -10.42). The 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. Not cropping the images to the areas found by the detection model causes an even more significant performance decrease (noCrop: mAP -18.28). For comparability, we used only the images for training where the MegaDetector had detected an animal with the minimum confidence. Omitting sampling during training lead to a decrease in performance only on the validation data. On the test data set, the performance actually slightly increased (noSamp: mAP +1.15), which we did not expect. We attribute this to the fact that our test data set contains mainly species for which sufficient training material was available. The sampling mainly helps the model to recognize the less frequent classes better. We therefore still use this measure in our final model.

4.2.4 Error Analysis: Finally, we examine some images that were challenging for the analysis by our models. In Figure 2, from left to right, two images each are shown for which our best model provided a correct Top1, Top5, Top10 classification or no correct classification at all. Many of the instances that are difficult to identify, even for humans, are correctly identified by the model. The false recognitions 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 distinguishing 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, e.g., the squirrel in the bottom right image. This shows that the models do not work perfectly and need to be better adapted to some cases, e.g., with more training data from the same domain.

5 Conclusion

We presented a deep learning approach for analyzing camera trap images that can distinguish not only mammals but also bird species. This can help researchers to analyze the large amounts of data generated when using camera traps to observe wildlife. We first localized the animals using Microsoft's MegaDetector and then used our trained EfficientNet and ConvNext models to determine the species. Our best classification model of the ConvNextBase architecture achieved a mAP of 96.89% % on a validation data set left out from our training data and a mAP of 93.88%. on a test data set we recorded in Hesse, Germany. 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. Since improvements would mainly be achieved by larger and more diverse amounts of training data, we plan to investigate methods to synthetically generate new or combine existing images. Furthermore, we will explore to what extent the images of the species not studied from the already used data sets can nevertheless be used for training. Finally, we will investigate whether there are better strategies to balance the amount of data per species during training that lead to higher performance improvements for both frequent and rare classes.

6 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).

7 References

1 Hüseyin Gökhan Akçay, Bekir Kabasakal, Duygugül Aksu, Nusret Demir, Melih Öz, and Ali Erdoğan. Automated bird counting with deep learning for regional bird distribution mapping. *Animals*, 10:1–24, 7 2020. doi: 10.3390/ani10071207.

2 Daphne Auer, Paul Bodesheim, Christian Fiderer, Marco Heurich, and Joachim Denzler. Minimizing the annotation effort for detecting wildlife in camera trap images with active learning. INFORMATIK, 2021.

3 Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *15th Eur. Conf. on Computer Vision (ECCV), Munich, Germany*, pages 472–489, 2018. doi: 10.1007/978-3-030-01270-0_28.

4 Sara Beery, Dan Morris, Siyu Yang, Marcel Simon, Arash Norouzzadeh, and Neel Joshi. Efficient pipeline for automating species id in new camera trap projects. *Biodiversity Inf. Science and Standards*, 3, 2019. doi: 10.3897/biss.3.37222.

5 Sara Beery, Grant van Horn, Oisin Mac Aodha, and Pietro Perona. The iwildcam 2018 challenge dataset. *arXiv preprint arXiv:1904.05986*, 4 2019.

6 Christin Carl, Fiona Schönfeld, Ingolf Profft, Alisa Klamm, and Dirk Landgraf. Automated detection of european wild mammal species in camera trap images with an existing and pre-trained computer vision model. *European Journal of Wildlife Research*, 66(4):1–7, 2020. doi: 10.1007/s10344-020-01404-y/Published.

7 Guobin Chen, Tony X Han, Zhihai He, Roland Kays, and Tavis Forrester. Deep convolutional neural network based species recognition for wild animal monitoring. In *IEEE Int. Conf. on Image Processing (ICIP)*, pages 858–862, 2014.

8 Mateusz Choiński, Mateusz Rogowski, Piotr Tynecki, Dries P. J. Kuijper, Marcin Churski, and Jakub W. Bubnicki. A first step towards automated species recognition from camera trap images of mammals using ai in a european temperate forest. In *Computer Information Systems and Industrial Management*, pages 299–310. Springer, 2021. ISBN 978-3-030-84340-3.

9 Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th Int. Conf. on Learning Representations (ICLR), Austria*. OpenReview.net, 2021.

10 Sandra Myrna Díaz 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. doi: 10.5281/zenodo.3553579.

11 André C. Ferreira, Liliana R. Silva, Francesco Renna, Hanja B. Brandl, Julien P. Renoult, Damien R. Farine, Rita Covas, and Claire Doutrelant. Deep learning-based methods for individual recognition in small birds. *Methods in Ecology and Evolution*, 11:1072–1085, 9 2020. doi: 10.1111/2041-210X.13436.

12 Kunihiko Fukushima. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological cybernetics*, 36(4):193–202, 1980.

13 Dawid Gradolewski, Damian Dziak, Milosz Martynow, Damian Kaniecki, Aleksandra Szurlej-Kielanska, Adam Jaworski, and Wlodek J. Kulesza. Comprehensive bird preservation at wind farms. *Sensors (Switzerland)*, 21:1–35, 1 2021. doi: 10.3390/s21010267.

14 Leslie W Gysel and Earle M Davis. A simple automatic photographic unit for wildlife research. *The Journal of Wildlife Management*, 20:451–453, 1956.

15 Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

16 Jonas Höchst, Hicham Bellafkir, Patrick Lampe, Markus Vogelbacher, Markus Mühling, Daniel Schneider, Kim Lindner, Sascha Rösner, Dana G Schabo, Nina Farwig, et al. Bird@edge: Bird species recognition at the edge. In *10th Int. Conf. on Networked Systems (NETYS), Virtual, May 17–19*, pages 69–86. Springer, 2022.

17 Suk Ju Hong, Yunhyeok Han, Sang Yeon Kim, Ah Yeong Lee, and Ghiseok Kim. Application of deep-learning methods to bird detection using unmanned aerial vehicle imagery. *Sensors*, 19, 4 2019. doi: 10.3390/s19071651.

18 David U. Hooper, E. Carol Adair, Bradley J. Cardinale, Jarrett E.K. Byrnes, Bruce A. Hungate, Kristin L. Matulich, Andrew Gonzalez, J. Emmett Duffy, Lars Gamfeldt, and Mary I. Connor. A global synthesis reveals biodiversity loss as a major driver of ecosystem change. *Nature*, 486:105–108, 6 2012. doi: 10.1038/nature11118.

19 Yo Ping Huang and Haobijam Basanta. Bird image retrieval and recognition using a deep learning platform. *IEEE Access*, 7:66980–66989, 2019. doi: 10.1109/ACCESS.2019.2918274.

20 C. Imholt, J. Jacob, A. Schlötelburg, and A. Geduhn. *Langfristige Populationsentwicklung krankheitsübertragender Nagetiere: Interaktion von Klimawandel, Landnutzung und Biodiversität : Abschlussbericht*. Climate change. Umweltbundesamt Deutschland, 2021.

21 Sazida B. Islam and Damian Valles. Identification of wild species in texas from camera-trap images using deep neural network for conservation monitoring. *2020 10th Annual Computing and Communication Workshop and Conference, CCWC 2020*, pages 537–542, 2020. doi: 10.1109/CCWC47524.2020.9031190.

22 I Jeena Jacob and P Ebby Darney. Design of deep learning algorithm for iot application by image based recognition. *Journal of ISMAC*, 3:276–290, 8 2021. doi: 10.36548/jismac.2021.3.008.

23 Liang Jia, Ye Tian, and Junguo Zhang. Domain-aware neural architecture search for classifying animals in camera trap images. *Animals*, 12, 2 2022. doi: 10.3390/ani12040437.

24 Stefan Kahl, Mary Clapp, W. Alexander Hopping, Hervé Goëau, Hervé Glotin, Robert Planqué, Willem-Pier Vellinga, and Alexis Joly. Overview of birdclef 2020: Bird sound recognition in complex acoustic environments. In *Conf. and Labs of the Evaluation Forum (CLEF), Thessaloniki*, volume 2696. CEUR-WS.org, 2020.

25 Stefan Kahl, Connor M. Wood, Maximilian Eibl, and Holger Klinck. Birdnet: A deep learning solution for avian diversity monitoring. *Ecological Informatics*, 61, 2021. doi: 10.1016/j.ecoinf.2021.101236.

26 Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012.

27 Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521: 436–444, 5 2015. doi: 10.1038/nature14539.

28 Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *IEEE/CVF Int. Conf. on Computer Vision*, pages 10012–10022, 2021.

29 Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 11976–11986, 2022.

30 Markus Mühling, Jakob Franz, Nikolaus Korfhage, and Bernd Freisleben. Bird species recognition via neural architecture search. In *Conf. and Labs of the Evaluation Forum (CLEF), Thessaloniki*, volume 2696. CEUR-WS.org, 2020.

31 Mohammad Sadegh Norouzzadeh, Anh Nguyen, Margaret Kosmala, Alexandra Swanson, Meredith S. Palmer, Craig Packer, and Jeff Clune. Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning. *Proceedings of the National Academy of Sciences*, 115, 6 2018. doi: 10.1073/pnas.1719367115.

32 Mohammad Sadegh Norouzzadeh, Dan Morris, Sara Beery, Neel Joshi, Nebojsa Jojic, and Jeff Clune. A deep active learning system for species identification and counting in camera trap images. *Methods in Ecology and Evolution*, 12:150–161, 1 2021. doi: 10.1111/2041-210X.13504.

33 Allan F O'Connell, James D Nichols, and K Ullas Karanth. *Camera traps in animal ecology: methods and analyses*, volume 271. Springer, 2011.

34 Nirosha Priyadarshani, Stephen Marsland, and Isabel Castro. Automated birdsong recognition in complex acoustic environments: a review. *Journal of Avian Biology*, 49(5), 2018. doi: 10.1111/jav.01447.

35 Satyam Raj, Saiaditya Garyali, Sanu Kumar, B E Scholar, and Sushila Shidnal. Image based bird species identification using convolutional neural network. *International Journal of Engineering Research & Technology (IJERT)*, 9, 2020.

36 Samuel R. P.-J. Ross, Darren P. O'Connell, Jessica L. Deichmann, Camille Desjonquères, Amandine Gasc, Jennifer N. Phillips, Sarab S. Sethi, Connor M. Wood, and Zuzana Burivalova. Passive acoustic monitoring provides a fresh perspective on fundamental ecological questions. *Functional Ecology*, 37(4):959–975, 2023. doi: 10.1111/1365-2435.14275.

37 Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015.

38 Stefan Schneider, Graham W. Taylor, and Stefan C. Kremer. Deep learning object detection methods for ecological camera trap data. In *15th Conf. on Computer and Robot Vision (CRV)*, pages 321–328, 2018. doi: 10.1109/CRV.2018.00052.

39 Stefan Schneider, Saul Greenberg, Graham W. Taylor, and Stefan C. Kremer. Three critical factors affecting automated image species recognition performance for camera traps. *Ecology and Evolution*, 10, 4 2020. doi: 10.1002/ece3.6147.

40 Andrew Shepley, Greg Falzon, Paul Meek, and Paul Kwan. Location invariant animal recognition using mixed source datasets and deep learning. *bioRxiv*, 2020. doi: 10.1101/2020.05.13.094896.

41 Julia Shonfield and Erin Bayne. Autonomous recording units in avian ecological research: current use and future applications. *Avian Conservation and Ecology*, 12 (1), 5 2017. doi: 10.5751/ACE-00974-120114.

42 Leandro Silveira, Anah T.A. Jácomo, and José Alexandre F. Diniz-Filho. Camera trap, line transect census and track surveys: A comparative evaluation. *Biol. Conserv.*, 114:351–355, 2003. doi: 10.1016/S0006-3207(03)00063-6.

43 Fanny Simões, Charles Bouveyron, and Frédéric Precioso. Deepwild: Wildlife identification, localisation and estimation on camera trap videos using deep learning. *Ecological Informatics*, 75, 2023. doi: 10.1016/j.ecoinf.2023.102095Ãf.

44 Alexandra Swanson, Margaret Kosmala, Chris Lintott, Robert Simpson, Arfon Smith, and Craig Packer. Snapshot serengeti, high-frequency annotated camera trap images of 40 mammalian species in an african savanna. *Scientific Data*, 2, 6 2015. doi: 10.1038/sdata.2015.26.

45 Michael A Tabak, Mohammad S Norouzzadeh, David W Wolfson, Steven J Sweeney, Kurt C VerCauteren, Nathan P Snow, Joseph M Halseth, Paul A Di Salvo, Jesse S Lewis, Michael D White, et al. Machine learning to classify animal species in camera trap images: Applications in ecology. *Methods in Ecology and Evolution*, 10:585–590, 4 2019. doi: 10.1111/2041-210X.13120.

46 Mingxing Tan and Quoc Le. EfficientNet: Rethinking model scaling for convolutional neural networks. In *36th Int. Conf. on Machine Learning (ICML)*, volume 97, pages 6105–6114. PMLR, 09–15 Jun 2019.

47 Mingxing Tan and Quoc Le. Efficientnetv2: Smaller models and faster training. In *Int. Conf. on Machine Learning (ICML)*, pages 10096–10106. PMLR, 2021.

48 Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30:5999–6009, 6 2017.

49 Alexander Gomez Villa, Augusto Salazar, and Francisco Vargas. Towards automatic wild animal monitoring: Identification of animal species in camera-trap images using very deep convolutional neural networks. *Ecological Informatics*, 41:24–32, 9 2017. doi: 10.1016/j.ecoinf.2017.07.004.

50 Hayder Yousif, Roland Kays, and Zhihai He. Dynamic programming selection of object proposals for sequence-level animal species classification in the wild. *IEEE Transactions on Circuits and Systems for Video Technology*, 2019.

51 Xiaoyuan Yu, Jiangping Wang, Roland Kays, Patrick A. Jansen, Tianjiang Wang, and Thomas Huang. Automated identification of animal species in camera trap images. *Eurasip Journal on Image and Video Processing*, 2013. doi: 10.1186/1687-5281-2013-52.

52 Zhi Zhang, Zhihai He, Guitao Cao, and Wenming Cao. Animal detection from highly cluttered natural scenes using spatiotemporal object region proposals and patch verification. *IEEE Transactions on Multimedia*, 18(10):2079–2092, 2016.