Improving Wildlife Management with AI: Species Detection and Classification from Camera Trap Data

Sead Mustafić[1,*], Dominik Dachs[2], Rainer Prüller[3], Florian Schöggl[3] and Roland Perko[1]

[1]*JOANNEUM RESEARCH, Graz, Austria*

[2]*Meles Wildbiologie, Großraming, Austria*

[3]*Pentamap GmbH, Graz, Austria*

Abstract

In this study, we explore advanced computer vision techniques to enhance wildlife management through the automatic detection and classification of animal species from camera trap images. Leveraging deep learning methods, our research focuses on the automated extraction of critical information from these images to support forest and wildlife management, biodiversity monitoring, and reintroduction program evaluations. We present a specialized data set with manually labeled and validated images and comprehensive metadata, including species identification, sex, age class, and unique IDs for individual animals. Our approach integrates both single-stage and two-stage detection and classification strategies, utilizing models such as YOLO and EfficientNet. Initial results demonstrate the effectiveness of our methods, achieving significant accuracies (up to 95%) and providing a user-friendly interface for further refinement of classifications. Future work will expand the data set and explore transformer-based deep neural networks to enhance the robustness and applicability of our wildlife classification system.

Keywords

Wildlife detection, wildlife classification, camera traps, computer vision, artificial intelligence

1. Introduction

Wildlife cameras have become essential tools in contemporary wildlife research and conservation efforts, being employed in a diverse range of applications. Their primary use involves monitoring wildlife populations, which includes both common game species such as red deer, roe deer, and wild boar, and more elusive or ecologically important species that play crucial roles in maintaining biodiversity [1, 2]. These cameras offer a non-intrusive means of gathering vital data, capturing images of animals in their natural habitats without interfering. Figure 1 shows a camera trap mounted on a tree, demonstrating the typical setup used to capture images of wildlife in their natural environments. This setup ensures minimal disturbance to the animals while providing a strategic view point for monitoring. Figure 2 illustrates a typical image captured by such a camera trap, showcasing its capability to provide detailed visual information.

Figure 1: Exemplary trail camera with a SIM card for near real-time data transmission and an infrared flash for night recordings, mounted on a tree to monitor forest activity.
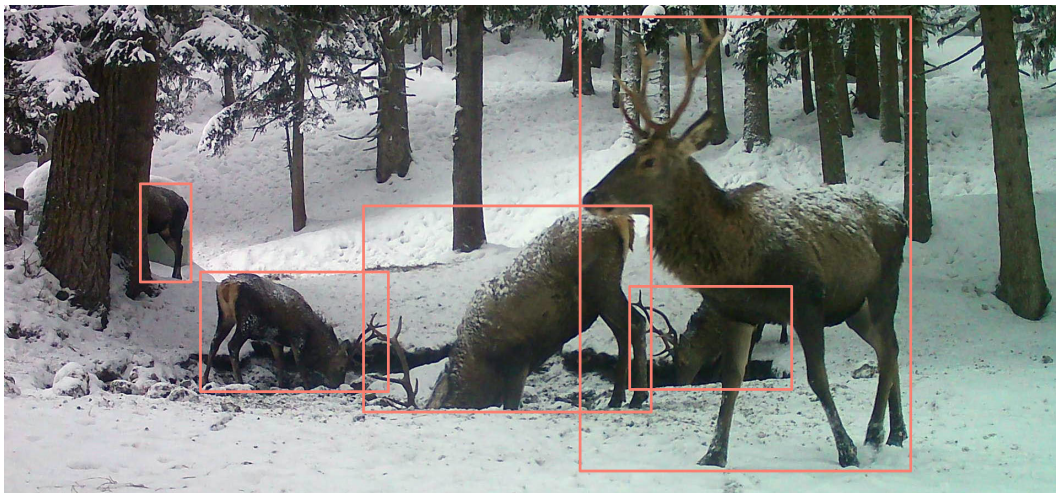
Figure 2: Exemplary labeled image from a red deer (*Cervus elaphus*) feedings site.

In this paper, we present a comprehensive investigation into the use of advanced computer vision techniques to assist biologists in classifying wildlife based on images obtained from camera traps. Our research leverages state-of-the-art machine learning methods to accurately identify and categorize various animal species, focusing on the automatic derivation of information from these images. The key stakeholders and beneficiaries of our approaches include:

1. Forest and wildlife management: Forest administrations, hunting authorities, and avalanche control agencies can benefit from our approaches for effective planning and implementation of forest protection strategies. The automatic information extraction from camera trap images aids in making informed decisions regarding wildlife management and habitat conservation.

2. Biodiversity monitoring: Ecologists, nature conservationists, and national park authorities can utilize our techniques to conduct comprehensive assessments of biodiversity. By automating the identification of species diversity and population dynamics, our research supports understanding the health of ecosystems and aids in the development of conservation policies.

3. Reintroduction program evaluation: Nature conservationists and national park administrators can use our automated methods to gain insights into the success of species reintroduction efforts. For example, tracking the reintroduction of predators such as wolves and lynxes requires precise monitoring, which our research facilitates by providing accurate identification and tracking of these species over time.

This study focuses on the initial and critical task of detecting and classifying all animal species present in the captured images. This step is essential for further advancements, where our approach will also identify individual animals and determine specific attributes such as sex and age class. This enhanced capability will enable more detailed ecological studies and management practices. Additionally, it is important to note that our method, similar to most camera trap-based computer vision techniques, also detects humans and vehicles. This aspect is crucial as it addresses privacy concerns by ensuring that images containing humans or vehicles (license plates) are promptly identified and deleted to protect individual privacy.

1.1. State of the Art

Various computer vision systems for wildlife detection and classification from camera trap images have been proposed (e.g., [3]) and evaluated (e.g., [4]). In this respect, Microsoft developed a specific detector for camera trap images called MegaDetector[1] [5], which yields bounding boxes for humans, vehicles, and animals and, thus, also classifies blank images. Up to version 4, it was based on Faster R-CNN [6], and from version 5 onwards, it utilized YOLOv5[2].

To achieve finer-grained classification of species, two methodologies are pursued: (1) training object detectors and classifiers specifically for this task, or (2) relying on a general detector for animals (like MegaDetector) and classifying the corresponding bounding boxes. The first approach is mostly based on one or two-stage detectors like Fast R-CNN [7], Faster R-CNN [6], EfficientDet [8], and various YOLO versions [9] (including YOLOv1 to YOLOv8, YOLOR, and YOLOX). The second approach is based on classifiers like VGG [10], Inception-v3 [11], Xception [12], or EfficientNet [13].

An example of an one-stage approach is presented within Trapper-AI[3]. It is based on YOLOv8[4] and was trained for 18 European mammal species. An example of a two-stage approach is DeepFaune [14], which can distinguish between 28 different species. It was trained on proprietary data from France and achieves relatively high accuracies on this data set. Additionally, Swarovski's AI-based binoculars AX Visio[5] represent a significant advancement in real-time wildlife observation, capable of distinguishing 9000 different bird species.

---

[1]https://github.com/microsoft/CameraTraps/blob/main/megadetector.md

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

[3]https://huggingface.co/OSCF/TrapperAI-v02.2024

[4]https://github.com/ultralytics/ultralytics

[5]https://www.swarovskioptik.com/us/en/birding/products/binoculars/ax-visio/ax-visio-binoculars/ax-visio

Several studies have further demonstrated the potential of deep learning for the detection and classification of animals from camera trap images [15, 16, 17, 18]. In most cases, these studies achieved higher accuracy in detecting the region of the animal compared to fine-grained classification [16, 18]. Additionally, research indicates that transfer learning, where a model trained on data from one region is applied to another, can be effective [16, 17]. While detection accuracy shows only minimal degradation, there is a notable decline in classification accuracy when models are transferred [16, 19]. This outcome is expected, as similar species across different regions may exhibit slight variations in appearance. Additionally, certain animal species may experience changes in appearance due to environmental factors, seasonal variations (e.g., winter vs. summer), or variations among subspecies.

Given the time-intensive nature of manually labeling camera trap images, recent research has increasingly focused on training models using only partially manually labeled data, or alternatively, fully automatically labeled or clustered data sets [20, 21]. This shift in focus aims to reduce the reliance on extensive human effort while maintaining or even improving the accuracy and efficiency of wildlife monitoring and species identification.

Overall, while methodologies for wildlife classification exist, they often lack comprehensive data sets for specific animal species. Furthermore, the areas of attribute classification and re-identification are not sufficiently addressed in the literature and represent a significant technological gap.

1.2. Contribution

Our research introduces two innovative aspects that advance the current state-of-the-art:

1. Customized Data Set: We present a specific data set comprising labeled camera trap images from various regions in Austria. This data set features manually validated images accompanied by comprehensive metadata. Unlike existing data sets, ours includes additional information such as the sex and age class of the animals, alongside species identification and unique IDs for re-identifying individual animals.

2. Automated Image Classification with User Adjustments: We introduce a computer vision approach designed to aid users in accurately classifying camera trap images. This method performs automatic detection and classification, while also allowing users to refine the results. It does so by presenting the top n categories with corresponding confidence levels through an intuitive graphical user interface (GUI). As illustrated in Figure 9, the GUI exemplifies how users can interact with the system, making corrections and adjustments to the automatically detected and classified images.

2. Input Data Sets

A robust data set is a fundamental prerequisite for the effective training of machine learning models. Therefore, the proposed data set has been carefully prepared and consists of three main components covering 31 animal species. Firstly, data sets from various projects by the Meles office were utilized. The images for these data sets were captured across 34 different locations, providing a diverse range of environments and conditions. Secondly, the project leveraged a

data set from Deermapper cameras by Pentamap[6], which includes millions of images taken at numerous locations, from which a subset was randomly selected for the project. Although these images predominantly feature common animal species, they also contain rare species that are invaluable for studies of this nature. Thirdly, we strategically placed cameras in zoos, specifically at the Hague Zoo and the Cumberland Wildlife Park Grünau, to capture targeted images of rare animal species like the sika deer and lynx. These species were underrepresented in the data sets from actual hunting areas. Overall, the data set is highly generic as it comprises images from different cameras positioned at various locations, showcasing diverse weather, daylight, and seasonal conditions (cf. Figure 3). The images were labeled using the TRAPPER software[7]. We adopted the open standard Camtrap DP [22] for data formatting. In the initial batch, 68.000 images containing 88.000 objects were labeled. Each bounding box containing animals was annotated with species information, while age class, sex, and animal ID were included whenever possible.

Figure 3: Data set for this study: Randomly selected crops of animals of our proposed data set. The data set is generic and holds image from different cameras at various standpoints, altering weather, daylight, and seasonal conditions.

Due to the highly unbalanced nature of the original data set, in this initial work, we only considered classes with more than 300 images. To maintain data balance, for all classes that had more than 3000 images, only 3000 randomly selected images were used for training, resulting in 13 final classes.

The class distribution, bounding box centroids, and bounding box dimensions are illustrated in Figure 4. Despite preselection and reduction to specific classes, an imbalance is still evident (Figure 4, left). However, this current imbalance is manageable with various training techniques, such as assigning higher weights to classes with fewer instances and applying targeted and varied augmentations to balance the data effectively.

The centroids of annotated bounding boxes are well-distributed within the images (Figure 4, middle), with a slightly lower density in the lower part of the image. This is expected in camera trap images, as these cameras are usually mounted relatively low on a tree, causing the centroids of animals further away to be positioned towards the middle or upper part of the image. Conversely, if an animal is very close to the camera trap, it appears relatively large in

the image, leading to larger bounding boxes with centroids closer to the center of the image.

To train the model optimally for all animals which are captured at various scales and different sizes within the camera trap images, the distribution of bounding box sizes (width and height) is crucial. It is important to have both large and small bounding boxes relative to the image size. Figure 4 (right) shows a relative good distribution in our data set. It is evident that a significant portion of the bounding boxes covers approximately a quarter of the image, but some bounding boxes extend over even larger parts of the image. Additionally, many bounding boxes span the entire image, as indicated by the yellow-red points at a width or height of around 1. These statistics and visualizations demonstrate that the quality of the data is sufficient for successful training.

The processed and partially balanced data set ultimately contains 25.000 labeled objects. This might not seem like a large amount, especially for deep learning. However, as is well known, the balance, quality, and diversity of the data are much more critical in training a model than having a large number of similar low-quality images. This has also been confirmed by our own investigations and tests.

To improve the accuracy of detection and classification and to detect and classify additional species, the data set will be further expanded throughout the project. The already trained model will be utilized to selectively choose images for underrepresented species from millions of images, excluding those that are heavily overrepresented, such as red deer, as candidates for manual labeling. This approach will significantly reduce the number of images requiring manual labeling, making the labeling process more efficient and resource-saving.

For training purposes, the described data set was divided into training, validation, and test sets, containing 80%, 10%, and 10% of the total images, respectively. This subdivision was performed randomly for each class to ensure a representative distribution.

Beery et al. [19] recommend using a location-based split (assigning locations exclusively to either the training or test data sets), in addition to a random split, to mitigate the impact of images captured from the same location with similar backgrounds, which can lead to inflated accuracy results. While this approach was not incorporated in the initial tests, it is planned for future experiments. The data set contains detailed location metadata for each image, which will be utilized in upcoming analyses specifically designed to apply and assess the impact of a location-based split. This approach is expected to further improve the model's generalizability and robustness across different environments.

Compared to object detection and classification in cities or in man-made environments, detecting and classifying animals in the wild is significantly more challenging. As seen in Figure 5 (left), animals can be quite difficult to spot in some images (actually two fawns are hiding in the meadow). They often hide or blend into their surroundings. Additionally, the background in wildlife images is highly diverse and can become even more unique and challenging due to numerous factors. The daily (shadow) and seasonal changes, along with phenology, further enhance the dynamic and diverse nature of the background.

Additionally, in some images only small parts of the animal are visible, as illustrated in Figure 5 (right, bounding boxes: a-g). Such images can be challenging to label even manually. Therefore, this type of images, along with other animals that do not fit into any predefined class, is categorized under the *unknown* class. The *unknown* class plays a crucial role in the model's learning process by helping it recognize and correctly handle images that do not belong
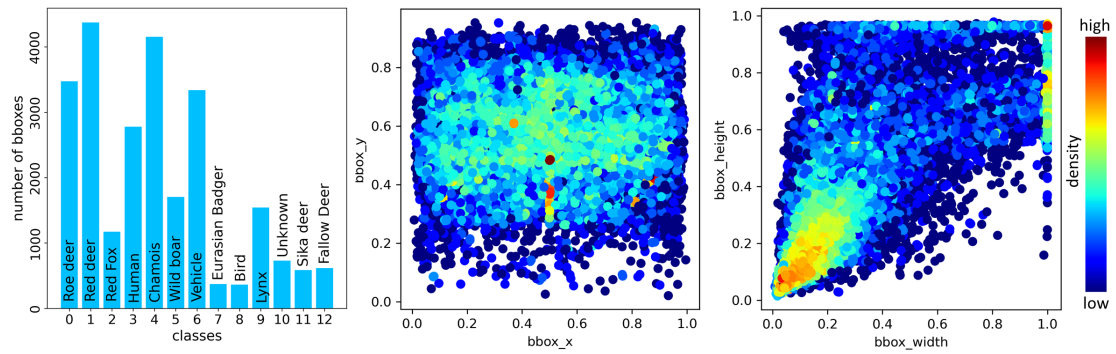
Figure 4: Custom data set used for this study. Shown are the class distribution (left), the distribution of the bounding box centroids holding animals within the images (middle), and the distribution of width and height (i.e. the size) of bounding boxes (right).

to any of the defined categories. By including this class, the model is trained to identify and avoid incorrectly predicting such images as belonging to one of the predefined classes. This prevents the model from making false positive predictions when encountering unfamiliar or ambiguous instances. During both training and evaluation, the *unknown* class is treated the same as any other class. It is included in the loss function and contributes to the model's overall performance metrics. This approach ensures that the model is robust in distinguishing between known classes and truly unknown or ambiguous examples, thereby enhancing its generalization capabilities.

The data set used for model training was sourced from multiple project partners and is not planned to be made publicly available.
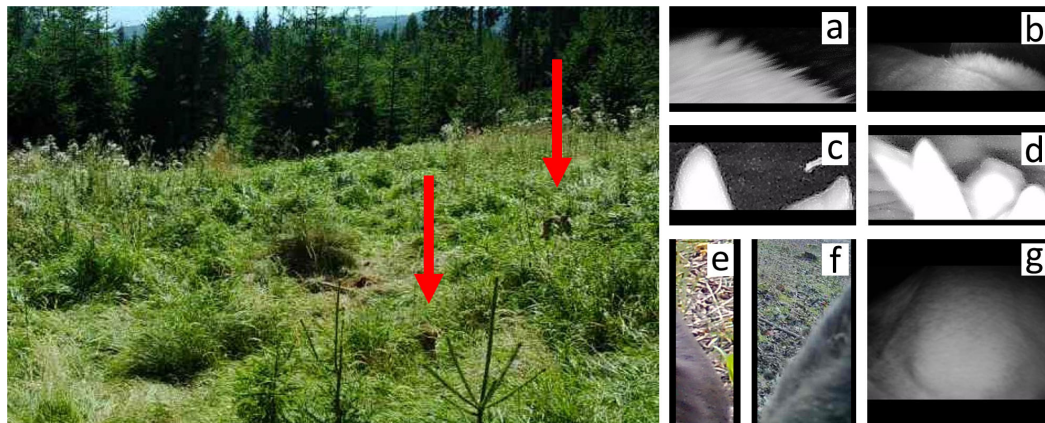
Figure 5: Challenging image where the two fawns indicated by the red arrows are barley visible (left) and exemplary bounding boxes of the *unknown* class (right: a-g).

3. Methodology

Our methodology for animal detection and species classification employs two distinct strategies, namely single-stage and two-stage approaches as depicted in Figure 6.

The single-stage approach involves training object detectors to perform both detection and classification simultaneously. We evaluated multiple advanced algorithms, such as Fast/Faster R-CNN, EfficientDet, and various models from the YOLO family. These models were assessed based on their runtime and accuracy on two devices: RTX A4000 GPU and NVIDIA Jetson TX2. To find the optimal configurations, we tested multiple models with different parameters, e.g., the input image size.

In the two-stage approach, the process is divided into two steps. Initially, a detector is trained to identify animals in the images. Subsequently, the detected bounding boxes are first cropped and then classified into specific species using different CNN architectures, including InceptionV3 [11], Xception [12], and EfficientNet (B0 to B7) [13]. Notably, we chose CNNs over transformer-based deep neural networks (DNNs) because the latter require significantly larger data sets to achieve comparable accuracy [23]. As the project progresses and more data becomes available, we plan first to explore combined CNN and transformer-based DNNs, such as CCT [24] and later pure transformer DNN like ViT [23] and SwinT [25].

The primary advantage of the two-stage approach is its flexibility. The animal detection model can remain unchanged while the classification model can be easily updated to include new species and attributes. Additionally, this method reduces the labeling workload for the detector, as it only requires bounding box annotations without needing detailed attribute labels.
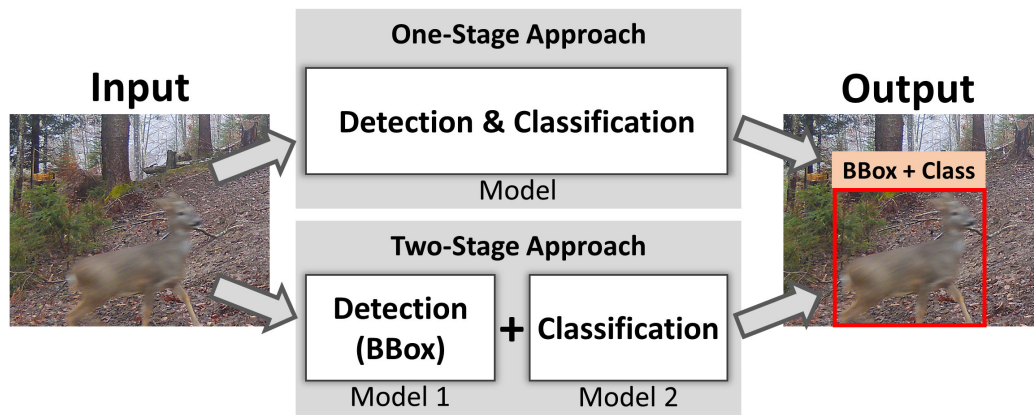
Figure 6: Detection and classification workflows: One-stage and two-stage approaches.

To facilitate user interaction, our approach not only provides the top prediction but also displays the best n species matches along with their confidence levels. This feature allows users to adjust labels if necessary. The user interface for the mobile application, designed to enable easy label modifications, is shown in Figure 9.

4. Results

Results are reported for one- and two-stage detection and classification of animal species, also including the developed mobile application.

4.1. One-Stage Detection and Classification

Initial evaluations revealed that among the various object detectors tested, some exhibited very slow performance (e.g., Fast/Faster R-CNN, EfficientDet). In contrast, scaled-YOLOv4 [26], YOLOX [27], and YOLOR [28] showed superior performance. Further testing indicated that YOLOR (specifically YOLOR with cross-stage partial connection) outperformed the others, followed by scaled-YOLOv4 and YOLOX, with all models using an input image resolution of 640 x 640 pixels. The inference speeds were 58.8/4.3 (YOLOR), 45.5/1.9 (scaled-YOLOv4), and 40.0/2.5 (YOLOX) frames per second on RTX A4000 / NVIDIA Jetson TX2, respectively.

The YOLOR-based detector was ultimately trained to recognize all animal species, as well as humans and vehicles. Evaluation on the test set demonstrated a weighted accuracy of 86.5% and an overall accuracy of 81.0% (cf. Figure 7, left). Certain classes, such as birds, had significantly lower accuracy, which also accounted for most of the undetected objects. Further analysis revealed that the bird class had relatively few available images, and the same bird species appeared in over 90% of these images. Additionally, the size of the birds in the images posed a challenge, as camera traps are sometimes triggered by birds that appear very small (only a few pixels) and blurry due to their distance from the camera.

The confusion matrix (cf. Figure 7, left) also shows misclassifications between visually similar species, such as roe deer, red deer, sika deer, and fallow deer. Moreover, animals are often only partially visible in the images, either at the corners, at considerable distances, or captured while moving rapidly, resulting in significant motion blur.

4.2. Two-Stage Detection and Classification

Initial tests of the two-stage approach revealed that EfficientNetB7 yielded the best performance, achieving a weighted accuracy of 76.8% and an overall accuracy of 76.6% (cf. Figure 7, right). The lower accuracies observed, compared to the one-stage approach, were anticipated due to the limited scope of hyperparameter tuning and specific fine-tuning conducted during the study. The following hyperparameters were varied during the experiments, with the final selected values highlighted in **bold**:

- **Optimizer: Adam**, AdamW, RMSprop
- **Learning Rate:** 0.01, **0.001**, 0.0001 in combination with **Weight Decay**: 0.001, **0.0001**, 0.00001
- **Batch Size:** 4, 6, **12**, 16, 32
- **Image Size:** 224, **299** (X- and Inception), **300** (EfficientNet), 528, 600
- **Training Epochs:** 25, **50**, 75, 100

The selection of these hyperparameters was not entirely arbitrary, although it did not result from an exhaustive optimization process. Instead, key hyperparameters were systematically varied,

and the most promising combinations were selected for the final models. Further automatic and extensive hyperparameter tuning, as well as adjustments to the neural network architecture, are expected to lead to improved accuracy and performance in future iterations.

Compared to the one-stage approach, similar challenges were encountered in this two-stage method, particularly with confusions between visually similar classes such as roe deer, red deer, sika deer, and fallow deer.

This suggests that the difficulties are inherent to the data set and the nature of the animal species rather than the detection method itself. For example, certain species were often partially visible, located at a distance, or captured while moving, resulting in motion blur, which presents a challenge for accurate classification. Future efforts will focus on refining the model through extensive hyperparameter tuning and exploring additional neural networks (e.g., transformer-based) to enhance performance. By addressing these challenges, we aim to improve the robustness and accuracy of the two-stage detection and classification system.
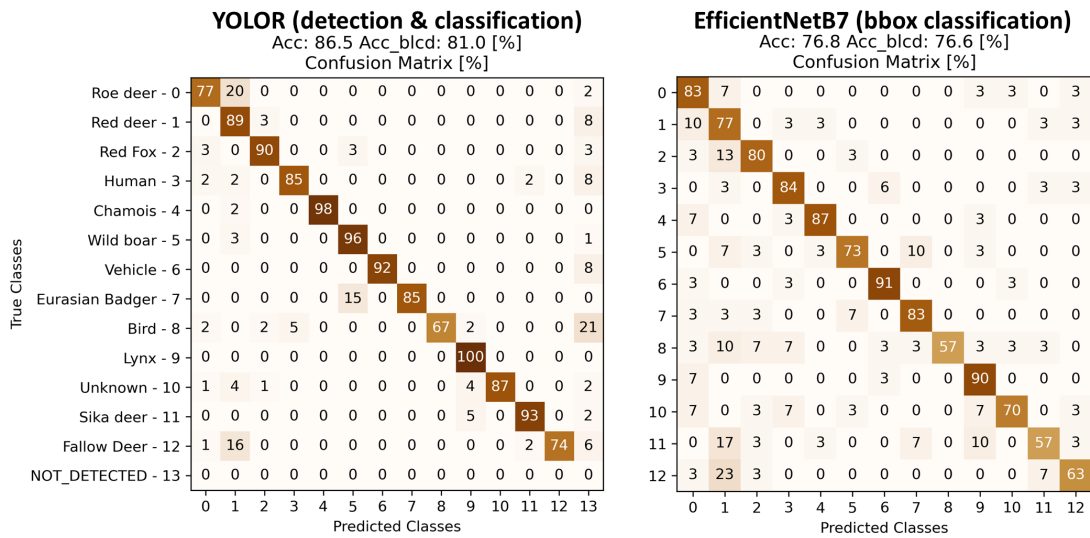
Figure 7: Results of one-stage detection and classification with YOLOR (left) and two-stage classification of reference bounding boxes using EfficientNetB7 (right).

Additionally, a detailed analysis was carried out to assess whether the correct species is included within the top 3 or top 5 predictions based on confidence scores. Figure 8 highlights the comparison using the confusion matrix and illustrates the accuracy improvements between the top 1, top 3, and top 5 predictions. The results show an accuracy increase of about 13% when considering the top 3 predictions compared to the top 1. Furthermore, there is an additional 6% gain in accuracy when expanding from the top 3 to the top 5 predictions. Overall, this results in a nearly 20% increase in accuracy when comparing the top 1 to the top 5 predictions. For some classes, the top 5 predictions achieved 100% accuracy, bringing the overall accuracy to 95%. However, despite the average accuracy improvement of nearly 20%, certain classes exhibit smaller accuracy gains. This is particularly evident in the *unknown* class, where the accuracy difference between the top 1 and top 5 predictions is only 7%. The likely reason for this modest

improvement is the high heterogeneity within the *unknown* class, which includes a wide range of animals that do not fit into any predefined category. This inherent diversity makes it more challenging for the model to accurately classify these instances. As a result, the *unknown* class remains more difficult to accurately predict than other more homogeneous classes.
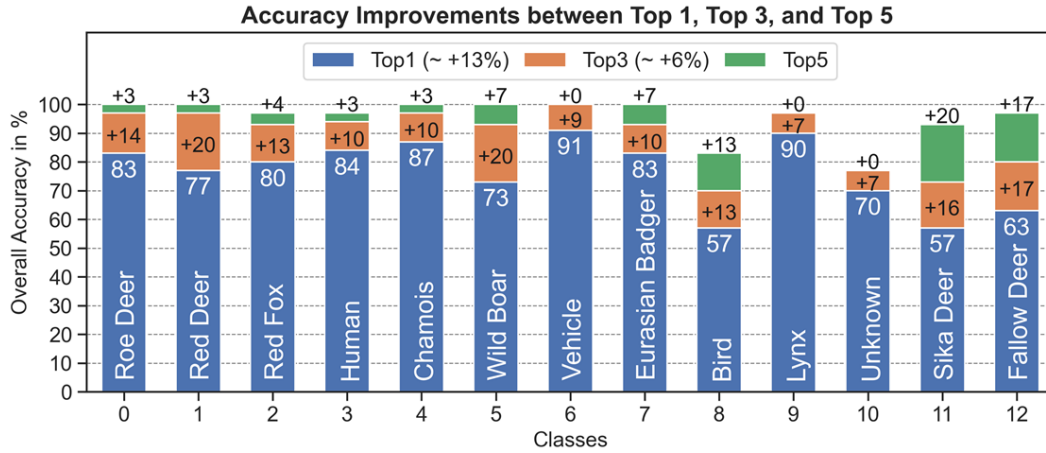
Figure 8: Classification of reference bounding boxes using EfficientNetB7 showing the improvements between top 1, top 3, and top 5.

4.3. Mobile Application

To demonstrate the practical benefits of considering the top 3 or top 5 predictions, we can look at specific examples from our data set. For instance, the fallow deer depicted in Figure 9 was initially misclassified as a red deer in the top 1 prediction. However, when considering the top 3 predictions, fallow deer was correctly identified as the second highest confidence prediction.

5. Conclusions and Outlook

This study provided valuable insights into wildlife classification, achieving species classification accuracies ranging from 77% to 87%. Both one-stage and two-stage approaches were explored, showing comparable performance, with the two-stage approach slightly lagging behind, likely due to the only partially exploited potential in hyperparameter optimization and fine-tuning. The strategy of using top 3 and top 5 predictions proved to be highly effective. Utilizing Top 3 predictions increased accuracy by 13% compared to top 1, and an additional 6% improvement was observed between top 3 and top 5, resulting in an overall accuracy of 95% with top 5 predictions.

The ability for users to correct automatically classified species (Figure 9) offered dual benefits: It allowed for easy and efficient correction of remaining misclassified images with minimal effort (just 1-2 clicks) and significantly enhanced the data set's value for retraining. Corrected images could then be incorporated into the training set, effectively implementing a kind of active learning approach to improve the models accuracy.

Figure 9: Mobile application which servers three purposes: Service for the users, validation of the users, and preselection of images to enhance the current data set.

Future developments will focus on optimizing and expanding the data set by including underrepresented and rare species. With an expanded data base, transformer-based DNNs will be employed and evaluated for classification tasks. Given that our current data sets already contain additional attributes such as age, sex, and partial individual identification, future methods will also aim at automatically determining these characteristics. Additionally, re-identification of individual animals will be a key area of development, further enhancing the robustness and applicability of the classification models.

By integrating these advancements, we aim to create a more accurate and versatile wildlife classification system that can adapt to new challenges and continuously improve through user interaction and expanded data.

Acknowledgments

References

[1] J. L. McCarthy, K. P. McCarthy, T. K. Fuller, T. M. McCarthy, Assessing variation in wildlife biodiversity in the Tien Shan mountains of Kyrgyzstan using ancillary camera-trap photos, Mountain Research and Development 30 (2010) 295–301.

[2] R. Y. Oliver, F. Iannarilli, J. Ahumada, E. Fegraus, N. Flores, R. Kays, T. Birch, A. Ranipeta, M. S.

Rogan, Y. V. Sica, et al., Camera trapping expands the view into global biodiversity and its change, Philosophical Transactions of the Royal Society B 378 (2023) 20220232.

[3] S. Leorna, T. Brinkman, Human vs. machine: Detecting wildlife in camera trap images, Ecological Informatics 72 (2022) 101876.

[4] J. Vélez, W. McShea, H. Shamon, P. J. Castiblanco-Camacho, M. A. Tabak, C. Chalmers, P. Fergus, J. Fieberg, An evaluation of platforms for processing camera-trap data using artificial intelligence, Methods in Ecology and Evolution 14 (2023) 459–477.

[5] S. Beery, D. Morris, S. Yang, Efficient pipeline for camera trap image review, in: Proceedings of the Data Mining and AI for Conservation Workshop at KDD19, 2019, pp. 1–2.

[6] S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: Towards real-time object detection with region proposal networks, Advances in Neural Information Processing Systems 28 (2015) 1–9.

[7] R. Girshick, Fast R-CNN, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1440–1448.

[8] M. Tan, R. Pang, Q. V. Le, EfficientDet: Scalable and efficient object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2020, pp. 10781–10790.

[9] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You only look once: Unified, real-time object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 779–788.

[10] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, in: Proceedings of the International Conference on Learning Representations, 2015, pp. 1–14.

[11] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the Inception architecture for computer vision, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2818–2826.

[12] F. Chollet, Xception: Deep learning with depthwise separable convolutions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1251–1258.

[13] M. Tan, Q. Le, EfficientNet: Rethinking model scaling for convolutional neural networks, in: International Conference on Machine Learning, 2019, pp. 6105–6114.

[14] N. Rigoudy, G. Dussert, A. Benyoub, A. Besnard, C. Birck, J. Boyer, Y. Bollet, Y. Bunz, G. Caussimont, E. Chetouane, et al., The DeepFaune initiative: A collaborative effort towards the automatic identification of European fauna in camera trap images, European Journal of Wildlife Research 69 (2023) 113.

[15] B. Dave, M. Mori, A. Bathani, P. Goel, Wild animal detection using yolov8, Procedia Computer Science 230 (2023) 100–111. URL: https://www.sciencedirect.com/science/article/pii/S1877050923020707. doi:https://doi.org/10.1016/j.procs.2023.12.065, 3rd International Conference on Evolutionary Computing and Mobile Sustainable Networks (ICECMSN 2023).

[16] C. Carl, F. Schönfeld, I. Profft, A. Klamm, D. 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 (2020) 62. URL: https://doi.org/10.1007/s10344-020-01404-y. doi:10.1007/s10344-020-01404-y.

[17] S. Schneider, G. W. Taylor, S. Kremer, Deep learning object detection methods for ecological camera trap data, in: Proceeding of Conference on Computer and Robot Vision (CRV), 2018, pp. 321–328. doi:10.1109/CRV.2018.00052.

[18] M. Tan, W. Chao, J.-K. Cheng, M. Zhou, Y. Ma, X. Jiang, J. Ge, L. Yu, L. Feng, Animal detection and classification from camera trap images using different mainstream object detection architectures, Animals 12 (2022). URL: https://www.mdpi.com/2076-2615/12/15/1976. doi:10.3390/ani12151976.

[19] S. Beery, G. Van Horn, P. Perona, Recognition in terra incognita, in: Proceedings of the European conference on computer vision (ECCV), 2018, pp. 456–473.

[20] T. Battu, D. S. Reddy Lakshmi, Animal image identification and classification using deep neural networks techniques, Measurement: Sensors 25 (2023) 100611. URL: https://www.sciencedirect.

com/science/article/pii/S2665917422002458. doi:`https://doi.org/10.1016/j.measen.2022.100611`.

[21] M. S. Norouzzadeh, A. Nguyen, M. Kosmala, A. Swanson, M. S. Palmer, C. Packer, J. Clune, Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning, Proceedings of the National Academy of Sciences 115 (2018) E5716–E5725. URL: https://www.pnas.org/doi/abs/10.1073/pnas.1719367115. doi:`10.1073/pnas.1719367115`. `arXiv:https://www.pnas.org/doi/pdf/10.1073/pnas.1719367115`.

[22] J. W. Bubnicki, B. Norton, S. J. Baskauf, T. Bruce, F. Cagnacci, J. Casaer, M. Churski, J. P. Cromsigt, S. D. Farra, C. Fiderer, et al., Camtrap DP: An open standard for the FAIR exchange and archiving of camera trap data, Remote Sensing in Ecology and Conservation (2023) 1–13.

[23] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16x16 words: Transformers for image recognition at scale, arXiv preprint arXiv:2010.11929 (2020) 1–22.

[24] A. Hassani, S. Walton, N. Shah, A. Abuduweili, J. Li, H. Shi, Escaping the big data paradigm with compact transformers, arXiv preprint arXiv:2104.05704 (2021) 1–18.

[25] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin transformer: Hierarchical vision transformer using shifted windows, in: International Conference on Computer Vision, 2021, pp. 10012–10022.

[26] C.-Y. Wang, A. Bochkovskiy, H.-Y. M. Liao, Scaled-YOLOv4: Scaling cross stage partial network, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2021, pp. 13029–13038.

[27] Z. Ge, S. Liu, F. Wang, Z. Li, J. Sun, YOLOX: Exceeding YOLO series in 2021, arXiv preprint arXiv:2107.08430 (2021) 1–7.

[28] C.-Y. Wang, I.-H. Yeh, H.-Y. M. Liao, You only learn one representation: Unified network for multiple tasks, arXiv preprint arXiv:2105.04206 (2021) 1–11.