Automated Visual Large Scale Monitoring of Faunal Biodiversity

Bernd Radig[a,*], Paul Bodesheim[b,], Dimitri Korsch[b,***], Joachim Denzler[b,****], Timm Haucke[c,*****], Morris Klasen[c,******], and Volker Steinhage[c,*******]**

[a] *Faculty of Informatics, Technical University of Munich, München, 80290 Germany*
[b] *Computer Vision Group, Friedrich Schiller University Jena, Jena, 07737 Germany*
[c] *Institut für Informatik, University Bonn, 53115 Bonn, Germany*
** e-mail: radig@in.tum.de*
*** e-mail: paul.bodesheim@uni-jena.de*
**** e-mail: dimitri.korsch@uni-jena.de*
***** e-mail: joachim.denzler@uni-jena.de*
****** e-mail: haucke@cs.uni-bonn.de*
******* e-mail: klasen@cs.uni-bonn.de*
******** e-mail: steinhag@cs.uni-bonn.de*

Abstract—To observe biodiversity, the variety of plant and animal life in the world or in a particular habitat, human observers make the most common examinations, often assisted by technical equipment. Measuring objectively the number of different species of animals, plants, fungi, and microbes that make up the ecosystem can be difficult. In order to monitor changes in biodiversity, data have to be compared across space and time. Cameras are an essential sensor to determine the species range, abundance, and behavior of animals. The millions of recordings from camera traps set up in natural environments can no longer be analyzed by biologists. We started research on doing this analysis automatically without human interaction. The focus of our present sensor is on image capture of wildlife and moths. Special hardware elements for the detection of different species are designed, implemented, tested, and improved, as well as the algorithms for classification and counting of samples from images and image sequences, e.g., to calculate presence, absence, and abundance values or the duration of characteristic activities related to the spatial mobilities. For this purpose, we are developing stereo camera traps that allow spatial reconstruction of the observed animals. This allows three-dimensional coordinates to be recorded and the shape to be characterized. With this additional feature data, species identification and movement detection are facilitated. To classify and count moths, they are attracted to an illuminated screen, which is then photographed at intervals by a high-resolution color camera. To greatly reduce the volume of data, redundant elements and elements that are consistent from image to image are eliminated. All design decisions take into account that at remote sites and in fully autonomous operation, power supply on the one hand and possibilities for data exchange with central servers on the other hand are limited. Installation at hard-to-reach locations requires a sophisticated and demanding system design with an optimal balance between power requirements, bandwidth for data transmission, required service and operation in all environmental conditions for at least ten years.

Keywords: biodiversity monitoring, species classification, animal detection, sensor networks, camera trapping with color and depth sensors, motion estimation in color and depth videos, fine-grained recognition, multi-object tracking, depth estimation, deep learning, novelty detection, anomaly detection, incremental learning, lifelong learning, active learning

DOI: 10.1134/S1054661821030214

INTRODUCTION

In order to obtain precise data on the development of faunal biodiversity, systematic, reproducible, and low error measurements must be carried out. To study large representative regions over a longer period of time, scientists are not sufficiently available. That is why the German government has launched a project called AMMOD (automated multisensor station for monitoring of biodiversity), which is developing prototypes for automated, networked, and self-sufficient measuring stations and testing them in the wild. These unique stations, when installed as a network, will generate a robust data pool for analyzing global changes in the faunal biosphere. The modules have the intrinsic ability to detect any sensor-specific signal to monitor birds, bats, mammals, and insects. Innovations include combining different sensors whose standardized signals complement each other and automating the devices and workflows. Technologies adapted for use in such an AMMOD station include automated

Received April 14, 2021; revised April 14, 2021;
accepted April 14, 2021

odor analysis that responds to minute concentrations of substances in the air, sampling robots for collecting organic particles and trapping insects, DNA metabarcoding for identifying species in bulk samples, bioacoustic monitoring primarily for detecting grasshoppers, bats, and birds, and, last but not least, camera systems for visual observation of, for example, mammals and insects. The plasticity and modular design of the base station enables the addition of other sensor types. Since AMMOD stations collect data from a radius in the immediate surroundings, only species occurring in this area will be detected. Especially for detecting mammals and birds using cameras and audio recorders, the positioning of recording devices (near the soil, in the canopy, along wildlife crossings) is crucial to increase the detection probability.

Species detection probability accounts for errors stemming from false positive and false negative detections from the classification programs, but accounts also for the incomplete information occurring inherently in every species survey.

In the case of time and species site occupancy rates, detection probability can be estimated by generating "detection histories," which are sequences of presence-absence records of a species per sampling location [8]. This can be easily derived by defining survey bouts of, e.g., one week intervals. In the case of species abundance, detection probability can be estimated either by establishing a sequential detection history (i.e., inference of abundance by repeated presence-absence surveys, [29]) or alternatively as a function of the detector-species detection distance (i.e., distance sampling methodology, [7, 17]). In the latter case, animal detection distances can be easily derived from stereoscopic images.

Species site occupancy estimation can in principle be applied to all AMMOD detectors and data output (e.g., [10, 18]). Additional verification of system output by humans to, e.g., remove false positive recordings can further improve the estimates [8]. As species site occupancy can change temporally (e.g., as a function of temperature), site occupancy can be estimated repeatedly over time to infer species activity and/or abundance.

The estimation of species abundance from presence-absence data can in principle be applied to all AMMOD detectors [11]. The estimation of species abundance based on distance sampling methodology is relatively straightforward when standard distance sampling methods can be applied. A direct count of insect individuals is possible, e.g., with the moth scanner. Finally, data from different detectors and additional data sources may be combined by using spatially integrated population models to infer abundance [9]. The synchrony of AMMOD signals is a unique basis for ecological analyses (Fig. 1).

"Camera traps are very widely used to monitor the presence of animals and record their behavior" [1]. In the Bavarian National Park alone, more than one million images have been collected from camera traps to date. A similar number was collected from cameras on so-called green bridges over highways [12, 26]. Statistics on the number of such images generated each year in Germany alone do not exist. However, only very few of them are evaluated. Common camera traps provide images for human observers rather than for automatic analysis. To design a suitable camera system to operate on AMMOD platforms in the wild, boundary conditions must be met.

- The systems must be self-sufficient in energy, e.g., from their own photovoltaic platforms (low power supply).
- They are wirelessly connected to external servers (high data volume vs. low bandwidth).
- There, the observed animals are classified. Abundances and densities of the species are calculated, combined with environmental information, and the data are transmitted to the servers.
- Unattended, uninterrupted stable operation, minor service requests, self-adaptation to all weather conditions and seasons.

Observed and analyzed data need to be available in or integrated into publicly accessible libraries for scientific, political, and economic use.

Our current task is to test the technology rather than developing large-scale biodiversity monitoring programs. Hence, AMMOD prototypes are deployed in fenced areas that are protected from vandalism, rich in species, accessible and—at least for the first trials—with electricity from the public grid. In future, when the stations are fully developed, they will need little maintenance and can be scattered over the country.

AMMOD CAMERA TRAPS

A high-end stereo imaging system will be developed for the AMMOD station to capture terrestrial species with appropriate magnification, resolution, and depth of field. The technical components of the so-called SpeciesSiteCam consist of two cameras for recording the stereo image pair, a detector of animal movements, an infrared lamp for the nocturnal illumination of the animals passing the detector, a stereo processor, a computer for preprocessing and cabling with the base station for power supply and data transmission (see Fig. 2).

Infrared illumination will be used where appropriate to enable night and twilight image capture without disturbing passing animals. The left and right cameras are triggered simultaneously by motion detectors and capture image sequences of adjustable length. The two images are passed to a stereo processor that computes a disparity image and a 3D image. The 3D image contains the computed three coordinates in the real world for each image point with a defined disparity value. Features of the animals such as distance, walking path,

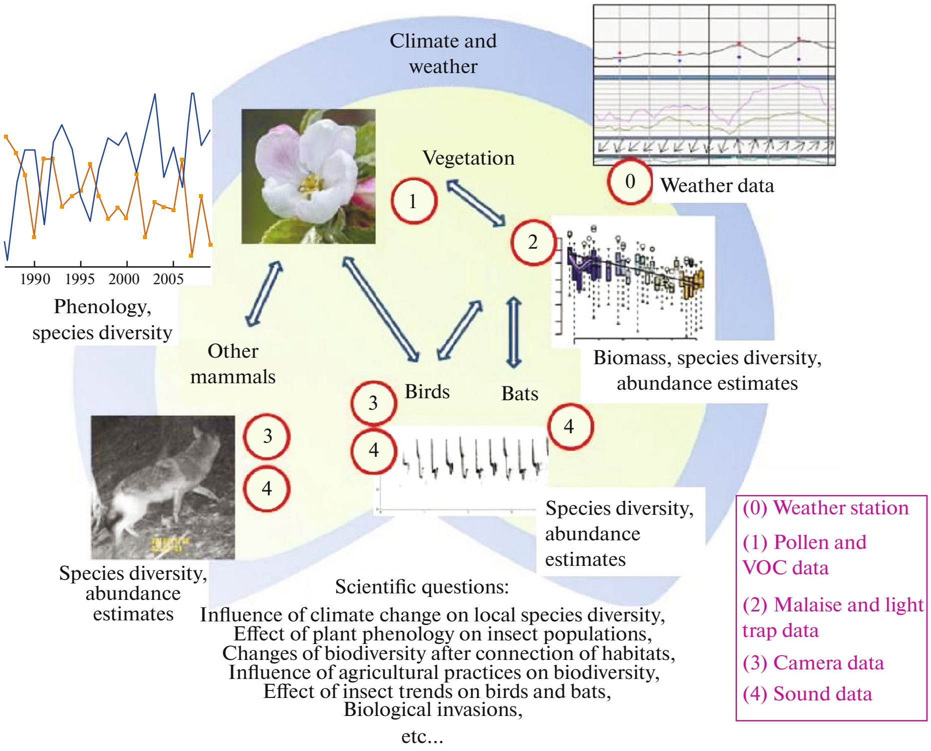
Climate and weather
Weather data
0
Vegetation
1
2
1990 1995 2000 2005
Phenology, species diversity
Other mammals
Birds
Bats
Biomass, species diversity, abundance estimates
3
4
3
4
4
Species diversity, abundance estimates
Species diversity, abundance estimates
Scientific questions:
Influence of climate change on local species diversity,
Effect of plant phenology on insect populations,
Changes of biodiversity after connection of habitats,
Influence of agricultural practices on biodiversity,
Effect of insect trends on birds and bats,
Biological invasions,
etc...
(0) Weather station
(1) Pollen and VOC data
(2) Malaise and light trap data
(3) Camera data
(4) Sound data

Fig. 1. Data integration over time and space on multi-sensor AMMOD stations.

body shape, and others can be determined from it. This data stream is transmitted for preprocessing to a powerful single-board computer on the base station. There, the data stream is broken down into its individual components and analyzed. Recordings are sorted out that do not contain animal images due to false alarms from the motion detector (typically 40–70%). A further reduction of the data volume is achieved by masking out image areas where no animals can appear.

An essential but critical component is the PIR (passive infrared) motion detector. It sends a trigger signal to the cameras when a moving object with a higher temperature than its surroundings moves through the observed area. The chosen detector is characterized by extensive programming options with which it can be adapted to the desired observation situation. These include, for example, the range or the angle of detection. Particularly important is the setting of parameters that suppress signals that erroneously report sudden temperature changes in the observed environment. It is essential to set the signal triggering for the time when animals are in the area of observation. Typical options are one-time triggering, video recording, or triggering a series of images at selected intervals. The observation area can be defined optically by placing Fresnel lenses, which are transparent to infrared light, in front of the actual temperature sensor. The motion detector selected so far is weatherproof, extremely low-power, and provides standardized trigger signals for easy integration into a camera system. The disadvantage is the time of about half a second that the detector needs for the reaction of the sensor element and the internal signal processing. In this time, animals may have already left the observation area, leading to empty images without any animal inside.

The first line of development being investigated is how to model the environment to be observed in more detail. In the meantime, PIR sensor components are available that only observe narrower angles but have a detection distance of up to 20 m. Several of these devices could then be distributed and aligned to look down paths or into areas where animals might approach the camera system in the first place. The preprocessing computer can then evaluate this set of signals to decide if and when the cameras will start to capture the images.

A second line of development is being pursued for use in AMMOD platforms that can be adequately powered. Here, the cameras can remain powered on (see Fig. 3).

The continuous image sequence can then be analyzed for the presence of animals. Initial attempts have been successful in using deep-learning techniques to learn a background image of the observation area. The appearance of an animal is then detected as an anomaly in that background. The latter works in tenths of a second. The accompanying figures show, from top to

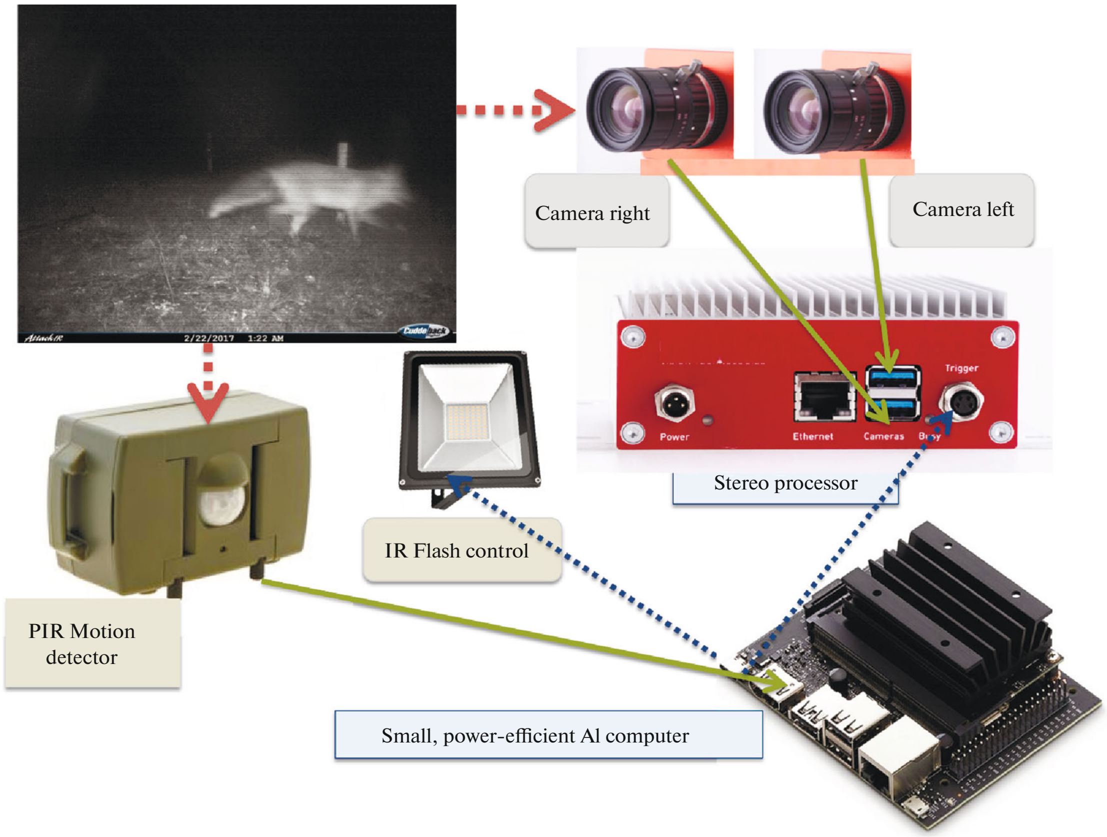

Fig. 2. System architecture with PIR motion detector.

bottom, a learned background, a passing animal, the image of the anomaly values (blue = low, red = high), the classified anomaly area bordered in red.

The inclusion of stereo data of the animal foreground and background will further advance the reliability and informativeness of this approach. Essential, however, is the ability, as long as lighting conditions or strong infrared illuminators allow to trigger and capture animal images from more or less any distance. To our knowledge, there are no suitable alternatives to either variant of triggering in technical or scientific publications to date.

AUTOMATIZED LONG-TERM WILDLIFE CLASSIFICATION

Since the AMMOD stations should operate for a longer period of time (several years), it is important to use recognition systems that can improve their performance over time by exploiting data that is recorded during operation. Hence, we aim at applying and improving lifelong learning algorithms for the identification of wildlife animal species that occur in the images from the camera traps. Lifelong learning, often also called continuous learning, continual learning, or incremental learning, denotes a process that enables updates of the classification models once new data is available. An important aspect that is different to conventional classification systems is the possibility to handle new, previously unseen categories, in our case new species unknown to the classifier. This is done by first applying novelty detection mechanisms to identify unknown species and second including human experts in the loop to obtain labels and valuable feedback that is integrated in an active learning fashion.

To start with an automatic recognition system, an initial classifier is trained in an initial learning stage using available training data from public databases that contain the relevant species. Then, novelty detection and active learning is applied to realize the lifelong learning cycle. Furthermore, active learning plays an important role for exploiting unlabeled sample sets. Of course, it is beneficial to include methods from domain adaptation, since the images from the initial training set of existing databases may differ clearly from images recorded by the AMMOD stations. Additionally, fine-grained recognition approaches are required that allow for distinguishing visually very similar species, like different bird species.

For the initial learning step, the collection of appropriate sets of labeled images is required to train an initial classifier from already digitized data. First, images of relevant species from existing labeled image datasets can be exploited, e.g., from ImageNet [23] or

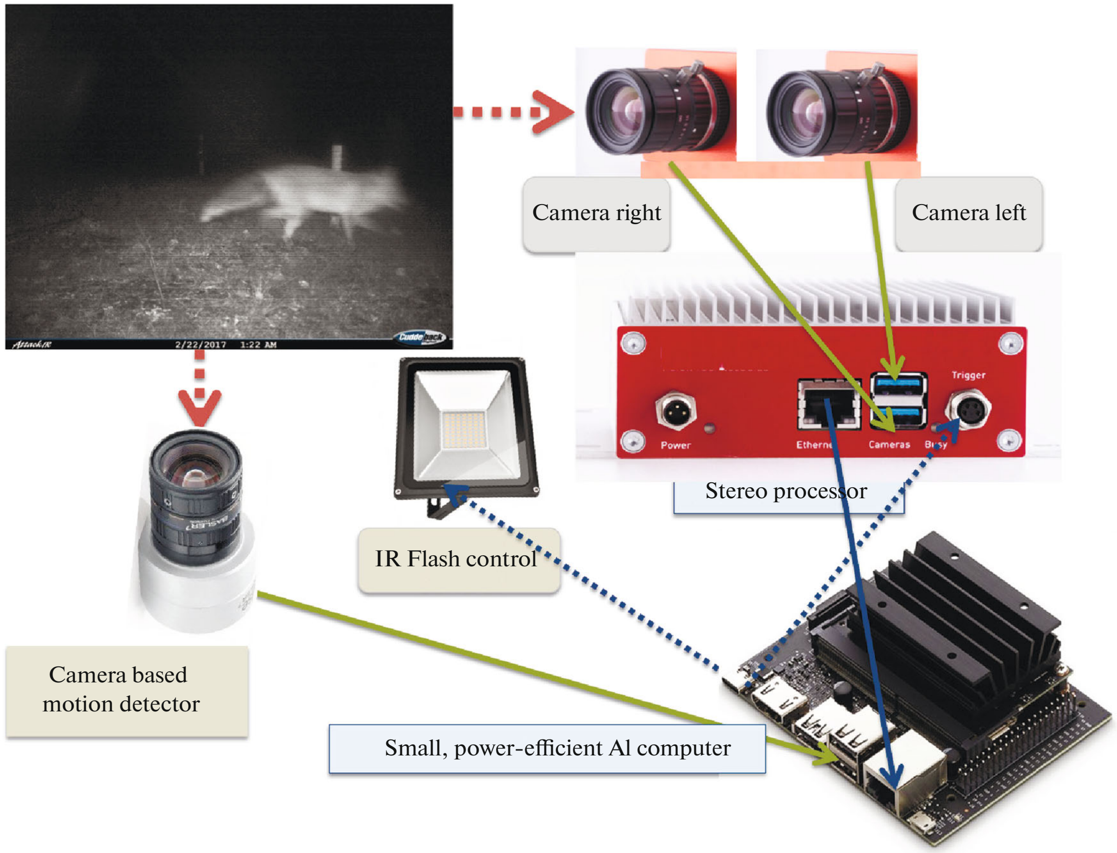

Fig. 3. Animal detection as an anomaly in the stream of images.

iNaturalist [33]. In addition, we use collections of annotated images from citizen scientists and natural history museums. Especially when exploiting images from collections of museums, methods from domain adaptation [13, 27] are required to compensate for the different image characteristics and camera setups. Furthermore, domain adaptation allows for integrating image data from different sensors (e.g., different camera types and images from the Internet), and can support the modeling of different environments with different background structures and lighting conditions.

Detection rates in the laboratory or at test data sets of more than 90% can generally not be achieved in the field under real, constantly changing conditions. For this reason, a feedback mechanism is provided, which involves the user in the optimization of the system, but at the same time minimizes the labeling effort to be made manually. This is known as active learning. In the actual application, the system will, at certain intervals, forward images to humans, and after the annotation these images will be used to train a classifier iteratively. Evaluation is carried out with a retained test set of images annotated manually.

Monitoring devices will likely also detect species that are unknown for the recognition system. Therefore, novelty detection is relevant [2, 3] which allows for automatic detection of animals that do not belong to the set of already known categories or species. In case that a known object has been incorrectly marked as "new" by the system, the corresponding feedback from an expert can contribute to improve the system performance. If a previously unknown species has actually been observed, the expert provides the corresponding label and this knowledge can be used to feed the classifier with the new type (incremental learning). Novelty detection can be evaluated by excluding images from a certain subset of classes in the training set, but including those classes in the test set.

When dealing with an increasing number of different animal species, some species will be visually more similar to each other, and some will be clearly different. The high visual similarity of related species is a particular challenge that is commonly addressed by fine-grained recognition approaches, especially when considering a particular domain like birds or moths and differentiating between different bird species or moth species. Furthermore, the visual appearance of related species can differ only slightly while individuals of the same species can look different, e.g., due to different viewing directions, poses, actions or background variation and occlusion. Fine-grained recognition treats this problem often via so-called part-based methods [20, 21]. Objects are modeled via com-

Table 1. Comparison of the daytime and nighttime modes of operation of our RGB-D camera trap [14]

	Daytime mode	Nighttime mode
Motion detection	Image-based	PIR
Depth acquisition	Active stereo	Passive stereo
Image acquisition	Color camera	Infrared camera
Illumination	Ambient light, active stereo pattern	Infrared lamp

ponents that are either explicitly extracted from annotated training data or are determined completely unsupervised by the system [31, 32]. Evaluation of the fine-grained classification methods are carried out on an annotated set of recorded data, that is split into training and test set, as well as on publicly available datasets for fine-grained recognition in the computer vision community.

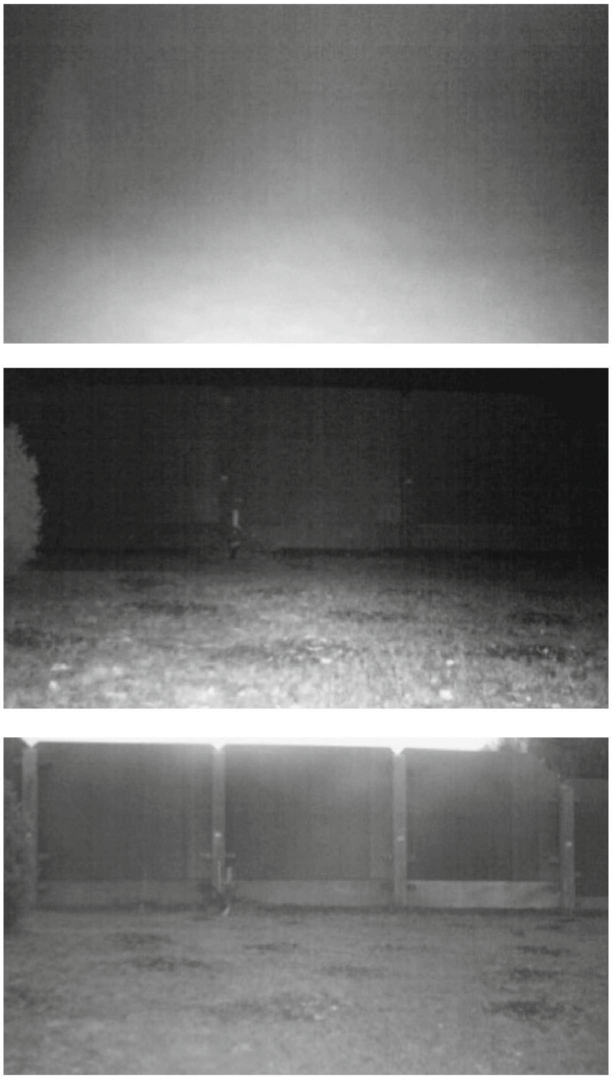

Fig. 4. Typical background training images.

STEREO FOR NEW QUALITY OF MONITORING

We derive depth information by applying stereo analysis. Depth is an important cue to discriminate observed animals efficiently from the environment (especially, for camouflaged or nontextured animals). Additionally, depth data allow deriving species abundance by estimating species detection probability as a function of the detector-species detection distance (i.e., distance sampling methodology [17]). We will exploit depth data in three ways. (1) Segmentation to separate animals from background and animals in different distances; (2) tracking of detected animals, to utilize their spatial movements for species identification; and (3) derivation of species abundances. Furthermore, depth information allows discarding irrelevant background areas and provides cues for the appropriate scale of object identification procedures.

We collect depth data using our RGB-D camera trap. It consists of an Intel®RealSense™ D435 active stereo camera [19], a NVIDIA® Jetson Nano™ developer kit for data processing, a passive infrared sensor (PIR) for animal detection at nighttime and an infrared lamp for nighttime illumination. Table 1 summarizes the two time-dependent modes of operation of the RGB-D camera trap.

We combine the depth data with images taken in the visible (daytime) or infrared (nighttime) spectrum. We map infrared images into the same RGB (red, green, blue) color-space as the color images. We refer to the combined images as RGB-D images.

We placed the RGB-D camera trap in a zoo scenario as this allows us to generate new images of animals at a higher frequency than in a purely natural scenario. The following figure shows a resulting RGB-D image.

We treat the problem of animal detection, fine-grained localization and classification as an instance segmentation problem. Most existing image understanding methods use solely color information. To make use of the additional depth information, we extend the Mask R-CNN [15] deep learning instance segmentation architecture to D-Mask R-CNN [14]. In Mask R-CNN, color image feature descriptors are determined by a *backbone* deep convolutional neural network (CNN) such as ResNet-50 [16], often configured as a feature pyramid network (FPN, [23]) to operate at different scales. We keep the ResNet-50 FPN color backbone and extend Mask R-CNN by an additional depth backbone. Both backbones have a similar architecture and are both initialized with the weights of ResNet-50 originally pretrained on the ImageNet dataset [30]. Depth images exhibit features similar to color images, such as strong gradients along object boundaries. We therefore argue that weights initially optimized for color images are also valuable for detecting such features in depth images. As the ResNet-50 architecture expects a three-channel RGB input image as input, its first convolutional layer has to

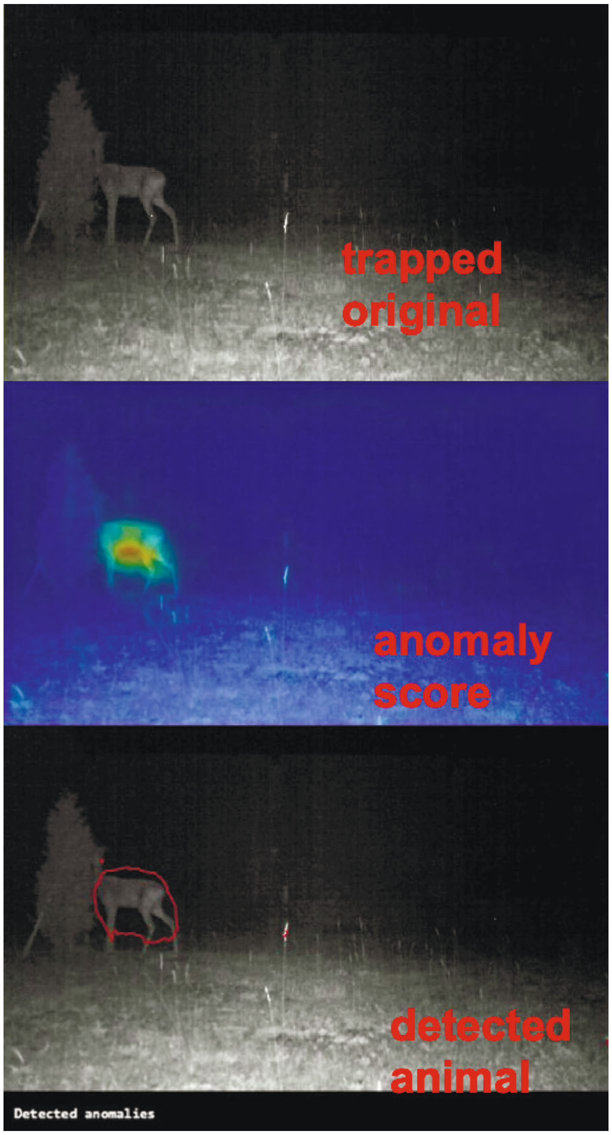

Fig. 5. Illustration of anomaly processing sequence.

be modified to take a single-channel depth image as an input. We achieve this by averaging the respective weights over the red, green and blue channels, which results in an unchanged expected activation in the following convolutional layers. Furthermore, we normalize all input channels (red, green, blue, and depth) by subtracting the respective mean and dividing by the respective standard deviation over the whole dataset. We also introduce a feature fusion module to combine the extracted features from both backbones using one 3 × 3 convolution per FPN scale and reduce the number of channels from 2 × 256 to 256. This allows us to use weights pre-trained on the Microsoft COCO dataset [24] for the region proposal network (RPN) classifier and mask head. Fig. 4 illustrates the general architecture of the resulting D-Mask R-CNN architecture.

Table 2. Average precision (AP) scores of the animal detection and instance segmentation on our camera trap dataset [14]

Bounding boxes	D-Mask R-CNN
AP	59.94%
AP50	94.50%
AP75	63.96%
Segmentation	D-Mask R-CNN
AP	37.27%
AP50	94.50%
AP75	13.25%

We restrict our evaluation of D-Mask R-CNN to instances of deers as it is the most common species in our RGB-D dataset. We quantify the results of D-Mask R-CNN using the average precision (AP), AP50, and AP75 metrics as defined by the Microsoft COCO dataset [24]. AP50 and AP75 denote the average precision at an intersection over union (*IoU*) threshold of 50 and 75%, respectively. AP denotes the average precision over *IoU* thresholds from 50 to 95% in 5% increments. The outlined metrics are summarized for bounding box detection and segmentation in Table 2.

For 3D multiobject tracking (MOT) of individuals we are using the RGB-D frames and intrinsic camera parameters to calculate 3D point clouds of the scene. With 2D mask projections and depth-expanded optical flow [34] we estimate instance-level scene-flow for each frame. By combination of frame-level scene-flow predictions and mask projections using the Hungarian matching algorithm we create individual animal tracklets (Fig. 8). In detail, large parts of this work dealt with preprocessing steps taken to calculate robust and usable point clouds. For temporal consistency of depth maps we employed a conditional median filter over the temporal component. We smooth out temporal inconsistency over a frame-window while constraining the peak-to-peak differences to preserve edges of actual animal movements. We aligned the point clouds to the world-coordinate system by estimating the ground plane using a RANSAC fitted plane, which is used to estimate the extrinsic camera parameters. And as the last processing step, we remove statistical outlier points of individual animals caused by blurry depth-map edges and hence imprecise projection to world-coordinates.

MOTH SPECIES DETERMINATION

The AMMOD Moth Scanner will take high-resolution images of insects that rest on an illuminated screen. For lighting, an LED lamp is used (Fig. 9), which can emit white light, ultraviolet light or both mixed [5]. This light is particularly attractive for moths [4].

A camera system images this screen at selected intervals and provides high-resolution images that are analyzed for the presence of identifiable insects

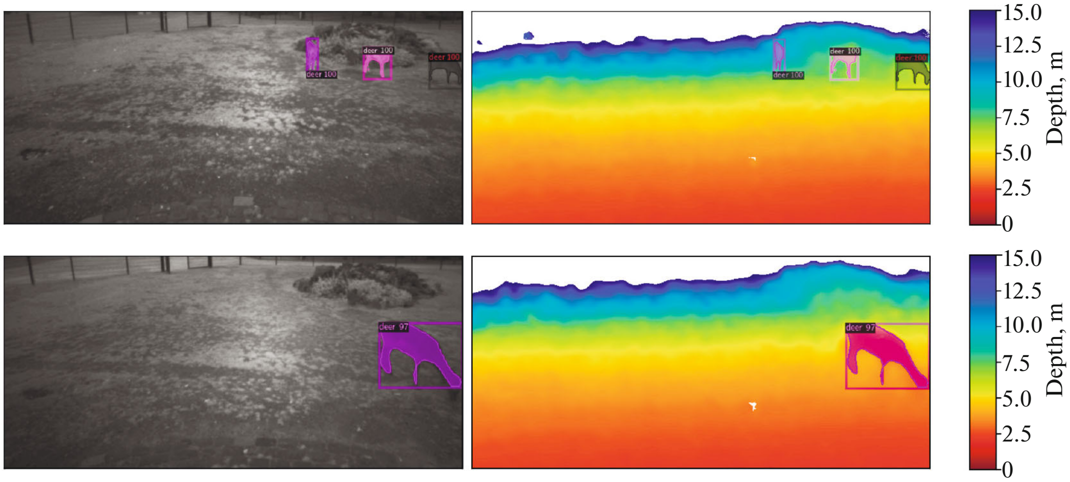

Fig. 6. Intensity images from zoo scenario mapped into an RGB image (left). Corresponding stereo depth estimations in the form of a so-called heat map, in which blue indicates more distant and red closer scene components (right). Animal instance segmentation (overlaid) with D-Mask R-CNN [14].

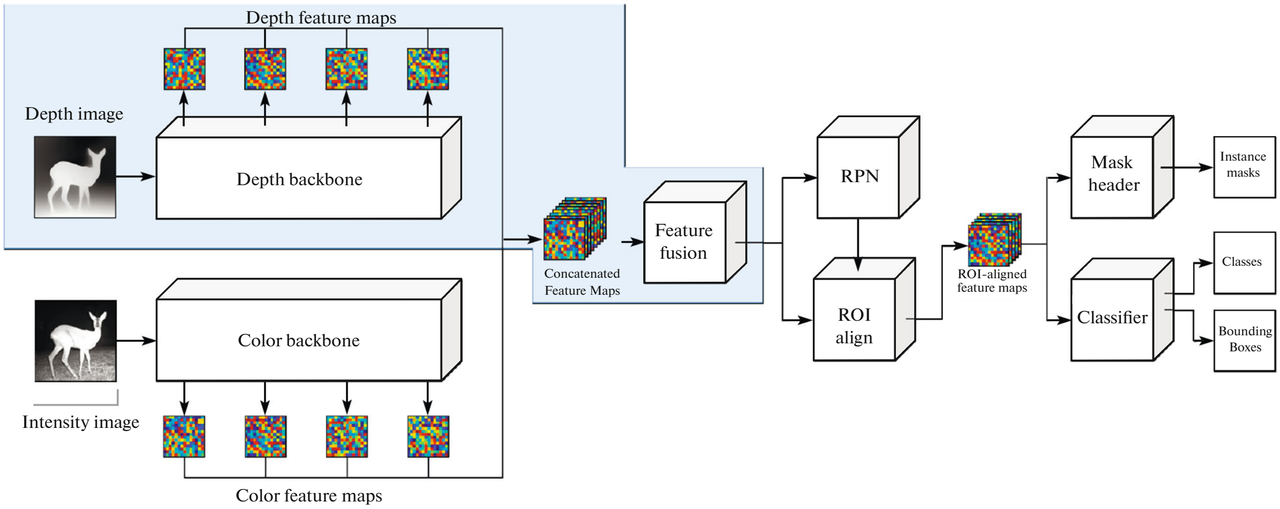

Fig. 7. The D-Mask R-CNN architecture. Modifications to the standard Mask R-CNN architecture are highlighted in blue.

(Fig. 10). A camera with 20 megapixel resolution is used. Images with LED flash are taken at regular, programmable intervals, typically several minutes. Subsequent analysis breaks down the image into dark areas that are approximately the area of the moths to be classified. Areas that are too small are ignored, and areas that are too large are examined to see if they remain for only one period of time (possibly feeding birds), remain unchanged (e.g., leaf blown on), or continue to be seen the next night (dirt on the screen). For such images, a message is generated to trigger human action if necessary (e.g., cleaning the screen).

Power consumption is a crucial aspect of the entire system. First, we intend to shut down the system to save power according to current weather data when moth flight is not expected. Furthermore, the system will be also turned off if the power consumption between the other sensors on the AMMOD platform is not sufficient for overall operation and must be fairly distributed. Finally, the station will turn off the camera system between scheduled exposures to preserve power, significantly reducing average power consumption. We also plan to turn off the light source at programmable intervals for a programmable time to allow moths to leave the screen. In the mentioned scenarios, theoretically, the entire system does not consume any power. Only corresponding areas are forwarded for further evaluation if they were not already contained unchanged in the previous image.

Early tests with mock samples of moth images have shown that currently automated identification success of species is about 80% (see also [28]). Accuracy is increasing very rapidly with improved algorithms every year. We will get valuable datasets for a large

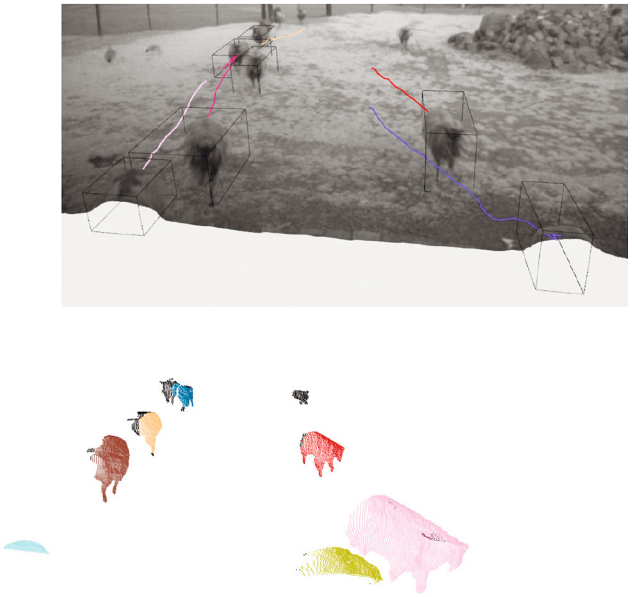

Fig. 8. 3D multiobject tracking using scene-flow (left). Instance clustering based on 2D mask projections (right).

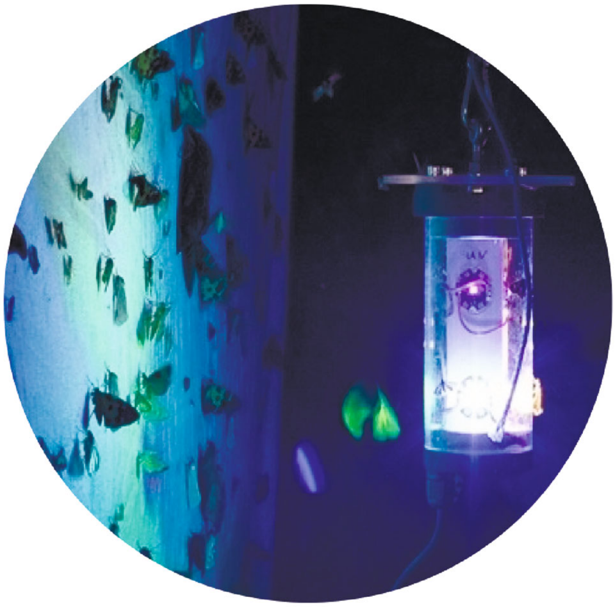

Fig. 9. LED Lamp attracting moths, © Gunnar Brehm, Phyletisches Museum Jena, 2017.

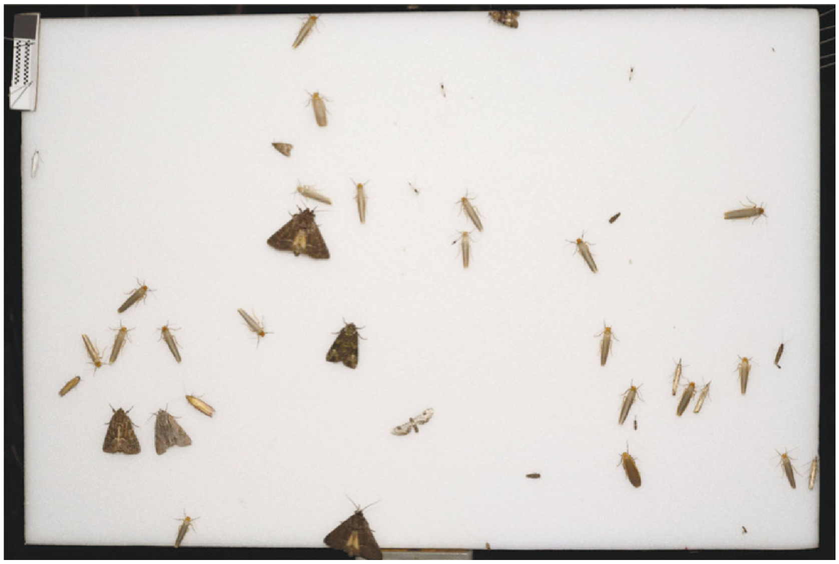

Fig. 10. Example of moth image, © Gunnar Brehm, Phyletisches Museum Jena, 2020.

number of species that can be identified because of their characteristic shape and color.

The goal of monitoring is not to detect every species, but to describe trends in local insect populations. It is important to know, for example, if species numbers and/or abundances drop by, e.g., 20% for all identified species in five years. Additionally, seasonal activity differences of single species, the correlation with weather conditions, and with vegetation phenology are crucial for ecological analyses.

In practice, some smaller species or species that do not stop flying will not be identified. However, the number of species considered is much higher than in current insect monitoring schemes (mainly diurnal butterflies, mostly presence-absence data) and the data will be comparable for all those species that can be detected. The year-round species monitoring is a novelty that allows for a much more detailed correlation with weather, plant phenology and the activity of other animals.

The functionality of the moth scanner consists of two parts. First, individual insects need to be detected (Fig. 11), i.e., a localization of single individuals in the image and description of their position and extent by circumscribing rectangles (bounding boxes). Then, for

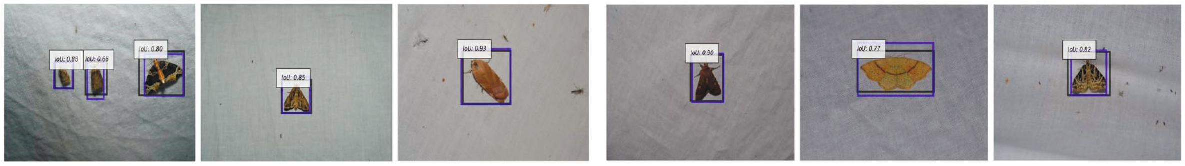

Fig. 11. Example images comparing automatically determined bounding boxes (black) with manual annotations (blue) for individual moths.

each detection, a prediction is made using a classification model to infer to which species the detected insect belongs to.

For the localization of moths in an image, we have trained a detector using the single-shot multibox detector (SSD) approach of [25] and annotated images of 200 different species. We achieved detection rates of 93.71 and 99.35% average precision (AP) at 75 and 50% IoU, respectively.

We then trained a standard deep learning classifier (ResNet-50 [16] pre-trained on ImageNet [30]) on the cropped bounding box images of the detected moths and achieved an average accuracy of 89.00% on an held-out test set. We also make use of additional images obtained from Internet image search engines to augment the training dataset. However, since this leads to noisy annotated images, filtering different types of label noise is required as proposed by [6]. This can further improve the recognition performance of the system, especially when the initial training set has been rather small with less than ten sample images per species.

CONCLUSIONS

In this paper, we have shown our current developments of visual monitoring systems for AMMOD stations that aim at recording animals in their surroundings using different sensors with the goal of observing trends and changes in species biodiversity. On the one hand, we explained the hardware setup for continuously recording images that has been chosen due to the various constraints and limitations for self-sustaining monitoring stations. On the other hand, we have detailed the individual software components that are required for an automatized monitoring of animal species.

For long term wildlife classification, we described our lifelong learning approach. In this approach, the species classifier should be updated in an incremental learning manner with new data recorded at the stations and labeled by experts in an active learning setup. Further components of the system denote novelty detection to handle species that are unknown for the initial classifier, since they were not present in the training set and domain adaptation to account for different imaging sensors and varying data sources.

With D-Mask R-CNN, we presented a novel approach to instance segmentation in RGB-D imagery which we evaluated using a proof-of-concept RGB-D camera trap setup in a zoo scenario. D-Mask R-CNN shows AP scores of 59.94 and 37.27% for animal detection by bounding boxes and segmentation masks, respectively. We plan to deploy stereo camera traps with larger baselines to improve depth estimation for more distant animals. Additionally we proposed RGB-D video processing to point clouds, enabling 3D multiobject tracking with improved characteristics over conventional 2D tracking.

Finally, we explained a specific part of the monitoring system called the moth scanner that is designed to record images of moths during the night. The moth scanner uses a light trap in the form of an illuminated screen. We also presented initial results for automatized moth localization (AP75 of 93.71%) and moth species identification (accuracy of 89.00%) in the resulting images with a deep learning detector and classifier.

FUNDING

This work was partially done within the project “Automated Multisensor Station for Monitoring of Species Diversity” (AMMOD), which is funded by the German Federal Ministry of Education and Research (Bundesministerium für Bildung und Forschung (BMBF), Bonn, Germany (FKZ 01LC1903B).

COMPLIANCE WITH ETHICAL STANDARDS

This manuscript is a completely original work of its authors; it has not been published before and will not be published in other sources.

CONFLICT OF INTEREST

The content of the article does not give grounds for raising the issue of a conflict of interest.

REFERENCES

1. P. Apps and J. W. McNutt, “How camera traps work and how to work them,” Afr. J. Ecol. **56** (4), 702–709 (2018).
2. P. Bodesheim, A. Freytag, E. Rodner, and J. Denzler, “Local novelty detection in multi-class recognition problems,” in *IEEE Winter Conference on Applications of Computer Vision (WACV)* (2015), pp. 813–820.
3. P. Bodesheim, A. Freytag, E. Rodner, M. Kemmler, and J. Denzler, “Kernel null space methods for novelty detection,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2013), pp. 3374–3381.

4. G. Brehm, J. Niermann, L. M. Jaimes Nino, D. Enseling, T. Jüstel, J. C. Axmacher, E. Warrant, and K. Fiedler, "Moths are strongly attracted to ultraviolet and blue radiation," Insect Conserv. Diversity **14** (2), 188–198 (2021).
5. G. Brehm, "New LED lamp for the collection of nocturnal Lepidoptera and a spectral comparison of light-trapping lamps," Nota Lepidopterol. **40**, 87–108 (2017).
6. J. Böhlke, D. Korsch, P. Bodesheim, and J. Denzler, "Lightweight filtering of noisy web data: Augmenting fine-grained datasets with selected internet images," in *International Conference on Computer Vision Theory and Applications (VISAPP)* (2021), pp. 466–477.
7. S. T. Buckland et al., *Distance Sampling: Methods and Applications* (Springer, New York, NY, 2015).
8. T. Chambert, D. A. W. Miller, and J. D. Nichols, "Modeling false positive detections in species occurrence data under different study designs," Ecology **96** (2), 332–339 (2015).
9. R. B. Chandler and J. D. Clark, "Spatially explicit integrated population models," Methods Ecol. Evol. **5** (12), 1351–1360 (2014).
10. A.-S. Crunchant et al., "Listening and watching: Do camera traps or acoustic sensors more efficiently detect wild chimpanzees in an open habitat?," Methods Ecol. Evol. **11** (4), 542–552 (2020).
11. I. Fiske and R. Chandler, "Unmarked: An R package for fitting hierarchical models of wildlife occurrence and abundance," J. Stat. Software **43** (10), 1–23 (2011).
12. P. Follmann and B. Radig, "Detecting animals in infrared images from camera-traps," Pattern Recognit. Image Anal. **28** (4), 740–746 (2018).
13. D. Haase, E. Rodner, and J. Denzler, "Instance-weighted transfer learning of active appearance models," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2014), pp. 1426–1433.
14. T. Haucke and V. Steinhage, "Exploiting depth information for wildlife monitoring," arXiv (2021). arXiv:2102.05607
15. K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (2017), pp. 2961–2969.
16. K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016), pp. 770–778.
17. E. J. Howe, S. T. Buckland, M. L. Després-Einspenner, and H. S. Kühl, "Distance sampling with camera traps," Methods Ecol. Evol. **8** (11), 1558–1565 (2017).
18. A. K. Kalan et al., "Towards the automated detection and occupancy estimation of primates using passive acoustic monitoring," Ecol. Indic. **54**, 217–226 (2015).
19. L. Keselman, J. Iselin Woodfill, A. Grunnet-Jepsen, and A. Bhowmik, "Intel real-sense stereoscopic depth cameras," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPR-WS)* (2017), pp. 1–10.
20. D. Korsch and J. Denzler, "In defense of active part selection for fine-grained classification," Pattern Recognit. Image Anal. **28** (4), 658–663 (2018).
21. D. Korsch, P. Bodesheim, and J. Denzler, "Classification-specific parts for improving fine-grained visual categorization," in *German Conference on Pattern Recognition (GCPR)* (2019), pp. 62–75.
22. A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems (NIPS)* (2012), pp. 1097–1105.
23. T. Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2017), pp. 2117–2125.
24. T. Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common Objects in Context," in *European Conference on Computer Vision (ECCV)* (2014), pp. 740–755.
25. W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *European Conference on Computer Vision (ECCV)* (2016), pp. 21–37.
26. B. Radig and P. Follmann, "Training a classifier for automatic flash detection in million images from camera-traps," in *International Conference on Pattern Recognition and Artificial Intelligence (ICPRAI 2018)* (2018), pp. 591–593.
27. E. Rodner and J. Denzler, "Learning with few examples for binary and multiclass classification using regularization of randomized trees," Pattern Recognit. Lett. **32** (2), 244–251 (2011).
28. E. Rodner, M. Simon, G. Brehm, S. Pietsch, J. W. Wägele, and J. Denzler, "Fine-grained recognition datasets for biodiversity analysis," in *CVPR Workshop on Fine-Grained Visual Classification (CVPR-WS)* (2015).
29. J. A. Royle and J. D. Nichols, "Estimating abundance from repeated presence–absence data or point counts," Ecology **84** (3), 777–790 (2003).
30. O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, and A. C. Berg, "ImageNet large scale visual recognition challenge," Int. J. Comput. Vision **115** (3), 211–252 (2015).
31. M. Simon and E. Rodner, "Neural activation constellations: Unsupervised part model discovery with convolutional networks," in *International Conference on Computer Vision (ICCV)* (2015), pp. 1143–1151.
32. M. Simon, E. Rodner, and J. Denzler, "Part detector discovery in deep convolutional neural networks," in *Asian Conference on Computer Vision (ACCV)* (2014), pp. 162–177.
33. Van G. Horn, O. Mac Aodha, Y. Song, Y. Cui, C. Sun, A. Shepard, H. Adam, P. Perona, and S. Belongie, "The iNaturalist Species Classification and Detection Dataset," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2018), pp. 8769–8778.
34. G. Yang and D. Ramanan, "Upgrading optical flow to 3D scene flow through optical expansion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2020), pp. 1334–1343.

Bernd Radig is Professor Emeritus at the Technical University of Munich (TUM), Department of Informatics, and TU Head of the Chair Cognitive Systems. 1986 he got a Chair for Image Understanding and Knowledge Based Systems at TUM. His research area is Artificial Intelligence, especially image and image sequence understanding. Research topics include tracking of cars in traffic scenes, analysis of football matches from television broadcasts, recognition of human emotions from image sequences of the face, human-robot communication, actions of persons in interior spaces, extension of driver assistance systems for road tracking and for automatic emergency braking, and currently on a large German infrastructure project of monitoring faunal biodiversity. Here, he leads a team to develop the visual sensors of fully automatized monitoring stations to be scattered all over Germany. Among other external positions he was the founder and chairman of the Bavarian Research Center for Knowledge Based Systems and served as a member of the board at the Excellence Cluster Cognition for Technical Systems.

Paul Bodesheim earned the degrees “Diplom-Informatiker” and “Dr.-Ing.” from the Friedrich Schiller University Jena, Germany, in years 2011 and 2017, respectively. He was a Research Associate at the Max Planck Institute for Biogeochemistry, Jena from 2015 to 2018. Since 2018, he is a Postdoctoral Researcher in the Computer Vision Group at the Friedrich Schiller University Jena, Germany, where he is currently a team leader for Computer Vision and Machine Learning. His research interests comprise novelty detection, open set recognition, and lifelong learning of visual object categories, including learning from small and imbalanced data, as well as fine-grained recognition with applications in biodiversity research.

Joachim Denzler earned the degrees “Diplom-Informatiker,” “Dr.-Ing.,” and “Habilitation” from the University of Erlangen, Germany, in years 1992, 1997, and 2003, respectively. Currently, he holds a position as Full Professor for Computer Science and is Head of the Computer Vision Group at the Friedrich Schiller University Jena, Germany. His research interests comprise the automatic analysis, fusion, and understanding of sensor data, especially development of methods for visual recognition tasks and dynamic scene analysis. He contributed in the area of active vision, 3D reconstruction, as well as object recognition and tracking. He is author and co-author of over 300 journal and conference papers as well as technical articles. He is a Member of IEEE, IEEE computer society, DAGM, and GI.

Morris Klasen received his BSc degree in 2020 at the Computer Science department of the University of Bonn. He is currently working in the German infrastructure project AMMOD (Automated Multisensor Station for Monitoring of Species Diversity). His research is specialized in 2D and 3D Multi Object Tracking of wildlife, RGB-D Video, and point-cloud processing techniques.

Timm Haucke received his BSc degree in 2019 at the Computer Science department of the University of Bonn. He is currently working in the German infrastructure project AMMOD (Automated Multisensor Station for Monitoring of Species Diversity) to classify and count animals in the wild. His research interest lies in the development of multimodal imaging systems and the implementation of multimodal image understanding methods in fauna and flora using deep learning.

Dimitri Korsch is a Research Associate in the Computer Vision Group at Friedrich Schiller University Jena, Germany. He received his BSc and MSc degree in IT-Systems Engineering from University of Potsdam in 2013 and 2016, respectively. His research interests include unsupervised machine learning, reinforcement learning as well as fine-grained visual categorization.

V. Steinhage is Head of the Intelligent Vision Systems working group of the Computer Science Department of Bonn University. He earned his PhD for his work on scene analysis at Bonn University. His professional focus is on artificial intelligence, machine learning, and computer vision. Beside other projects, he is involved in the German infrastructure project AMMOD (Automated Multisensor Station for Monitoring of Species Diversity), where he is heading a subproject that utilizes depth information to improve the detection and analysis of animals within wildlife monitoring.