DiversityScanner: Robotic discovery of small invertebrates with machine learning methods

Lorenz Wührl[1], Christian Pylatiuk[1,*], Matthias Giersch[1], Florian Lapp[1], Thomas von Rintelen[3], Michael Balke[4], Stefan Schmidt[4], Pierfilippo Cerretti[5], and Rudolf Meier[2,*]

[1]**Institute for Automation and Applied Informatics (IAI), Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany**
[2]**Department of Biological Science, National University of Singapore (NUS), Singapore**
[3]**Museum für Naturkunde, Leibniz-Institut für Evolutions- und Biodiversitätsforschung, Berlin, Germany**
[4]**SNSB – Zoologische Staatssammlung München, Munich, Germany**
[5]**Sapienza University of Rome, Rome, Italy**
[*]**Correspondence: *pylatiuk@kit.edu & Rudolf.Meier@mfn.berlin***

ABSTRACT

Invertebrate biodiversity remains poorly explored although it comprises much of the terrestrial animal biomass, more than 90% of the species-level diversity, supplies many ecosystem services. Increasing anthropogenic threads also require regular monitoring of invertebrate communities. The main obstacle is specimen- and species-rich samples consisting of thousands of small specimens. Traditional sorting techniques require manual handling based on morphology and are too slow and labor-intensive. Molecular techniques based on metabarcoding struggle with obtaining reliable abundance information. We here present a fully automated sorting robot for small specimens that are detected in the mixed sample using a convolutional neural network. Each specimen is then moved from the mixed sample to a well of a 96-well microplate in preparation for DNA barcoding. Prior to movement, the specimen is being photographed and assigned to 14 particularly common "classes" of insects in Malaise trap samples. The average assignment precision for the classes is 91.4 % (75-100 %) based on a preliminary neural network that is expected to improve further as more images are used for training. In order to obtain biomass information, the specimen images are also used to measure the specimen length and estimate the body volume. We outline how the "DiversityScanner" robot can be a key component for tackling and monitoring invertebrate diversity by generating large numbers of images that become training sets for species-, genus-, or family-level convolutional neural networks, once the imaged specimens are classified with DNA barcodes. The robot also allows for taxon-specific subsampling of large invertebrate samples. We conclude that the combination of automation, machine learning, and DNA barcoding has the potential to tackle invertebrate diversity at an unprecedented scale.

Keywords: biodiversity, classification, convolutional neural network, insects, machine learning

1 INTRODUCTION

Biodiversity science is currently at an inflection point. For decades, biodiversity declines had been mostly an academic concern although many biologists already predicted that these declines would eventually threaten whole ecosystems. Unfortunately, we are now at this stage which explains why the World Economic Forum considers biodiversity decline as one of the top three global risks based on likelihood and impact for the next 10 years [1]. This new urgency is also leading to a reassessment of research priorities in biodiversity science. Biologists have traditionally focused on charismatic taxa (e.g., vertebrates, vascular plants, butterflies) with a preference for endangered species because these taxa have more data (historical and current) and are favored by grantors and journals. However, with regard to quantitative arguments relating to ecosystem health, these taxon biases are poorly justified. For example, if one were to adopt a biomass point of view to terrestrial animal diversity, wild vertebrates would receive next to no attention because they only contribute very little biomass [2]. Indeed, endangered species would receive

the least attention because many are functionally extinct. The same conclusions is supported when one adopts a species diversity perspective. The largest number of multicellular species are fungi and invertebrates. The same groups would also be research priorities if one were to adopt a functional or an evolutionary point of view given that many fungal and invertebrate clades are much older and diverse than those taxa that contain most of the charismatic species. All these points of views suggest that it will be critical to have efficient tools for assessing and monitor non-charismatic taxa that provide numerous ecosystem services.

One major obstacle to pivoting attention towards those taxa that are important from a quantitative point of view are lack of biodiversity data on many of the relevant taxa. More than 10 years ago, Robert May [3] summarized the state-of-affairs as follows: "We are astonishingly ignorant about how many species are alive on earth today, and even more ignorant about how many we can lose (and) yet still maintain ecosystem services that humanity ultimately depends upon." He highlighted that the discovery and description of earth's biodiversity is one of the large, outstanding tasks in biology but he also anticipated that neglecting this task is perilous. Most of the undiscovered and undescribed diversity is in those invertebrate clades that are nowadays often called "dark taxa". Hartop et al. [4] recently defined these clades as those "for which the undescribed fauna is estimated to exceed the described fauna by at least one order of magnitude and the total diversity exceeds 1.000 species." They dominate many biodiversity samples and contribute most of the undescribed species-level diversity. Species discovery in these taxa is particularly difficult because it requires the sorting of thousands of usually very small specimens that need to be dissected for careful morphological examination.

Fortunately, there are three technical developments that promise relief. The first is already widely used. It is cost-effective DNA sequencing with 2nd and 3rd generational sequencing technologies, which have revolutionized microbial ecology, but can also be applied to invertebrate specimens [5]–[7]. In particular, portable nanopore sequencers by Oxford Nanopore Technologies are in the process of democratizing access to DNA sequence data [8]–[10]. However, the two remaining developments remain underutilized in biodiversity science. They are automation and data processing with neural networks. Currently, automation mostly exists in the form of pipetting robots in molecular laboratories, while data processing with neural networks is only widely used for the monitoring of charismatic species. Bulk invertebrate samples that include most of the undiscovered and unmonitoried biodiversity remain orphaned although thousands of samples are collected every day. They include plankton samples in marine biology, macroinvertebrate samples used for assessing freshwater quality, and insect samples obtained with pitfall- and Malaise traps [11]–[14]. Automation and data processing with artificial intelligence have the potential to greatly increase the amount of information that can be obtained from such samples [15]. The desirable end goal should be convolutional neural nets that use images (1) to identify the specimens to species, (2) provide specimen and species counts, (3) measure the biomass, and (4) compare the results to samples previously obtained from the same sites.

Manual sorting and identification of specimen- and species-rich invertebrate samples is time-consuming and prone to error. Processing with metabarcoding mostly yields presence/absence information but struggles with yielding abundance information and can be affected by taxonomic bias [16]. New systems are needed that yield comprehensive information. Fortunately,

computer-based identification systems for invertebrates are starting to yield promising results [17]–[19]. Particularly attractive are deep convolutional neural nets with transfer learning [15], but they require reasonably large sets of training images which are hard to obtain for invertebrates given that most species are undescribed and/or difficult to identify. It is here that robotics can have an important impact if robotic handling of specimens can be combined with taxonomic identifications based on DNA barcodes. First steps in this direction have been taken. One system was developed for processing macroinvertebrate samples that are routinely obtained for freshwater quality assessment. This system can size and identify stoneflies (Plecoptera) [20]. Another system focused on soil mesofauna [21]. However, these systems used a robotic arm which made them comparatively expensive. Many other insect sorting robots have been designed for more specific purposes. Some are for sorting mealworm larvae (*Tenebrio molitor*) and can separate healthy mealworm larvae from skins, feces, and dead worms. Another commercially available robot can sort mosquitoes [22] and is capable of distinguishing the gender of target species. However, all these machines lack the ability to recognize a wide variety of insect specimens preserved in ethanol. A machine that is closer to achieving this goal is the BIODISCOVER, a "robot-enabled image-based identification machine" by Arje et al. [23] which can identify ethanol-preserved specimens which, however, have to be fed into the machine manually one by one. After identification all specimen are returned into the same container.

We here describe a new system that overcomes some of these shortcomings. It recognizes insect specimens based on an overview image of a sample. Specimens below 3 mm body length are then imaged and moved into the wells of a 96-well microplate. We demonstrate that the images are of sufficient quality for training convolutional neural nets to common taxa. Furthermore, the images are used to derive length measurements and a coarse estimation of biomass based on specimen volume. Please note that we refer to the term "classification" in the machine learning context, as assigning objects (specimens) to different classes.

2 CONCEPT AND METHODS

The aim of the project was to develop an insect classification and sorting robot that is compact and that works reliably. It should also be easily reproducible, so that several systems can be set up and operated in parallel to allow high-throughput taxonomic identification. For this purpose, a design was developed that integrates as many standard parts as possible to ensure robustness. Furthermore, all connecting parts for the robot were designed to be produced by a standard 3D printer. The basic design with a cube-shaped frame and 3 linear drives with accurately positioning stepper motors is based on a zebrafish embryo handling robot [24]. The robot was equipped with two high-resolution cameras with customized lenses, suitable LED lighting and image recognition software. Furthermore, a transport system based on a suction pump was integrated to transfer detected insects into the wells of a standard 96-well microwell plate. Thus, the robot system can be divided into: (1) the Transport System, (2) the Image Acquisition System and (3) the Image Processing will be described in detail in the following. A free parts list and the assembly instructions is provided on request.

For the purpose of insect handling, a petri-dish with full-ethanol preserved insects is placed in the robot, which are then

classified, measured and sorted in a microwell plate. The setup provided for this purpose consists of a 50 x 50 x 50cm main frame, in which all components except the control panel are located. Figure 1 shows the sorting robot. The the x-, y-, and z-axis can be seen as well as the petri dish and the micro wellplate. For the operation of the robot by the user, a touch screen with graphical user interface (GUI) is mounted on the front side.

2.1 Transport System

The transportation system is based on a three-axes robot for transferring insects from a petri dish to a microwell plate and to position a camera for a detailed view (C2) of a single specimen. The transportation system with its three axes is illustrated in Figure 1.

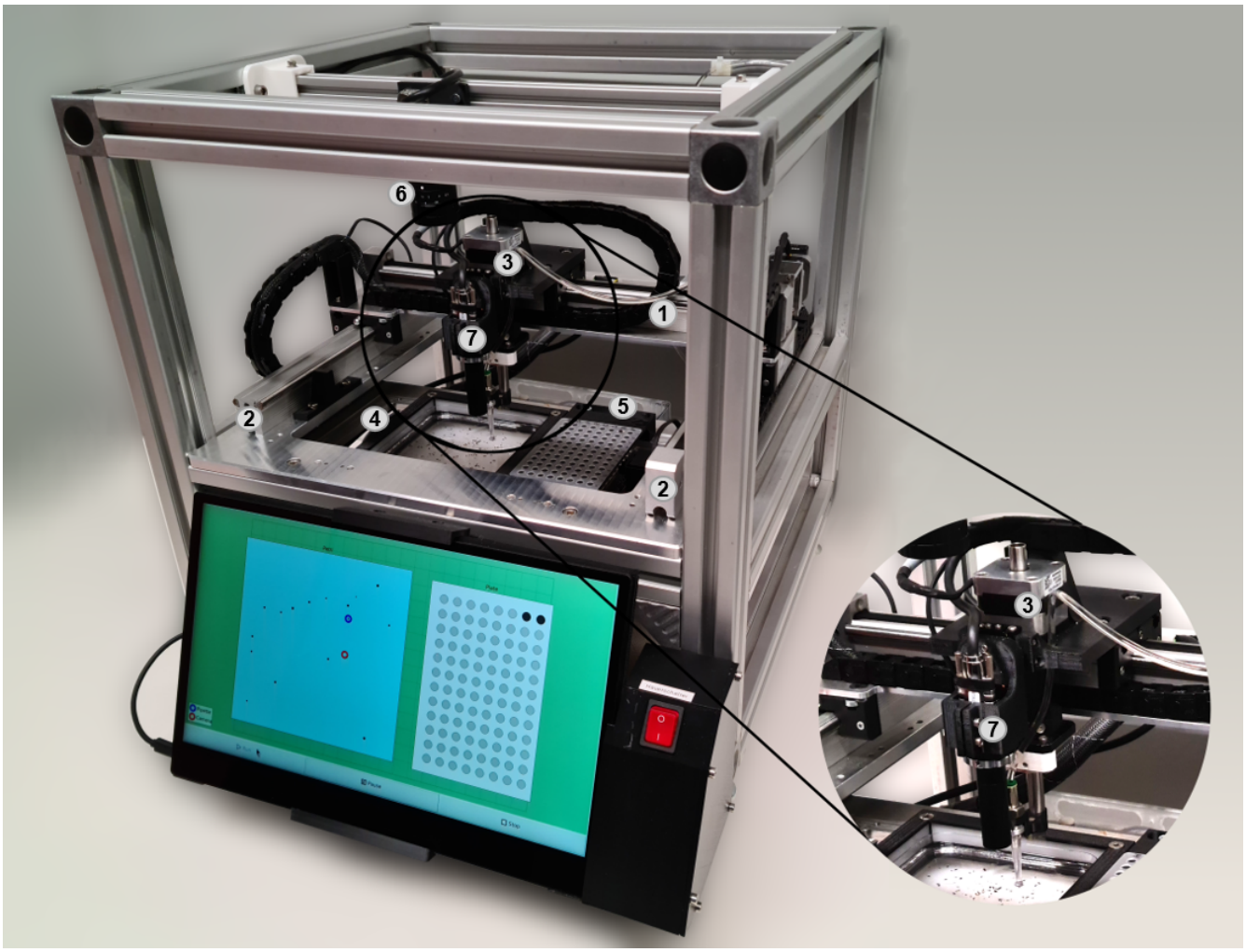


Figure 1. The DiversityScanner with 1: x-axis; 2: y-axis; 3: z-axis; 4: Petri dish; 5: Micro wellplate; 6: Overview camera (C1), 7: Detail camera (C2). The electronics box with Raspberry Pi, motor control unit, and the syringe pump are in the lower part of the sorting robot and therefore not visible in this view. The status of both, insect position determination and status of the sorting process are displayed on a touch screen, where the sorting process can also be started and stopped.

The x- and y-axes of the robot are realised by LEZ1 linear drives (Isel AG, Eichenzell, Germany) and connected to the outer frame of the robot at half height. Both linear drives are driven by high-precision stepper motors with little tolerance to ensure

good positioning accuracy. The y-axis is connected orthogonally to the shaft slide of the x-axis and is transported by it. The shaft slide of the y-axis transports both, the camera (C2) and the z-axis with the suction hose. In order to move the suction hose in the z-direction (=up and down) the z-axis is driven by a AR42H50 spindle drive with stepper motor (Nanotec Electronic GmbH & Co. KG, Feldkirchen, Germany). All three axes are controlled by a single TMCM-3110 motor controller (Trinamic, Hamburg, Germany) that allows for precise, fast and smooth movements of the axes. The motor controller was located in a box at the bottom of the robot along with other electronics, so that it is protected from water and ethanol droplets. The transport system is controlled by a Raspberry Pi single-board computer that was programmed in Python software, specially developed for the sorting robot. In order to pick up insects from a petri dish and discharge them in a well of a 96-well microplate a suction hose with a pipette tip is positioned by the transportation system. The hose is connected to a LA100 syringe pump (Landgraf Laborsysteme HLL GmbH, Langenhagen, Germany), that is also controlled by the Raspberry Pi. The sorting process is illustrated in Figure 2.

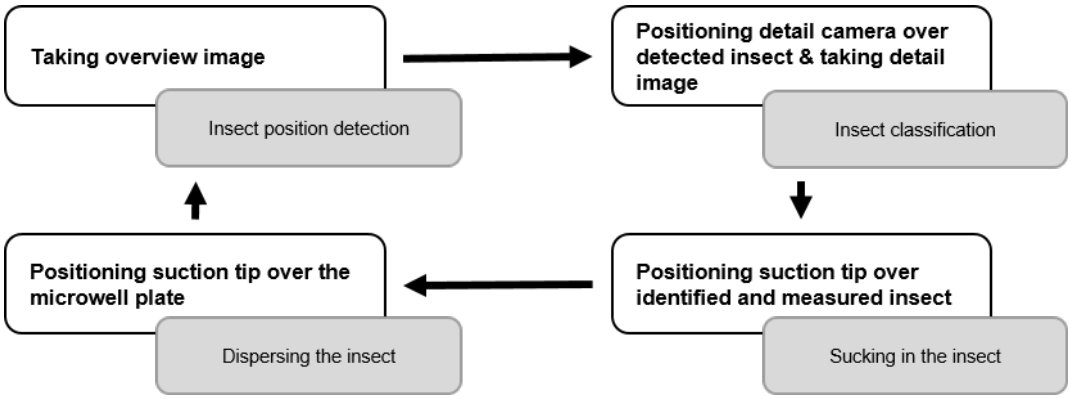


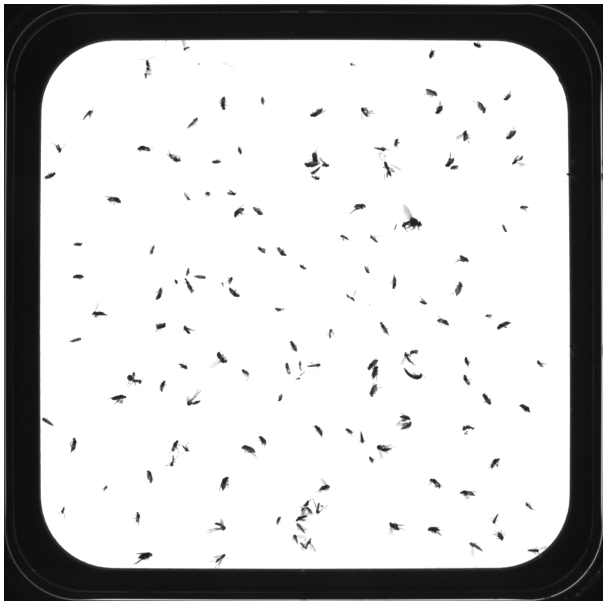
Figure 2. Process-chain for the classification and sorting process.

The sorting system includes two cameras with different lenses: the overview camera (C1) and the detailed view camera (C2). The first camera (C1) is a Ximea MQ042CG-CM camera with a CK12M1628S11 lens (Lensation GmbH, Karlsruhe, Germany) with a focal length of 16mm and an aperture of 2.8 is positioned directly above the petri dish to take a detailed overview image of all insects inside. This image is used for detecting insects and their position within the Petri dish for the sorting process. Figure 3 (a) shows an exemplary image of the overview camera.

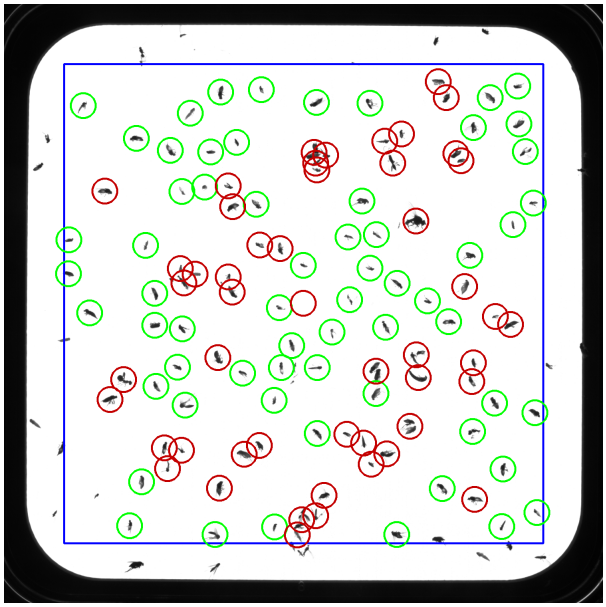
The second camera (C2) ia a Ximea MQ013CG-E2 camera with a telecentric Lensation TCST-10-40 lens with a magnification of 1x. This camera has to be moved by the x and y axes of the robot above the position of an insect to take a detailed image of it for classification, measuring and length determination. Figures 4 and 6 show examplary images from the detail camera.

2.2 Image Processing Software

Three different software algorithms are used: The first algorithm determines the position of each object within the square petri dish. The second one measures the length and volume of each insect. The third algorithm is based on an artificial neural net to classify insects into different classes.



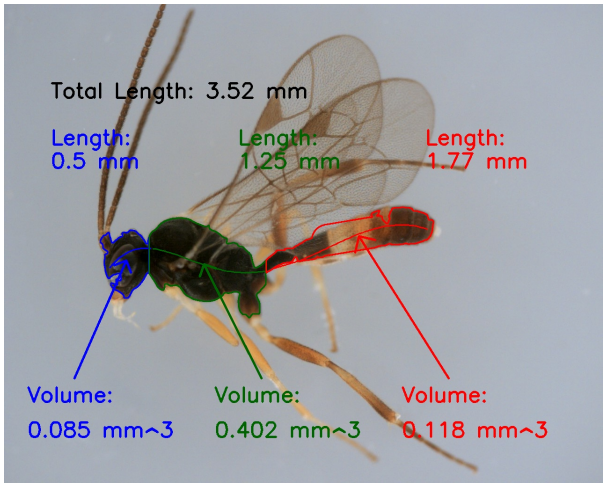
(a) The native image of the square Petri dish has a size of 120x120mm



(b) After image processing a blue line was drawn into the image, located 10 mm from the edge of the Petri dish to define the area in which the object positions are determined. Green circles represent the positions of detected objects that meet the conditions of size and sufficient distance from other objects.

Figure 3. Sample image obtained with the detail camera (C2) before and after processing.

(a) The native image has a size of 6.4x4.8mm and shows a Hymenoptera Ichneumonidae



(b) After image processing the length and volume of the head, mesosoma, metasoma and the total length are displayed

Figure 4. Specimen image obtained with the detail camera (C2) before and after processing.

Determination of Object Position: Most objects are insects, or parts of insects, but there can also be debris or other objects. After the overview image is taken, various image processing operations have to be performed to detect the objects: (1) A median filter removes noise from the image, (2) a conversion from a RGB-image to a gray scale image is performed, (3) an adaptive threshold filter segregates the objects and (4) a contour finder identifies the boundaries of all objects. Two conditions must be met for objects to be detected: first, their size must be within a specified interval, and second, the distance between an

object and neighbouring objects must exceed a minimum threshold value. If a cluster of objects is present, then the objects in the cluster fall below the specified minimum distance and are therefore not considered until they are separated. This ensures that only one single object is picked up during pipetting. Additionally, an accessible area within the petri dish has been defined that has a distance of ten millimeters from the edge to ensure that the insects can be reached (blue line in Figure 3) (b). Finally, all objects are color-coded. The coordinates of the detected objects are stored in a list, which is then used to control the position of the pipetting tip. After an object is removed, a new overview image is taken to determine the new coordinates of the objects, as they might have moved due to the pipetting of an object. This position identifying process continues until no more objects are detected or all wells of the 96-well microplate are filled with one insect each.

Object Dimensions: The length and volume of the insect body should be determined automatically and stored for estimating biomass. So, several image processing operations are then performed on each specimen image. First, the contour is determined using morphological operators. Only those surfaces are selected which have a minimum value. If more than one surface is found (e.g. two body parts of the same specimen separated by a light area), they are connected so that there is only one contour. Within this contour, points are placed randomly, which are used to create a regression. The more points are used, the more accurate the regression and thus the estimate of the insect length is. To find the dividing lines of the head, thorax and abdomen, straight lines are placed at right angles to and along the regression line. Only those points of a line are considered that lie within the contour in the process. Subsequently, the dividing line between the head and thorax or between the thorax and abdomen is determined by examining the changes in length. To estimate the volume, a straight line is drawn through each body part. After that additional perpendicular straight lines are drawn which must be within the body contour. Now the distance and length of the straight lines can be used to determine the volume slice by slice. The determined lengths and volumes of the individual body parts as well as the total length are displayed on the screen of the sorting robot and the measurements are stored. All operations are implemented using the free OpenCV program library (version 4.5.1) and the Python programming language (version 3.8.6). Please note that the results for volume estimation are only accurate if the body parts are rotationally symmetrical. This works relatively well for Hymenoptera, but yields less precise measurements for Diptera. A correct determination for all insect classes is only possible if a second detail camera were to take another image from a right-angle perspective. This is not yet implemented.

2.3 Insect Classification

In order to recognize different classes of insects and identify specimens to classes, machine learning algorithms were applied, based on convolutional neural nets (CNN).

Data Set: In a first trial only images from other image databases (e.g. Biodiversity of Singapore Image Database and Zoologische Staatssammlung München were used. However, the first classification results were poor which was presumably due to differences in the morphology of imaged specimens. We subsequently used our own images with the detailed camera for the training image data set. We used 5 Malaise trap samples from 3 different locations in Germany near the small towns and

villages of Rastatt, Kitzing and Framersbach and 2 from the Province of L'Aquila, Italy: Valle di Teve and Foresta Demaniale Chiarano-Sparvera. Thus, a mix of own images from different Malaise trap samples was used. The images for our target taxa were not equally distributed but reflected the abundances of each taxon in the Malaise trap samples. [11]. In total 4,325 color images in 15 classes were used for training, while 1,115 images were used for testing.

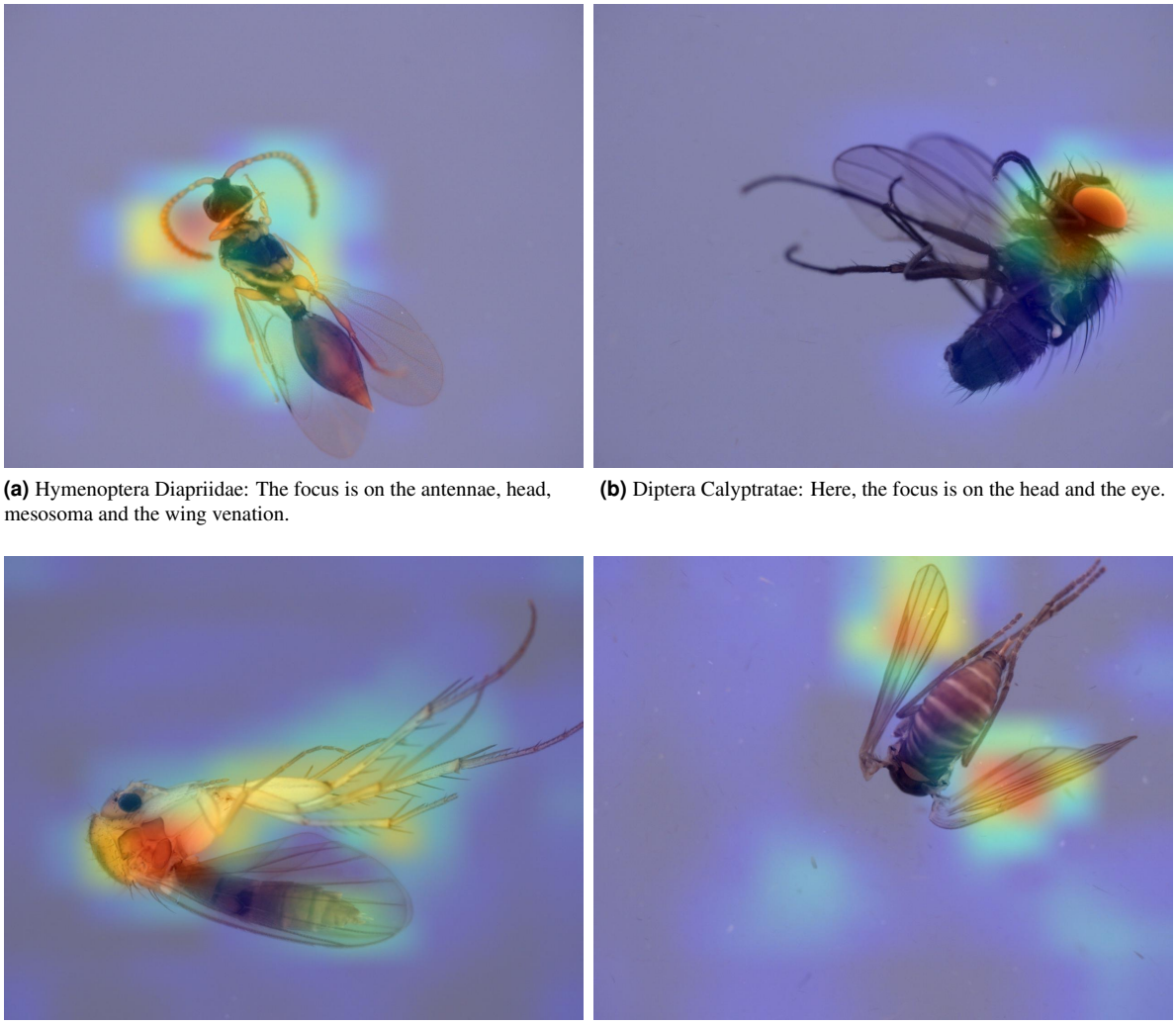
Class	Number of images	Class	Number of images
Diptera Acalyptratae	594	Diptera Calyptratae	79
Diptera Cecidomyiidae	467	Diptera Chironomidae	192
Diptera Dolichopodidae	140	Diptera Empididae & Hybotidae	446
Diptera Mycetophilidae & Keroplatidae	440	Diptera Phoridae	837
Diptera Psychodidae	129	Diptera Sciaridae	363
Hemiptera Cicadellidae	137	Hymenoptera Braconidae	113
Hymenoptera Diapriidae	255	Hymenoptera Ichneumonidae	133

Table 1. Classes and the number of images that were available for training, validation and testing.

Each sample contains a wide variety of insect taxa but only the common ones can be covered by the trained CNN. To be able to process images of insects that do not belong to any of the 14 classes, an additional residual class is created. This class consists of different taxa and images of body parts (mainly legs and wings), each of which has too few images for its own class. In total there are 693 images in this residual class.

Data Augmentation: Since the database consists of only 5,018 images for training the CNN, data augmentation was performed to increase both, the number of images and the invariance within a class. The following image processing operations were applied randomly to the images: rotation, width shift, height shift, shear, zoom, horizontal flip and fill mode nearest.

Network Architecture: As a base model for classification, the VGG19 architecture was used [25]. To apply transfer-learning, the model was initialized with pre-trained ImageNet weights and the last layer was removed. For the new classification layer, a global average pooling, a dense layer with 1,024 units and a relu-activation, and a linear layer with a dropout rate of 0.4 were added. For the final classification, a softmax and a L2-regularization with a value of 0.02 are applied. In total the model has about 20.5 million parameters and the input size of an image is 224x224 pixels. The number of nodes in the last layer corresponds to the number of classes in the experiment. For training, the parameters of the original model were frozen and only the classification layer was trained. Afterwards, the whole model was optimized, where training was applied to all layers. To get an impression whether the neural network selects the decisive features of an insect for classification, heat maps were generated. These class activation maps are obtained by a global average pooling layer. Figure 5 a-d show examples for four different specimen. The warmer the colour, the more crucial they are for classifying an insect. The network focused on these areas for classification.



(a) Hymenoptera Diapriidae: The focus is on the antennae, head, mesosoma and the wing venation.

(b) Diptera Calyptratae: Here, the focus is on the head and the eye.

(c) Diptera Keroplatidae and Mycetophilidae: The focus is on the thorax and the legs

(d) Diptera Psychodidae: In this class the focus is only on the wings

Figure 5. Heatmaps (Class activation maps) of four different insect classes.

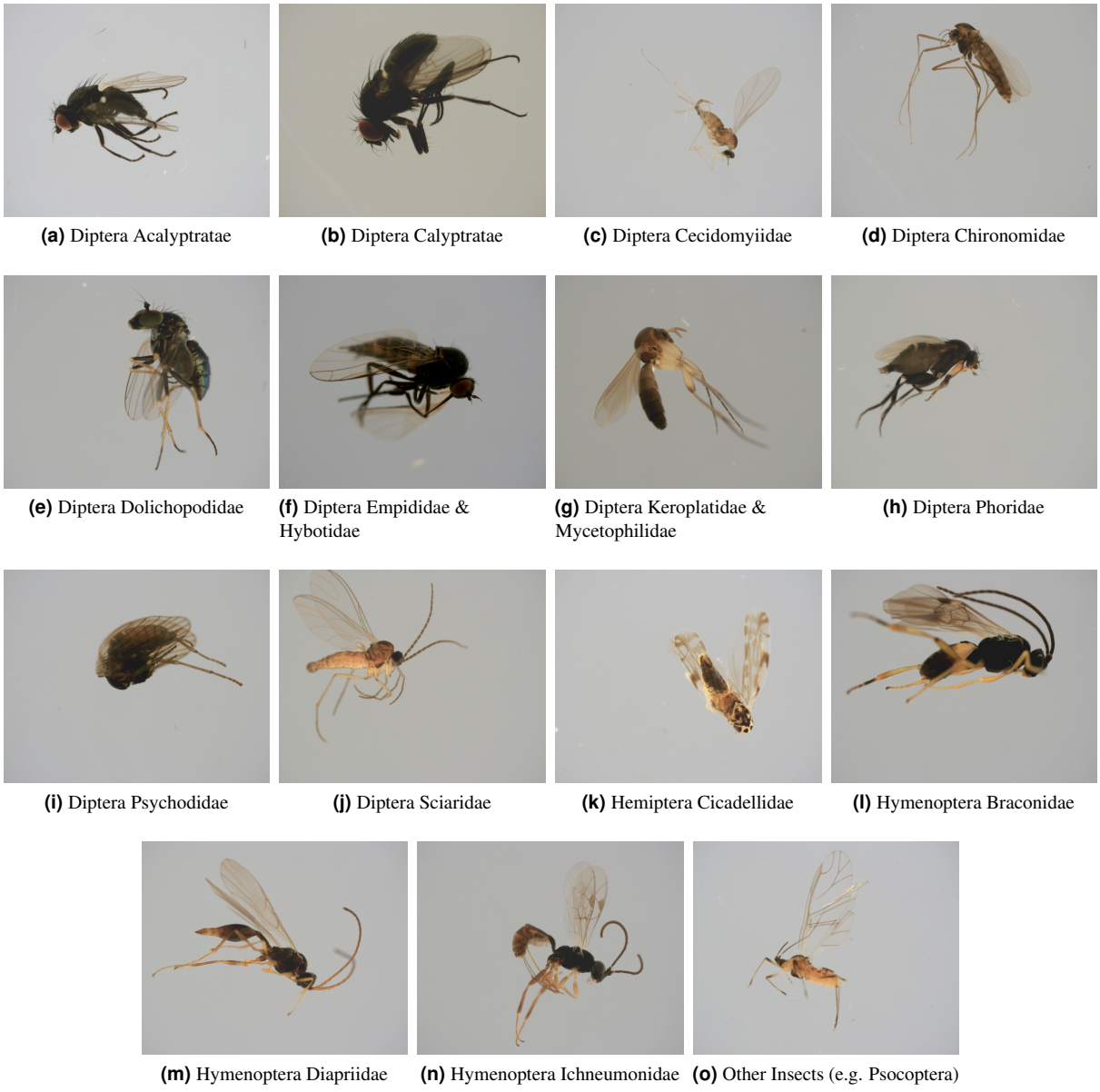
Setup: The model is implemented in Keras (version 2.4.3) based on Tensorflow (version 2.2.1) and all experiments are conducted in the Python programming language (version 3.8.6). The networks were trained on a single board computer (Nvidia, Santa Clara, California, USA) as well as on more powerful GPUs using the online tool Colabatory.

3 RESULTS

Currently, the sorting robot can pipette insects up to 3 mm length, as larger insects do not fit through the pipetting tip. Detected insects can be classified by the algorithm into 14 different classes of insects. All other insect classes and non-insect objects are combined in the class ”other”. The classification results are provided in Table 2. Examples of insects from the different classes are displayed in Figure 6 a-o. The overall working principles of the DiversityScanner are summarized by the following video clip: https://www.youtube.com/watch?v=ElJ5VSHa4OI.

Class (Taxon)	Result	Class (Taxon)	Result
Diptera Acalyptratae	91%	Diptera Psychodidae	89%
Diptera Calyptratae	83%	Diptera Sciaridae	92%
Diptera Cecidomyiidae	91%	Hemiptera Cicadellidae	100%
Diptera Chironomidae	97%	Hymenoptera Braconidae	82%
Diptera Dolichopodidae	86%	Hymenoptera Diapriidae	100%
Diptera Empididae & Hybotidae	87%	Hymenoptera Ichneumonidae	75%
Diptera Keroplatidae & Mycetophilidae	99%	Other	81%
Diptera Phoridae	97%		

Table 2. Classification results for the 15 classes. These classes include 14 one or more higher insect taxa and one class for all other objects and not specified insects.



(a) Diptera Acalyptratae **(b)** Diptera Calyptratae **(c)** Diptera Cecidomyiidae **(d)** Diptera Chironomidae

(e) Diptera Dolichopodidae **(f)** Diptera Empididae & Hybotidae **(g)** Diptera Keroplatidae & Mycetophilidae **(h)** Diptera Phoridae

(i) Diptera Psychodidae **(j)** Diptera Sciaridae **(k)** Hemiptera Cicadellidae **(l)** Hymenoptera Braconidae

(m) Hymenoptera Diapriidae **(n)** Hymenoptera Ichneumonidae **(o)** Other Insects (e.g. Psocoptera)

Figure 6. Examples of 14 classes and one class of other insects.

The best classification result was achieved for the classes of Hymenoptera Diapriidae and Hemiptera Cicadellidae, where all insects were correctly classified, whereas insects of the class Diptera Dolichopodidae had the lowest correct classification rate. Two different automated sorting processes are possible: Either one insect after the other can be classified and sorted until the last well of the 96-well microplate is filled, or only insects of a predefined class are pipetted into the well plates until no insect of this class can be found.

4 DISCUSSION

The use of CNNs for the identification of charismatic species is starting to be routine [26]–[28]. However, these methods are largely unavailable for small invertebrates although they comprise most of the multicellular animal species and contribute many ecosystem services. The main problem is not the availability of invertebrate samples, but the lack of CNNs which cannot be trained because there are few sets of training images. We believe that the best strategy for changing this undesirable situation is by combining automated imaging with DNA barcoding. Each "DiversityScanner" robot can process several invertebrate samples per day. Each contains thousands of specimens that can be imaged with minimal manual labour. After imaging, the specimens are moved into microplates for DNA barcoding. Once barcoded, the images can be re-labeled with approximately species-level identifications given that most animal species have species-specific barcodes, or they can be assigned to family- or genus-level based on DNA sequence similarities. Common species, genera, and families rapidly acquire sufficiently large sets of images that can then be used for training CNNs. Indeed, for the most common "classes" of insects in Malaise traps, we already had enough images for creating such networks after partially imaging only five Malaise trap samples.

Some biologists doubt that CNNs will be sufficiently powerful to yield species-level identifications for closely related species and we agree that it remains unclear whether species-level identifications can be achieved [15], [19]. However, we believe that the main limitation is not the CNN but the image quality and orientation of the insects. Fortunately, these limitations can be overcome by using high-quality cameras and obtaining large numbers of specimen images in different orientations. This is particularly straightforward once specimens have been pre-sorted to putative species based on DNA barcodes. As illustrated by the BIODISCOVER robot, large numbers of images can be obtained rapidly for the same set of specimens by inserting them into a cuvette; i.e., one could obtain a sufficiently large number of training images even for fairly rare species. Once the CNNs have been trained for a sufficiently large number of species, the DiversityScanner could identify most specimens in routine samples based on images. DNA barcoding would only be needed for those specimens that are not identifiable based on visual information. These are more likely to belong to rare and new species so that the DiversityScanner would also become a powerful tool for discovering new species in samples. This ability would be particularly important in the 21st century because new species continue to arrive at well-characterized sampling sites. Some of these species recently shifted their distribution in response to climate change while others may be new anthropogenic introductions. For both it would be desirable to have an early-warning system based on automated workflows.

We designed the automatic classification and sorting robot for smaller insects, because they are particularly abundant. The design of the robot focused on reproducibility and low-cost (<5,000€), so that many robots can sort a large number of insects simultaneously. This makes the robot an attractive alternative to manual identification and sorting. After modification, the DiversityScanner will also be suitable for many additional purposes. For example, larger specimens could be handled by modifying the suction tip diameters or installing a gripper with a sensor-based feedback system that ensures that the specimens are not damaged. A particularly attractive modification would also be the ability to subsample a sample. For example, some invertebrate samples are dominated by a few taxa whose exhaustive treatment may not be needed for monitoring. The robot could then be instructed to only fill/identify 2-3 microplates' worth of specimens for these taxa. Conversely, the user could specify that only certain taxa should be moved to microplates or different taxa should be moved to different microplates. The latter would be particularly useful if the specimens are supposed to be barcoded using different molecular markers or taxon-specific DNA extraction or PCR recipes should be used. Many additional modifications are conceivable. For example, only specimens belonging to one gender could be selected given that often only the morphology of one sex is species-specific.

Thus, we believe that robots like the DiversityScanner have the potential to solve some of the problems that were outlined by Robert May. Biodiversity discovery and monitoring can be greatly expedited and accelerated, in particular for the "dark taxa" that have been largely ignored in the past, because of the problems associated with their handling and identification. Of course, the DiversityScanner can only address some of the challenges. For example, newly discovered species will still have to be described and described species matched to types. Even when all the species have been described or identified, we will still know very little about the ecological roles that the species play within ecosystems. Fortunately, molecular approaches to diet analysis and life history stage matching can help [29], [30], but ecosystems routinely consist of thousands of species. This means that automation and data analysis with the tools of AI will become increasingly important.

Acknowledgments We would like to specially thank Daniel Moser and Stefan Vollmannshauser for their support with manufacturing the mechanical parts and helping us with connecting the electronic circuits. Mr Leshon Lee prepared the video documenting the working principles of the DiversityScanner. Funding was provided by the Center for Integrative Biodiversity Discovery at the Museum für Naturkunde Berlin.

Author Contributions Conceptualization: R.M., T.v.R., L.W. and C.P.; writing original draft preparation: L.W., R.M. and M.G.; writing review and editing: C.P., R.M., S.S., P.C., M.B. and T.v.R.; visualization: L.W. and M.G. and S.S.; supervision: C.P., R.M. and T.v.R.; funding acquisition: C.P., T.v.R. and R.M.; L.W. and C.P. contributed equally. All authors have read and agreed to the published version of the manuscript.

Data Availability Statement All image data that were used for training and testing are accessible at the media repository of the Museum für Naturkunde Berlin..

All files for printing the robot parts and the software code are accessible at the repository of the Open Science Framework.

Supplementary Materials Video

REFERENCES

[1] T. World Economic Forum's Global Risk Initiative, *The global risks report 2020*, `http://www3.weforum.org/docs/WEF_Global_Risk_Report_2020.pdf`, 2020.

[2] Y. M. Bar-On, R. Phillips, and R. Milo, "The biomass distribution on earth," *Proceedings of the National Academy of Sciences*, vol. 115, no. 25, pp. 6506–6511, 2018.

[3] R. M. May, "Why worry about how many species and their loss?" *PLoS Biol*, vol. 9, no. 8, e1001130, 2011.

[4] E. Hartop, A. Srivathsan, F. Ronquist, and R. Meier, "Large-scale integrative taxonomy (lit): Resolving the data conundrum for dark taxa," *bioRxiv*, 2021.

[5] P. D. Hebert, T. W. Braukmann, S. W. Prosser, S. Ratnasingham, J. R. DeWaard, N. V. Ivanova, D. H. Janzen, W. Hallwachs, S. Naik, J. E. Sones, *et al.*, "A sequel to sanger: Amplicon sequencing that scales," *BMC genomics*, vol. 19, no. 1, pp. 1–14, 2018.

[6] A. Srivathsan, E. Hartop, J. Puniamoorthy, W. T. Lee, S. N. Kutty, O. Kurina, and R. Meier, "Rapid, large-scale species discovery in hyperdiverse taxa using 1d minion sequencing," *BMC biology*, vol. 17, no. 1, pp. 1–20, 2019, https://doi.org/10.1186/s12915-019-0706-9.

[7] W. Y. Wang, A. Srivathsan, M. Foo, S. K. Yamane, and R. Meier, "Sorting specimen-rich invertebrate samples with cost-effective ngs barcodes: Validating a reverse workflow for specimen processing," *Molecular ecology resources*, vol. 18, no. 3, pp. 490–501, 2018.

[8] A. Srivathsan, L. Lee, K. Katoh, E. Hartop, S. N. Kutty, J. Wong, D. Yeo, and R. Meier, "Minion barcodes: Biodiversity discovery and identification by everyone, for everyone," *bioRxiv*, 2021.

[9] M. Watsa, G. A. Erkenswick, A. Pomerantz, and S. Prost, "Portable sequencing as a teaching tool in conservation and biodiversity research," *PLoS biology*, vol. 18, no. 4, e3000667, 2020.

[10] A. Pomerantz, N. Peñafiel, A. Arteaga, L. Bustamante, F. Pichardo, L. A. Coloma, C. L. Barrio-Amorós, D. Salazar-Valenzuela, and S. Prost, "Real-time dna barcoding in a rainforest using nanopore sequencing: Opportunities for rapid biodiversity assessments and local capacity building," *GigaScience*, vol. 7, no. 4, giy033, 2018.

[11] D. Karlsson, E. Hartop, M. Forshage, M. Jaschhof, and F. Ronquist, "The swedish malaise trap project: A 15 year retrospective on a countrywide insect inventory," *Biodiversity Data Journal*, vol. 8, 2020, https://doi.org/10.3897/BDJ.8.e47255.

[12] B. V. Brown, A. Borkent, P. H. Adler, D. de Souza Amorim, K. Barber, D. Bickel, S. Boucher, S. E. Brooks, J. Burger, Z. L. Burington, *et al.*, "Comprehensive inventory of true flies (diptera) at a tropical site," *Communications biology*, vol. 1, no. 1, pp. 1–8, 2018, https://doi.org/10.1038/s42003-018-0022-x.

[13] A. Borkent, B. V. Brown, *et al.*, "How to inventory tropical flies (diptera)-one of the megadiverse orders of insects," *Zootaxa*, vol. 3949, no. 3, pp. 301–322, 2015.

[14] B. V. Brown, "Malaise trap catches and the crisis in neotropical dipterology," *American Entomologist*, vol. 51, no. 3, pp. 180–183, 2005, https://doi.org/10.1093/ae/51.3.180.

[15] J. Ärje, J. Raitoharju, A. Iosifidis, V. Tirronen, K. Meissner, M. Gabbouj, S. Kiranyaz, and S. Kärkkäinen, "Human experts vs. machines in taxa recognition," *Signal Processing: Image Communication*, vol. 87, p. 115 917, 2020, https://doi.org/10.1016/j.image.2020.115917.

[16] T. J. Creedy, W. S. Ng, and A. P. Vogler, "Toward accurate species-level metabarcoding of arthropod communities from the tropical forest canopy," *Ecology and evolution*, vol. 9, no. 6, pp. 3105–3116, 2019.

[17] P Perre, F. A. Faria, L. Jorge, A Rocha, R. d. S. Torres, M. Souza-Filho, T. Lewinsohn, and R. A. Zucchi, "Toward an automated identification of anastrepha fruit flies in the fraterculus group (diptera, tephritidae)," *Neotropical entomology*, vol. 45, no. 5, pp. 554–558, 2016, https://doi.org/10.1007/s13744-016-0403-0.

[18] L. Feng, B. Bhanu, and J. Heraty, "A software system for automated identification and retrieval of moth images based on wing attributes," *Pattern Recognition*, vol. 51, pp. 225–241, 2016, https://doi.org/10.1016/j.patcog.2015.09.012.

[19] A. Knyshov, S. Hoang, and C. Weirauch, "Pretrained convolutional neural networks perform well in a challenging test case: Identification of plant bugs (hemiptera: Miridae) using a small number of training images," *Insect Systematics and Diversity*, vol. 5, no. 2, p. 3, 2021.

[20] M. Sarpola, R. Paasch, E. Mortensen, T. Dietterich, D. Lytle, A. Moldenke, and L. Shapiro, "An aquatic insect imaging system to automate insect classification," *Transactions of the ASABE*, vol. 51, no. 6, pp. 2217–2225, 2008, https://doi.org/10.13031/2013.25375.

[21] M. A. Chamblin, R. Paasch, D. Lytle, A. Moldenke, L. Shapiro, and T. Dietterich, "Design of an automated system for imaging and sorting soil mesofauna," *Biological Engineering Transactions*, vol. 4, no. 1, pp. 17–41, 2011, https://doi.org/10.13031/2013.37174.

[22] H. Lepek, T. Nave, Y. Fleischmann, R. Eisenberg, B. E. Karlin, and I. Tirosh, *Method for sex sorting of mosquitoes and apparatus therefor*, US Patent App. 16/479,648, 2020.

[23] J. Ärje, C. Melvad, M. R. Jeppesen, S. A. Madsen, J. Raitoharju, M. S. Rasmussen, A. Iosifidis, V. Tirronen, M. Gabbouj, K. Meissner, *et al.*, "Automatic image-based identification and biomass estimation of invertebrates," *Methods in Ecology and Evolution*, vol. 11, no. 8, pp. 922–931, 2020, https://doi.org/10.1111/2041-210X.13428.

[24] A. Pfriem, C. Pylatiuk, R. Alshut, B. Ziegener, S. Schulz, and G. Bretthauer, "A modular, low-cost robot for zebrafish handling," in *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, https://doi.org/10.1109/EMBC.2012.6346097, 2012, pp. 980–983.

[25] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[26] M. A. Tabak, M. S. Norouzzadeh, D. W. Wolfson, S. J. Sweeney, K. C. VerCauteren, N. P. Snow, J. M. Halseth, P. A. Di Salvo, J. S. Lewis, M. D. White, *et al.*, "Machine learning to classify animal species in camera trap images: Applications in ecology," *Methods in Ecology and Evolution*, vol. 10, no. 4, pp. 585–590, 2019.

[27] D. Stowell, M. D. Wood, H. Pamuła, Y. Stylianou, and H. Glotin, "Automatic acoustic detection of birds through deep learning: The first bird audio detection challenge," *Methods in Ecology and Evolution*, vol. 10, no. 3, pp. 368–380, 2019.

[28] A. J. Fairbrass, M. Firman, C. Williams, G. J. Brostow, H. Titheridge, and K. E. Jones, "Citynet—deep learning tools for urban ecoacoustic assessment," *Methods in ecology and evolution*, vol. 10, no. 2, pp. 186–197, 2019.

[29] D. Yeo, J. Puniamoorthy, R. W. J. Ngiam, and R. Meier, "Towards holomorphology in entomology: Rapid and cost-effective adult–larva matching using ngs barcodes," *Systematic entomology*, vol. 43, no. 4, pp. 678–691, 2018.

[30] A. Srivathsan, N. Nagarajan, and R. Meier, "Boosting natural history research via metagenomic clean-up of crowdsourced feces," *PLoS biology*, vol. 17, no. 11, e3000517, 2019.