







Towards a Taxonomy Machine: A Training Set of 5.6 Million Arthropod Images

Dirk Steinke ^{1,2,*}, Sujeevan Ratnasingham ^{1,2}, Jireh Agda ¹, Hamzah Ait Boutou ¹, Isaiah C. H. Box ¹, Mary Boyle ¹, Dean Chan ¹, Corey Feng ¹, Scott C. Lowe ³, Jaclyn T. A. McKeown ¹, Joschka McLeod ¹, Alan Sanchez ¹, Ian Smith ¹, Spencer Walker ¹, Catherine Y.-Y. Wei ¹ and Paul D. N. Hebert ^{1,2}

¹ Centre for Biodiversity Genomics, University of Guelph, 50 Stone Road East, Guelph, ON N1G 2W1, Canada; sratnasi@uoguelph.ca (S.R.); agdaj@uoguelph.ca (J.A.); aitboutou@gmail.com (H.A.B.); mboyle14@uoguelph.ca (M.B.); mrsixcount@gmail.com (D.C.); cfeng01@uoguelph.ca (C.F.); mccormij@uoguelph.ca (J.T.A.M.); joschka.mcleod@gmail.com (J.M.); asanch04@uoguelph.ca (A.S.); ismith@uoguelph.ca (I.S.); swalke19@uoguelph.ca (S.W.); cwei@uoguelph.ca (C.Y.-Y.W.); phebert@uoguelph.ca (P.D.N.H.)

² Department of Integrative Biology, University of Guelph, 50 Stone Road East, Guelph, ON N1G 2W1, Canada

³ Vector Institute, Toronto, ON M5G 1M1, Canada; scott.lowe@vectorinstitute.ai

* Correspondence: dsteinke@uoguelph.ca

Abstract: The taxonomic identification of organisms from images is an active research area within the machine learning community. Current algorithms are very effective for object recognition and discrimination, but they require extensive training datasets to generate reliable assignments. This study releases 5.6 million images with representatives from 10 arthropod classes and 26 insect orders. All images were taken using a Keyence VHX-7000 Digital Microscope system with an automatic stage to permit high-resolution (4K) microphotography. Providing phenotypic data for 324,000 species derived from 48 countries, this release represents, by far, the largest dataset of standardized arthropod images. As such, this dataset is well suited for testing the efficacy of machine learning algorithms for identifying specimens into higher taxonomic categories.

Keywords: insects; machine learning; object recognition; image-based classification; biodiversity



Citation: Steinke, D.; Ratnasingham, S.; Agda, J.; Ait Boutou, H.; Box, I.C.H.; Boyle, M.; Chan, D.; Feng, C.; Lowe, S.C.; McKeown, J.T.A.; et al. Towards a Taxonomy Machine: A Training Set of 5.6 Million Arthropod Images. *Data* **2024**, *9*, 122. <https://doi.org/10.3390/data9110122>

Received: 18 September 2024

Revised: 10 October 2024

Accepted: 22 October 2024

Published: 25 October 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Summary

The identification of organisms is a fundamental part of recognizing and describing biodiversity. The development of automated methods, which can identify specimens without involving taxonomists, is critical given the taxonomic impediment [1,2]. One effective solution involves the adoption of identification systems based on the analysis of sequence variation in short, standardized DNA regions [3]. In addition, digitization initiatives at major natural history collections [4] and imaging linked to large DNA barcoding projects [5,6] are providing the basis for image-based identification systems driven by machine learning algorithms.

Images can be used to build classification systems capable of identifying species [7–9]. Various methods and datasets have been proposed to advance image-based identifications for arthropods [10,11]. Most past work has considered arthropod classification from (1) the context of integrated pest management [12–14], (2) as part of crowd-sourced citizen science efforts such as iNaturalist [15,16], or (3) interpreting data acquired by camera traps [17,18].

Machine learning algorithms, especially convolutional artificial neural networks and their variants, have emerged as the most effective method for object recognition and detection [10,19,20]. However, datasets of sufficient size for properly training these models are scarce [20–23], as creating them is labour-intensive and expensive. On the other hand, they are urgently needed to avoid performance issues and limitations for vision models resulting from insufficient training [24]. For instance, most available biodiversity image

datasets have poor representation of arthropod species even though these represent a large fraction of eukaryotic life. In recent years, substantial efforts have been invested into the digitization of arthropod natural history collections, which have imaged both individual specimens [25–27] and entire drawers of specimens [6,28,29]. This work provides training material for classification algorithms, but the images are structurally uniform. Specimens are usually mounted and therefore captured in a single standard orientation (mostly dorsal view), limiting their application in less regimented contexts [24]. In addition, there are biases in selecting insects, which translate into the availability of specimens for digitization or images gathered by community science [30,31].

In this study, we present a dataset of 5.6 million arthropod images gathered during a large-scale DNA barcoding project. As specimens were not placed in a standard orientation before photography, any machine learning-based object detection must interpret specimens in varying orientations to classify them taxonomically, an approach that requires massive training sets.

2. Data Generation

The present dataset was generated as part of an ongoing effort to build a global DNA barcode reference library for terrestrial arthropods [32]. The workflow involves placing each small (<5 mm) specimen into a well of a round-bottom 96-well microplate before DNA extraction. The specimens are individually photographed at 4K resolution in plate format using a Keyence VHX-7000 Digital Microscope system (Keyence, Osaka, Japan) with a fully integrated head and an automatic stage (Figure 1). The setup uses an adjustable Keyence illumination adapter to ensure uniform light conditions and a scanning stage equipped with a custom-engineered mount that holds each plate (Figure 1, Supplementary File S1). This system can capture 95 images within 15 min by controlling stage movements in X-Y coordinates. It also has the capability to automatically control the height of the stage with a precision of 0.1 μm . By moving the lens throughout the different focal planes of each specimen, the VHX system captures every pixel that comes into focus at each level and combines them into a single, fully focused image. For all the images, the system was set to a brightness of 27.8 ms, and colours were set to R1.7 G1.0 B2.19.



Figure 1. Keyence VHX-7000 Digital Microscope system. The inset shows a microplate within the custom-engineered mount.

The resulting images are packaged with a Python script (Supplementary File S2) and uploaded onto the Barcode of Life Data Systems V4 (BOLD) [33] where they are automatically associated with individual specimen records. BOLD is a cloud-based data storage and analysis platform that supports the assembly and use of DNA barcode data. Users can generate and populate the specimen records. BOLD enables its users to upload and store many collaterals such as specimen-associated taxonomic labels, images, and DNA sequences [33]. A backup copy of each image is subsequently transferred to a third-party cloud service using an automated script (Supplementary File S3).

3. Data Description

The present dataset includes 5,675,731 images, mostly of terrestrial arthropods. Associated metadata include information on provenance and taxonomy (Supplementary File S4). The dataset includes 1.13 million images from a previous release [20] together with 4.54M images generated from 2019 to 2022. The size of each image is 2880×2160 pixels, creating an average file size of 17.9 MB for a TIFF file and 1.88 MB for a JPEG file. Figure 2 shows an array of 95 images taken with the Keyence system (the empty space at the lower right corresponds to an empty control well in the microplate).



Figure 2. Panel of example images taken with the Keyence setup. The empty space at the lower right corresponds to an empty control well in the microplate.

3.1. Geographic Coverage

The dataset contains images for specimens from 1698 sites in 48 countries (Figure 3A). Most specimens were derived from Costa Rica (62%), followed by South Africa (6%), United States (5%), and Thailand (3%) (Figure 3B, Supplementary File S4). Most specimens were collected using Malaise traps, but about 1500 were captured using plankton and dip nets [34].

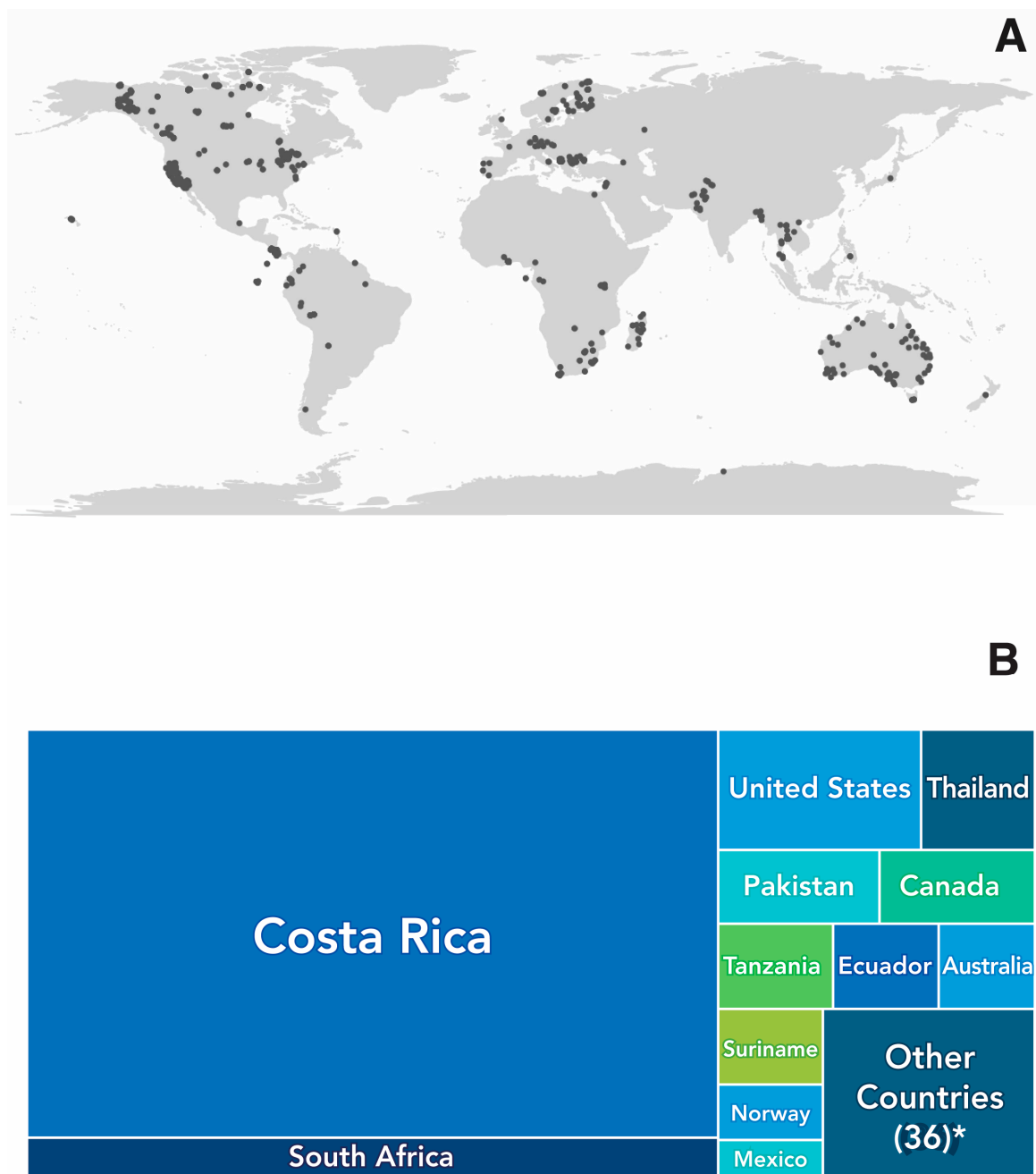


Figure 3. (A) Sampling locations for photographed specimens. (B) Tree map of countries of origin. * a complete list of countries can be found in Supplementary File S5.

3.2. Taxonomic Coverage

Most of the images show individuals from the class Insecta (98%) (Figure 4A). Figure 4B shows the distribution of insect orders within the dataset. A very high proportion of the complete set of specimens has taxonomic assignments at the ordinal (99.4%) and family (85.9%) levels, but only 20.3% possess a generic assignment. Just 7.6% of the specimens (430,036) have a Linnean species designation, but 90.6% ($N = 5,143,970$) possess a Barcode Index Number (BIN) assignment. The BIN system [35] is a key feature of BOLD. It employs an algorithm that combines single linkage and Markov clustering to group DNA barcode sequences into Operational Taxonomic Units that represent good species proxies [35]. A total of 324,427 such BINs are represented in this dataset.

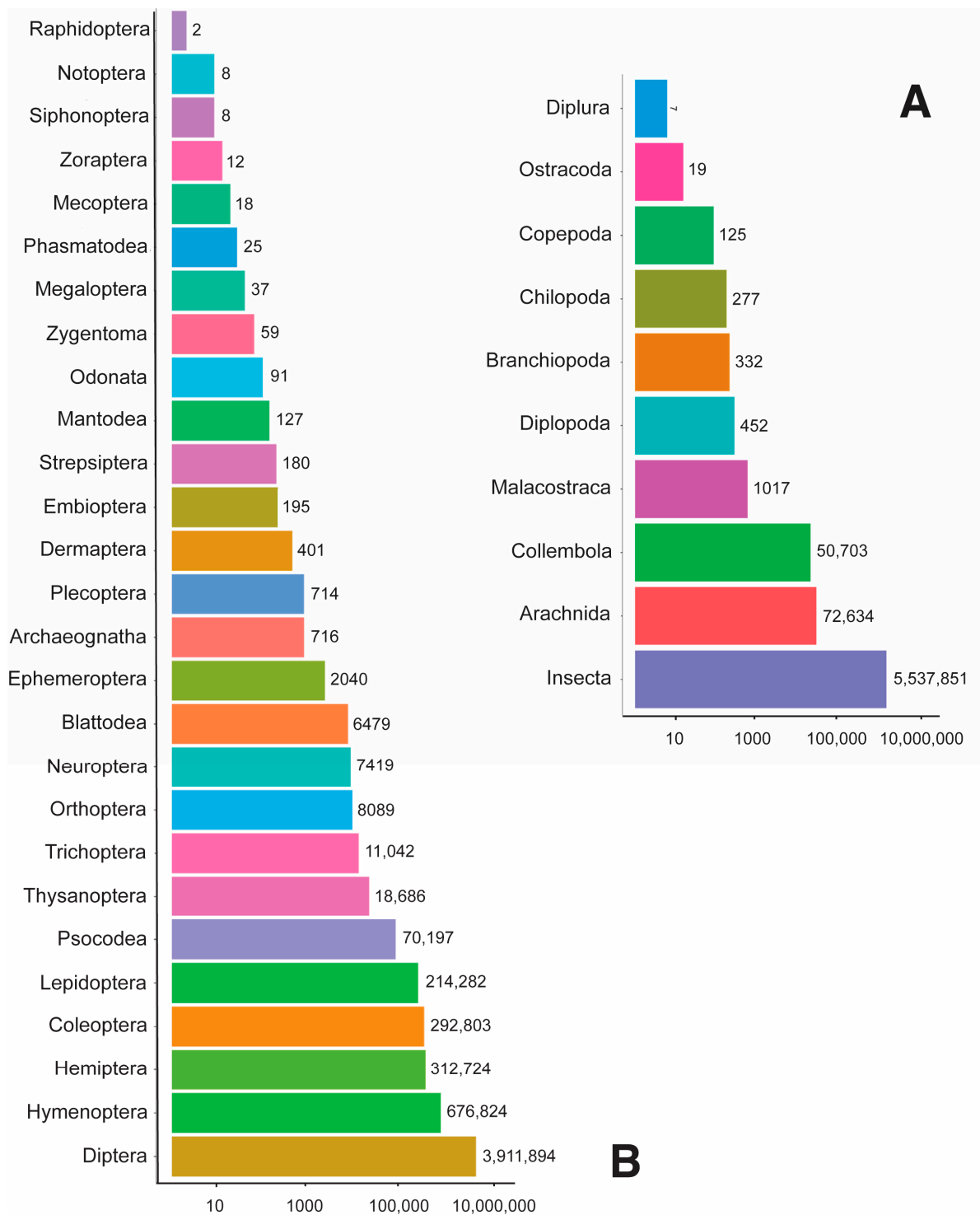


Figure 4. Log-scale plot showing coverage for 10 arthropod classes (A) and 27 insect orders (B).

4. Conclusions

This dataset of 5.6 million images representing over 320,000 species of arthropods, mostly insects, is the largest of its kind. It is also unique in the geographic distribution of the photographed samples. As such, it represents a dataset that should aid the development of machine learning algorithms for object recognition and classification despite a considerable class imbalance (69% Diptera). Data from a previous release were successfully used for several classification tasks across arthropod orders [20]. In addition, a larger subset of this dataset has been combined with DNA sequence information to develop and benchmark

multimodal classification experiments [36]. These images can also be used to estimate biomass [37,38] or even abundance when entire communities are digitized [39,40]. Each image in the dataset is also available as an element of the collateral data for a barcode record on BOLD [33] providing support for taxonomic assignment and to enable direct visual comparisons between individuals.

Generating high quality images at the described rate can be difficult with any automated system given the manifold differences in shape and size of specimens [6]. Some large and some very small individuals might be outside the standard stacking depth, which can result in an out-of-focus image. Furthermore, mistakes during the placement of individuals onto the microplates can lead to accidental imaging of an empty well or, rarely, the placement of more than one individual into one well.

By using three Keyence VHX-7000 systems for 50 h per week, the Centre for Biodiversity Genomics generates three million images per year. The deployment of Keyence systems at a few core facilities could readily generate ten million images per year, allowing rapid growth in the training sets required to hone AI-enabled identification systems, including routines that automatically flag instances of lower quality as described above.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/data9110122/s1>: Supplementary File S1: 3D-print file for custom plate mount; Supplementary File S2: Python script to pack and upload images; Supplementary File S3: Python script for file transfer; Supplementary File S4: List of countries of origin for photographed specimens; Supplementary File S5: Metadata field definitions.

Author Contributions: Conceptualization, P.D.N.H., S.R. and D.S.; software and database, S.R., J.A., H.A.B., D.C., A.S., S.W. and C.Y.-Y.W.; formal analysis, D.S.; data generation and curation, I.C.H.B., M.B., C.F., J.T.A.M., J.M., I.S. and S.C.L.; writing—original draft preparation, D.S.; writing—review and editing, P.D.N.H., S.R. and J.T.A.M.; visualization, D.S.; project administration, P.D.N.H., S.R. and D.S.; funding acquisition, P.D.N.H. All authors have read and agreed to the published version of the manuscript.

Funding: This study was enabled by awards to PDNH from the Ontario Ministry of Economic Development, Job Creation and Trade, the Canada Foundation for Innovation, Genome Canada and Ontario Genomics (OGI-208), the New Frontiers in Research Fund (NFRFT-2020-00073), Polar Knowledge Canada under the Northern Science and Technology Programme (NST-1819-0039), the Walder Foundation, and the Guanacaste Dry Forest Conservation Fund and by a grant from the Canada First Research Excellence Fund to the University of Guelph’s “Food From Thought” research program (Project 000054).

Data Availability Statement: The dataset and a separate metadata file in *.tsv format (Supplementary File S5 contains definitions of all metadata fields) can be accessed at <https://biodiversitygenomics.net/5M-insects/> (accessed on 20 October 2024).

Acknowledgments: All images were generated at the Centre for Biodiversity Genomics as one component of an integrated program aiming to construct a DNA barcode reference library. As such, the assembly of this dataset benefitted from the contributions of all CBG staff as well as colleagues around the world who are aiding in the progress in this work. We thank Suzanne Bateson for graphics support.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Gaston, K.J.; O’Neill, M.A. Automated species identification: Why not? *Philos. Trans. R. Soc. Lond. B* **2004**, *359*, 655–667. [[CrossRef](#)] [[PubMed](#)]
2. Godfray, H.C.J. Linnaeus in the information age. *Nature* **2007**, *446*, 259–260. [[CrossRef](#)] [[PubMed](#)]
3. Hebert, P.D.N.; Cywinska, A.; Ball, S.L.; deWaard, J.R. Biological identifications through DNA barcodes. *Proc. R. Soc. B* **2003**, *270*, 312–321. [[CrossRef](#)] [[PubMed](#)]
4. Blagoderov, V.; Kitching, I.J.; Livermore, L.; Simonsen, T.J.; Smith, V.S. No specimen left behind: Industrial scale digitization of natural history collections. *ZooKeys* **2012**, *209*, 133–146. [[CrossRef](#)] [[PubMed](#)]

5. Hebert, P.D.N.; Ratnasingham, S.; Zakharov, E.V.; Telfer, A.C.; Levesque-Beaudin, V.; Milton, M.A.; Pedersen, S.; Janetta, P.; de Waard, J.R. Counting animal species with DNA barcodes: Canadian insects. *Philos. Trans. R. Soc. Lond. B* **2016**, *371*, 20150333. [[CrossRef](#)]
6. deWaard, J.R.; Ratnasingham, S.; Zakharov, E.V.; Borisenko, A.V.; Steinke, D.; Telfer, A.C.; Perez, K.H.J.; Sones, J.E.; Young, M.R.; Levesque-Beaudin, V.; et al. A reference library for Canadian invertebrates with 1.5 million barcodes, voucher specimens, and DNA samples. *Sci. Data* **2019**, *6*, 308. [[CrossRef](#)]
7. Farnsworth, E.J.; Chu, M.; Kress, J.; Neill, A.K.; Best, J.H.; Pickering, J.; Stevenson, R.D.; Courtney, G.W.; VanDyk, J.K.; Ellison, A.M. Next-generation field guides. *BioScience* **2013**, *63*, 891–899. [[CrossRef](#)]
8. Seeland, M.; Rzanny, M.; Alaqraa, N.; Wäldchen, J.; Mäder, P. Plant species classification using flower images—A comparative study of local feature representations. *PLoS ONE* **2017**, *12*, e0170629. [[CrossRef](#)]
9. Wäldchen, J.; Mäder, P. Machine learning for image based species identification. *Methods Ecol. Evol.* **2018**, *9*, 2216–2225. [[CrossRef](#)]
10. Martineau, C.; Conte, D.; Raveaux, R.; Arnault, I.; Munier, D.; Venturini, G. A survey on image-based insect classification. *Pattern Recognit.* **2017**, *65*, 273–284. [[CrossRef](#)]
11. De Cesaro, T., Jr.; Rider, R. Automatic identification of insects from digital images: A survey. *Comput. Electron. Agric.* **2020**, *178*, 105784. [[CrossRef](#)]
12. da Silveira, F.A.G.; Castela, E.; Astolfi, G.; Bessada Costa, A.; Paraguassu Amorim, W. Performance analysis of YOLOv3 for real-time detection of pests in soybeans. In Proceedings of the Intelligent Systems: 10th Brazilian Conference, BRACIS 2021, Virtual Event, 29 November–3 December 2021; Proceedings, Part II. Springer: Berlin/Heidelberg, Germany; pp. 265–279. [[CrossRef](#)]
13. Li, W.; Zheng, T.; Yang, Z.; Li, M.; Sun, C.; Yang, X. Classification and detection of insects from field images using deep learning for smart pest management: A systematic review. *Ecol. Inform.* **2021**, *66*, 101460. [[CrossRef](#)]
14. Xing, S.; Lee, H.J. Crop pests and diseases recognition using DANet with TLDP. *Comput. Electron. Agric.* **2022**, *199*, 107144. [[CrossRef](#)]
15. van Horn, G.; Mac Aodha, O.; Song, Y.; Cui, Y.; Sun, C.; Shepard, A.; Adam, H.; Perona, P.; Belongie, S. The iNaturalist species classification and detection dataset. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 8769–8778. [[CrossRef](#)]
16. van Horn, G.; Cole, E.; Beery, S.; Wilber, K.; Belongie, S.; Mac Aodha, O. Benchmarking representation learning for natural world image collections. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 12884–12893. [[CrossRef](#)]
17. Schneider, S.; Taylor, G.W.; Linquist, S.; Kremer, S.C. Past, present and future approaches using computer vision for animal re-identification from camera trap data. *Methods Ecol. Evol.* **2018**, *10*, 461–470. [[CrossRef](#)]
18. Bothmann, L.; Wimmer, L.; Charrakh, O.; Werber, T.; Edelhoff, H.; Peters, W.; Nguyen, H.; Benjammin, C.; Menzel, A. Automated wildlife image classification: An active learning tool for ecological applications. *Ecol. Inform.* **2023**, *77*, 102231. [[CrossRef](#)]
19. Ding, W.; Taylor, G.W. Automatic moth detection from trap images for pest management. *Comput. Electron. Agric.* **2016**, *123*, 17–28. [[CrossRef](#)]
20. Gharaee, Z.; Gong, Z.; Pellegrino, N.; Zarubiieva, I.; Haurum, J.B.; Lowe, S.C.; McKeown, J.T.A.; Ho, C.C.Y.; McLeod, J.; Wei, Y.C.; et al. A step towards worldwide biodiversity assessment: The BIOSCAN-1M insect dataset. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: New York, NY, USA, 2023; Volume 37. [[CrossRef](#)]
21. Sun, Y.; Liu, X.; Yuan, M.; Ren, L.; Wang, J.; Chen, Z. Automatic in-trap pest detection using learning for pheromone-based *Dendroctonus valens* monitoring. *Biosyst. Eng.* **2018**, *176*, 140–150. [[CrossRef](#)]
22. Wu, X.; Zhan, C.; Lai, Y.-K.; Cheng, M.-M.; Yang, J. IP102: A large-scale benchmark dataset for insect pest recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 8787–8796. [[CrossRef](#)]
23. Badirli, S.; Akata, Z.; Mohler, G.; Picard, C.; Dundar, M. Fine-Grained Zero-Shot learning with DNA as side information. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: New York, NY, USA, 2021; Volume 34. [[CrossRef](#)]
24. Yang, H.-P.; Ma, C.-S.; Wen, H.; Zhan, Q.-B.; Wang, X.-L. A tool for developing an automatic insect identification system based on wing outlines. *Sci. Rep.* **2015**, *5*, 12786. [[CrossRef](#)]
25. Heerlien, M.; van Leusen, J.; Schnörr, S.; de Jong-Kole, S.; Raes, R.; van Hulsen, K. The natural history production line: An industrial approach to the digitization of scientific collections. *ACM J. Comput. Cult. Herit.* **2015**, *8*, 3. [[CrossRef](#)]
26. Ströbel, B.; Schmelzle, S.; Blüthgen, N.; Heethoff, M. An automated device for the digitization and 3D modelling of insects, combining extended-depth-of-field and all-side multi-view imaging. *ZooKeys* **2018**, *759*, 1–27. [[CrossRef](#)]
27. Tegelberg, R.; Kahanpää, J.; Karppinen, J.; Mononen, T.; Wu, Z.; Saarenmaa, H. Mass digitization of individual pinned insects using conveyor-driven imaging. In Proceedings of the 2017 IEEE 13th International Conference on e-Science (e-Science), Auckland, New Zealand, 24–27 October 2017; pp. 523–527. [[CrossRef](#)]
28. Mantle, B.L.; Salle, J.L.; Fisher, N. Whole-drawer imaging for digital management and curation of a large entomological collection. *ZooKeys* **2012**, *209*, 147–163. [[CrossRef](#)] [[PubMed](#)]
29. Holovachov, O.; Zatushevsky, A.; Shydlovsky, I. Whole-drawer imaging of entomological collections: Benefits, limitations, and alternative applications. *J. Conserv. Mus. Stud.* **2014**, *12*, 9. [[CrossRef](#)]
30. Small, E. The new Noah’s ark: Beautiful and useful species only. Part 2. The chosen species. *Biodiversity* **2012**, *12*, 37–53. [[CrossRef](#)]

31. Leandro, C.; Jay-Robert, P.; Vergnes, A. Bias and perspectives in insect conservation: A European scale analysis. *Biol. Conserv.* **2017**, *215*, 213–224. [[CrossRef](#)]
32. Hobern, D.; Hebert, P.D.N. BIOSCAN—Revealing Eukaryote Diversity, Dynamics, and Interactions. *Biodivers. Inf. Sci. Stand.* **2019**, *3*, e37333. [[CrossRef](#)]
33. Ratnasingham, S.; Wei, C.; Chan, D.; Agda, J.; Agda, J.; Ballesteros-Mejia, L.; Ait Boutou, H.; El Bastami, Z.M.; Ma, E.; Manjunath, R.; et al. BOLD v4: A Centralized Bioinformatics Platform for DNA-Based Biodiversity Data. In *DNA Barcoding: Methods and Protocols*; Springer: New York, NY, USA, 2024; Chapter 26; pp. 403–441.
34. Nowosad, D.S.J.; Hogg, I.D.; Cottenie, K.; Lear, C.; Elliott, T.A.; deWaard, J.R.; Steinke, D.; Adamowicz, S.J. High diversity of freshwater invertebrates on Inuinnait Nuna, the Canadian Arctic, revealed using mitochondrial DNA barcodes. *Polar Biol.* **2024**. [[CrossRef](#)]
35. Ratnasingham, S.; Hebert, P.D.N. A DNA-based registry for all animal species: The Barcode Index Number (BIN) System. *PLoS ONE* **2013**, *8*, e66213. [[CrossRef](#)]
36. Gharaee, Z.; Lowe, S.C.; Gong, Z.M.; Arias, P.M.; Pellegrino, N.; Wang, A.T.; Haurum, J.B.; Zarubiieva, I.; Kari, L.; Steinke, D.; et al. BIOSCAN-5M: A Multimodal Dataset for Insect Biodiversity. *arXiv* **2024**, arXiv:2406.12723. [[CrossRef](#)]
37. Ärje, J.; Melvad, C.; Jeppesen, M.R.; Madsen, S.A.; Raitoharju, J.; Rasmussen, M.S.; Iosifidis, A.; Tirronen, V.; Meissner, K.; Gabbouj, M.; et al. Automatic image-based identification and biomass estimation of invertebrates. *Mol. Ecol. Resour.* **2021**, *11*, 922–931. [[CrossRef](#)]
38. Wührl, L.; Pylatiuk, C.; Giersch, M.; Lapp, F.; von Rintelen, T.; Balke, M.; Schmidt, S.; Cerretti, P.; Meier, R. Diversityscanner: Robotic handling of small invertebrates with machine learning methods. *Mol. Ecol. Resour.* **2022**, *22*, 1626–1638. [[CrossRef](#)]
39. Schneider, S.; Tayler, G.W.; Kremer, S.C.; Burgess, P.; McGroarty, J.; Mitsui, K.; Zhuang, A.; deWaard, J.R.; Fryxell, J.M. Bulk arthropod abundance, biomass and diversity estimation using deep learning for computer vision. *Methods Ecol. Evol.* **2021**, *13*, 346–357. [[CrossRef](#)]
40. Schneider, S.; Taylor, G.W.; Kremer, S.C.; Fryxell, J.M. Getting the bugs out of AI: Advancing ecological research on arthropods through computer vision. *Ecol. Lett.* **2023**, *26*, 1247–1258. [[CrossRef](#)] [[PubMed](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.