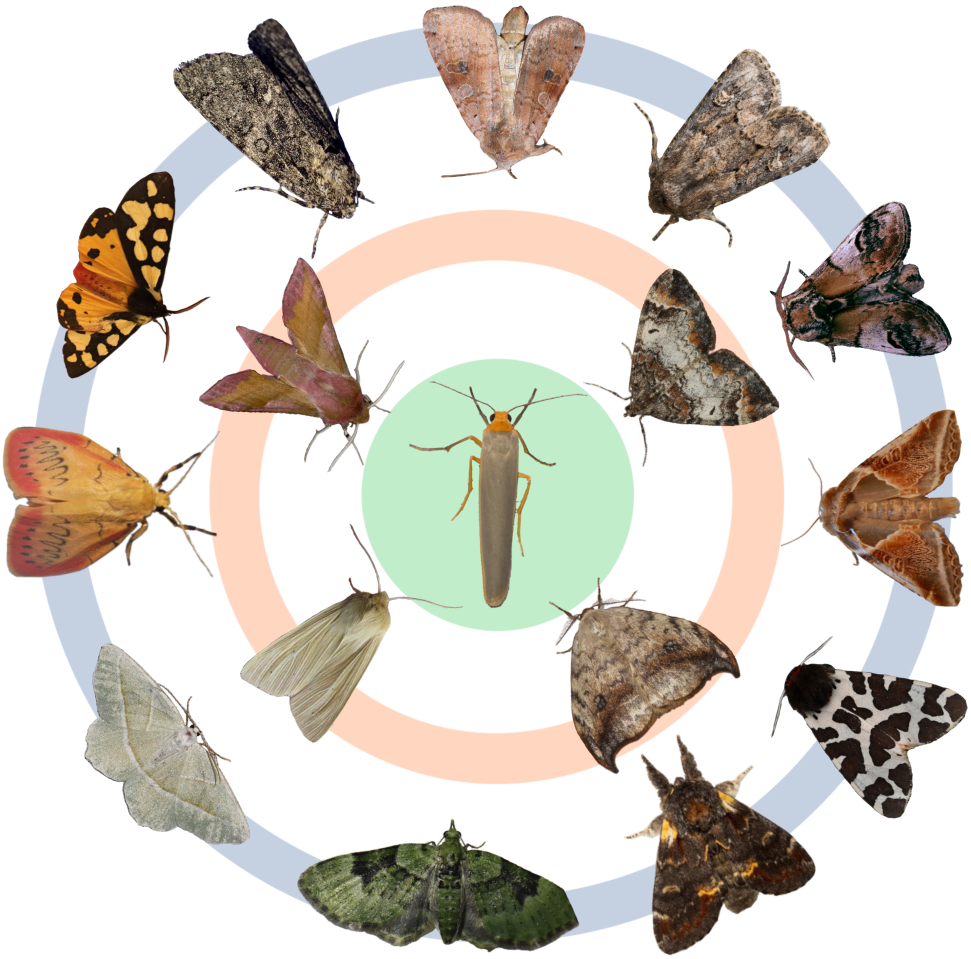
Scalable Biodiversity Monitoring & Analysis with Computer Vision



DEPARTMENT OF ECOSCIENCE

PhD Mid Term Progress Report

Title	Scalable Biodiversity Monitoring & Analysis with Computer Vision
Student	Asger Svenning
Study no.	201907234
Main Supervisor	Toke Thomas Høye
Co-Supervisor	Kim Bjerge
Programme	Ecoscience

SOLIDUM PETIT IN PROFUNDIS · UNIVERSITAS ARHUSIENSIS ·

Contents

1 Introduction

Conservation of Nature is not a recent phenomenon in the modern context[3], however biodiversity and the global 'biodiversity crisis' is a recent phenomenon in the public mind and serious policy[35]. This is particularly evident in the danish context with the highly proclaimed 'green tripartite' that has garnered wide political and public support through its' focus on climate, environment, nature, and biodiversity. At the European and global scale the increased focus and priority to conservation and restoration takes the form of multilateral treaties such as CBD[1][44] and the EU Nature Restoration Law[34], which in turn are translated into both large-scale and local, biodiversity and nature, projects and programmes. As countries and organizations continue to increase their efforts in the continuation and implementation of these treaties, projects, and programmes, there also follows increasing calls for accountability and transparency around the current state and progress towards the stated goals of halting and reversing biodiversity decline[6],[32],[36],[37]. The proposed conservation and restoration tools are many and often controversial due to conflicting priorities and unresolved scientific questions[11],[24],[41], which only underscores the necessity of effective and scalable methods for quantifying the current and continuing state of biodiversity. Accommodating the ambitious goals and ending the biodiversity crisis must then necessitate the development and implementation of truly *scalable biodiversity monitoring and analyses*.

1.1 Project

As the title of my PhD project 'Scalable Biodiversity Monitoring & Analysis with Computer Vision' suggests, the project is aimed towards research which furthers large-scale biodiversity monitoring schemes such as the Biodiversa+ pilot projects on automated biodiversity monitoring stations and invasive alien species monitoring[2] or the demonstration sites in the Horizon Europe project, MAMBO. In practice, the project is focused towards image-based monitoring of insects, particularly moths and other light-attracted nocturnal insects, with data deriving mainly from the insect camera trap AMI (Automated Monitoring of Insects)[36] deployed in projects across Denmark and Europe.

In the first part of the project, my research have been focused broadly on finding solutions to image-processing, including both detection, localization and classification, of AMI images in Denmark and Europe, as well as in the XPRIZE Rainforest competition in Brazil. Solutions, which are not only powerful and generally applicable, but also scalable and efficient to be directly or with small effort implemented in large-scale projects with millions of images.

[1] The Convention on Biological Diversity
[2] https://www.biodiversa.eu/biodiversity-monitoring/pilots/

As such the focal problems of the project are aligned with the axes of scale in biodiversity monitoring; spatiotemporal, taxonomic, abundance, ecology, and data:

- The spatiotemporal, taxonomic and abundance axes of scale are 'simple'; comprehensive biodiversity monitoring must cover large areas and the full breadth of species richness, which in turn requires detection of huge numbers of individuals to ensure high coverage.
- The ecological axis of scale is less intuitive, but can be explained concisely in the terms that monitoring must take into account and cover not only the geographical space over the monitoring area, but also the ecological space defined through ecosystems, ecological niches, and species communities.
- The data axis of scale in biodiversity monitoring is unique and distinct taken in comparison with the latter axes, as it is the only scale which is not inherently defined in terms of biology, but instead through computer science, statistics and mathematics. It can be understood simply as an 'umbrella-axis' which covers the diversity of data sources such as citizen science, government monitoring, commercial monitoring and scientific enquiry, as well as the size of the datasets which are collected and analyzed as biodiversity monitoring programmes are upscaled.

In the project these axes of scale are tackled through three core pillars (1) machine learning, (2) biology, and (3) deployment. Through machine learning, I use data from research projects such as Biodiversa+ and MAMBO that are derived 'in-house', as well as publically available images primarily through the Global Biodiversity Information Facility (GBIF). So far, I have used these images successfully to develop a state-of-the-art insect detection system in my first major first-author paper `flatbug`: *'A General Method for Detection and Segmentation of Terrestrial Arthropods in Images'*[42], which is currently available as a preprint and in the submission process. I have also used these images to develop taxonomical hierarchical species classification models, both for the XPRIZE Rainforest competition in Brazil, as well as for an upcoming paper, which I am co-leading with postdoc Guillaume Mougeot in our lab. Through biology I use the results from the aforementioned projects and methods, as well as other biodiversity monitoring projects such as NOVANA, to derive biodiversity and ecology insights and inference applicable to both furthering our understanding of nature and ecology, as well as biodiversity policy and management. Lastly, with deployment I take part in the use and development of deployable models and pipelines, which are directly integrated into large-scale biodiversity monitoring programmes and projects. This can be exemplified with the XPRIZE Rainforest competition, wherein I participated as a member of an international team of 50+ scientists and practitioners in the ETH BiodivX team, and was the major contributor to the image-processing pipeline which processed images from three remotely deployed and controlled canopy camera light traps in the field in the Amazon rainforest near Manaus, Brazil.

2 Project Aim

The aim of my PhD project *"Scalable Biodiversity Monitoring & Analysis"* can be summarized as:

> "Advancing biodiversity monitoring through automated computer vision techniques with a focus on classifying and localizing insect species from images."

To that end the first major sub-aim of the project is to develop deep learning models and pipelines capable of processing images from biodiversity monitoring projects and programmes at a sufficient quality for downstream ecological inference and decision-making. This sub-aim is two-pronged in nature, including research on novel methods capable of handling the axes of scale mentioned in Section 1.1, and the development and implementation of both these and existing methods. The second major sub-aim of the project is use the results generated by the models and pipelines from the first sub-aim, and perform biologically relevant analyses at scale. Mirroring the first sub-aim, this second sub-aim also includes both developing new or improved methods capable of handling data of the scale produced by the biodiversity monitoring pipelines, and applying these methods for ecology and biodiversity research, and management questions.

In summary my PhD project aims to advance the application of automated image-based biodiversity monitoring schemes—particularly of insects—from raw images to the downstream biodiversity questions.

3 Methods, Results & Conclusions

Throughout the first two years of the project I have taken part in a multitude of different projects across an array of different research scopes and groups.

- December 2023-ongoing:
  `flatbug`—*'A General Method for Detection and Segmentation of Terrestrial Arthropods in Images'*
- May 2024-November 2024:
  XPRIZE Rainforest with ETH BiodivX
- March 2025-ongoing:
  Hierarchical Classification with Citizen Science Data
- December 2024-ongoing:[†]
  NOVANA trend analysis

[†]Related research projects that are not a part of the core PhD project.

3.1 `flatbug`: A General Method for Detection and Segmentation of Terrestrial Arthropods in Images

The project on the method coined as `flatbug` and currently available in preprint *'A General Method for Detection and Segmentation of Terrestrial Arthropods in Images'*[3][42] under submission at the journal 'Methods in Ecology and Evolution', was motivated by the growing number of projects within Toke T. Høye and Quentin Geissmanns' groups which produced images of insects, and the lack of a general-purpose insect detection model for these types of images. The project was kickstarted by Quentin Geissmann as a sort of successor to his earlier work in Geissmann *et al.* [(14)] and as an improvement on the widespread use of `YOLO`[4] + `SAHI`[1] for flexible-size inference. I spent the first 4-5 months of 2024 solving the various technical hurdles regarding the inference pipeline, integration of `YOLOv8`, and collection and annotation of data. Meanwhile, I was brought on to produce the needed image-processing pipeline for the XPRIZE Rainforest competion finale described in Section 3.3, the first real project where the—at the time—prototype of `flatbug` was implemented and used. Then, in the fall of 2024 I polishing the `flatbug` method and setting up and executing the experiments we defined for the paper, following which the paper was written, published as a preprint and submitted to Methods in Ecology and Evolution in the spring of 2025.

3.2 Hierarchical Classification with Citizen Science Data

Developing robust, accurate species-classification models with a broad taxonomic scope is challenging for several reasons: (1) limited training data, (2) extreme class imbalance, and (3) an extremely large number of classes (species). The advent of large-scale citizen science platforms (e.g., iNaturalist, Pl@ntNet, Arter.dk, Merlin) and the availability of their observation data via the Global Biodiversity Information Facility (GBIF) have high potential to address the first challenge. However, these resources by themselves do not solve the second and third problems. One research avenue proposed in the literature is the use of hierarchical computer vision models. This idea has been explored by Jia Deng *et al.* [(21)], Bjerge *et al.* [(2)], Luca Bertinetto *et al.* [(26)], among others, but its specific benefits for species classification—particularly for insects and even more so for moths—remain underexplored.

With this motivation in mind, I began working on hierarchical classification in the spring of 2024 as a side project, and I presented my preliminary results at the biennial Nordic Society Oikos (NSO) conference. I also deployed these preliminary methods as part of the image-processing pipeline I built for the XPRIZE Rainforest competition finals in Brazil. In that pipeline, I developed a hierarchical model covering $\sim 8,000$ Brazilian insect species.[5] Subsequently, during discussions on developing large-scale species classification models, the research was narrowed to focus on a key question: Do hierarchical deep-learning image classifiers achieve higher overall accuracy on citizen science images than traditional 'flat' models? And do they produce more domain-adaptable models with fewer taxonomically catastrophic misclassifications?

In this project, we compare three types of models: a traditional 'flat' classifier, a hierarchical multi-label classifier, and a novel hierarchical multi-layer classifier. We evaluate each model using both standard classification metrics and custom domain-specific metrics. Specifically, a 'flat' classifier is trained using one-hot labels for each leaf class (species) in the taxonomy. A hierarchical multi-label classifier is trained on concatenated one-hot label vectors for each taxonomic level (e.g., species, genus, family), essentially corresponding to independent classification heads for each taxonomic level. Finally, the hierarchical multi-layer classifier is a novel architecture[6] that outputs a logit vector for the leaf classes (species) like a flat classifier, but also computes additional outputs for each higher taxonomic level (genus, family). These additional output layers aggregate

[3]The manuscript is attached as part of this progress report in Section 7.
[4]Over the years the specific version of YOLO; `YOLOv5`, `YOLOv8` and `YOLOv11` has changed.
[5]See Section 3.3.
[6]A very similar approach is briefly described in the preprint Shkodrani *et al.* [(38)] for the joint detection-classification task.

the species-level logits such that, after the softmax, the probability of any higher-level taxon equals the sum of the probabilities of its child classes.

We have developed the infrastructure needed to assemble large-scale citizen science datasets and have set up training pipelines for all the model architectures under investigation.[7] We are now in the process of defining domain-specific performance metrics to compare these models in the context of large-scale insect monitoring using light-attracting camera traps. This will allow us to quantitatively evaluate how well each model addresses the challenges of this specific use-case.

Our primary hypothesis is that models will more effectively leverage the highly imbalanced citizen science datasets: By explicitly learning higher-level 'meta-classes' (genera and families)—an easier task that provides many more training examples per class—these models can both reinforce robust feature representations and simplify the species-level classification task. We expect this to lead to better cross-domain performance and higher accuracy at the species level. We further hypothesize that hierarchical classifiers will be significantly more accurate than flat classifiers at higher taxonomic levels (genus, family). This improvement at the broader levels should reduce the frequency of 'taxonomically catastrophic' misclassifications.

3.3 XPRIZE Rainforest Finals

As stated on the official website of the XPRIZE Rainforest competition:

> "XPRIZE Rainforest is a global 5-year, $10 million competition that convenes innovators and experts across disciplines — from conservationists and Indigenous scientists to engineers and roboticists — and challenges them to use novel technologies to expedite the monitoring of tropical biodiversity."
>
> — *XPRIZE communication*

the aim of the XPRIZE Rainforest competition was to advance novel biodiversity monitoring methods in the tropics; a goal which follows that of this PhD project very closely, only differing in the biogeographical focus.

I participated in the project as the main researcher involved in image-processing for detection and classification of insects from the light-attracting insect camera canopy raft seen in Fig. 2, as part of the ETH BiodivX team. I developed and implemented[8] a complete, three-stage data analysis pipeline to convert raw light trap images into structured individual-based species detections. For the initial detection step, I adapted an early version of `flatbug`, where the outputs from stage fed into a custom clustering module designed to track individual insects over time. The second tracking stage employed a graph-based clustering approach for individual-tracking; by establishing an 'adjacency criterion' based on strict thresholds for spatial (100 pixels), temporal (300 seconds), and morphological (90% cosine similarity) proximity, with the latter being calculated from a 1280-dimensional feature vector, the tracking problem was reduced to a simple connected component problem, which we solved by calculating the time-chunked transitive closure of the adjacency matrix between detections. For the final classification stage, I built and trained a hierarchical classification model using a dataset of over $813,000$ images compiled from GBIF using a species list obtained via the Catálogo Taxonômico da Fauna do Brasil[47]. This classifier utilized an `EfficientNetV2m` backbone with a bespoke hierarchical head based on the method described in Section 3.2. All three stages were then integrated in an end-to-end pipeline that consumed an image time-series and produced a comprehensive final dataset that includes a detailed table with classifications, track IDs, and bounding boxes, as well as folders of cropped and sorted images for each individual arthropod, ready for downstream analysis.[9] The entire pipeline was built to run in-situ in the Amazonian countryside (Fig. 2) on a consumer-grade laptop and resulted in a total of $685,611$ detections from the three insect-camera traps deployed over the single night of data collection for the finale. Although the methods I developed for the XPRIZE Rainforest finale were certainly not perfect, they provide a microcosm for the aims of my PhD project, and I was also able to produce biologically relevant inferences from the data within the allotted 48 hours of the finale, such as temporal activity curves and species counts as can be seen in Fig. 1.

Through the collaboration of the 50+ members of the ETH BiodivX team, the image-processing pipeline I developed played a part in the us being one of the four named winning teams:

[7] Data assembly was done primarily by Guillaume Mougeot using his tool `gbifxdl`.
[8] With great help and support from Guillaume Mougeot
[9] This section is adapted from the method description written with Toke T. Høye and Guillaume Mougeot for XPRIZE.

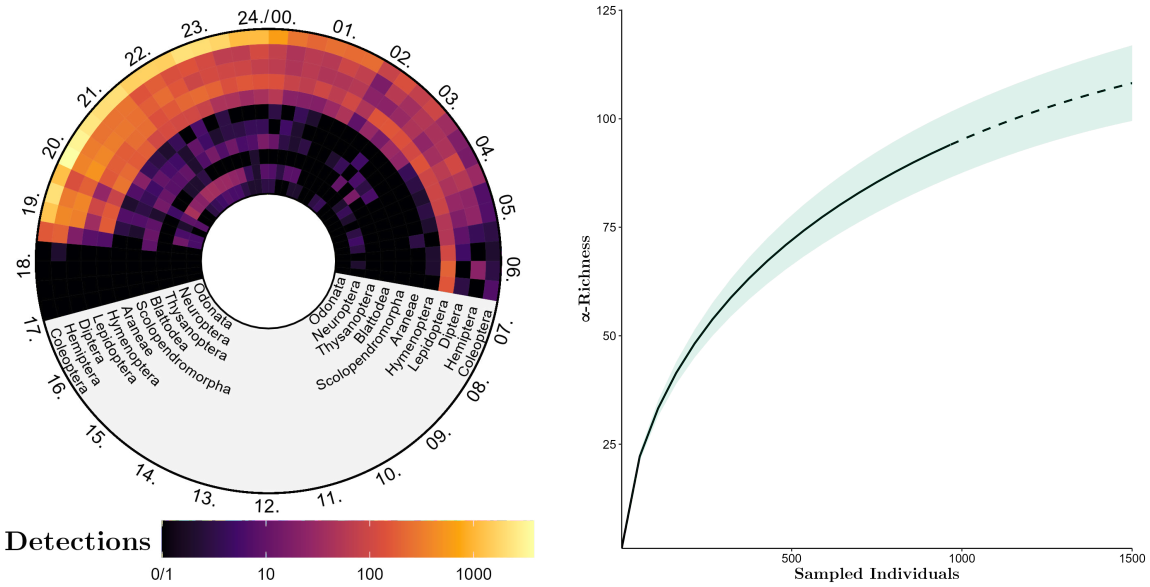


Figure 1: Examples of biological inference from XPRIZE Rainforest data from the three insect camera canopy rafts. Left: Order-level activity curves in 20 minute intervals from one of the rafts (axis is 24 hour, time-of-day). Right: Species rarefaction curve[7],[19] after heavy detection and classification quality filtering. Adapted from XPRIZE Finals results.

> “ Team ETH BiodivX received a $250,000 bonus prize at the discretion of the Judging Panel for their technology which collects large amounts of eDNA, images and sounds through semi-autonomous drones and analyzes the data through a live dashboard utilizing advanced AI algorithms and a global community of Indigenous and forest community citizen scientists. ”
>
> — *XPRIZE Rainforest communication*

Later this also lead to my most far-reaching dissemination activity by participating in the national TV show ‘Go’ Morgen Danmark’ with Professor Claus Melvad from the Department of Mechanical and Production Engineering at Aarhus University in November 2024.

4 Future plans

In Part A of my PhD, I focused on two fundamental problems: individual insect detection and hierarchical classification (with an emphasis on moths). However, the approaches I used were essentially conventional, domain-agnostic classification methods, and did not address a key issue: the identification of organisms cannot be boiled down to rote taxonomic classification. Instead, as biologists we ‘identify’ an organism through multiple lenses; taxonomic, ecological, molecular, and functional. For example, a biologist observing a nut weevil (*Curculio nucum*) on a hazel bush (*Corylus avellana*) might identify it as a ‘medium-sized, hazel-coloured, hazel-herbivorous weevil with a darker snout’, while an eDNA survey of the same area might detect sequences from *Corylus avellana* and *Curculio sp.* (weevil). Yet, most machine learning methods for identifying organisms rely strictly on the species concept and current taxonomy, with prominent examples from the literature such as Jain *et al.* [(20)], Van Horn *et al.* [(46)], Valan *et al.* [(45)], Gu *et al.* [(18)]), as well as in fact my own project described in Section 3.2. Notable exceptions exist, however, in recent multi-modal and integrated approaches. For example, the `CLIBD`[17] model combines image, text, and molecular (DNA barcode) data, while `BioCLIP`[40] and `BioCLIP2`[18] link images with textual descriptions composed of the taxonomic and vernacular species names. Similarly, the `BIOSCAN-1M`[15] and `BIOSCAN-5M`[16] provide image datasets paired with taxonomic information and DNA barcodes.

The following projects in Part B of my PhD aim to close this gap; training organism image ‘classification’ models to not just memorize species (or taxa), but classify individual organisms through multiple different descriptive biological lenses. This broad research direction has two main goals: First, to produce more robust models, while using data and computational resources more efficiently. Second, to determine whether image species classification models (conventional or novel) implicitly learn biologically meaningful features (i.e. traits). I will approach this second goal as both an explainability problem: discovering which traits are most discrimi-

Figure 2: Images from my participation in the XPRIZE Rainforest competition in Manaus, Brazil. Top left: The light-attracting insect camera 'canopy raft' developed by Claus Melvads' group at the department of Mechanical and Production Engineering at Aarhus University. Top right: A representative image of the Brazilian rainforest very close to the site where the camera was deployed. Bottom left: The drone used for deployment of the cameras operated by Georg Strunk in preparation for the competition finale. Bottom right: Me (center) in the progress of in-situ deployment of the image-processing pipeline developed for the competition with ETH BiodivX team lead Kristy Deiner (left), co-lead of insights David Dao (center-right), and local team member Gabriel Nunes (right). Credit: Asger Svenning (1-3), ETH BiodivX on YouTube (4).

native in the learned representation (latent/embedding space), and an optimization problem: 'controlling' the training of these models towards using based on a priori known discriminative traits. While the first main goal is more practical and the second is more academic in nature, I hypothesize and intend to test through the projects described below, that they are closely intertwined and follow from each other.

As shown in Fig. 3, the first planned project (B) under this umbrella relates to hierarchical taxonomic classification — arguably the simplest additional lens that can be added. This project is already well underway and should be completed within the first few months of Part B. The second project (C) aims to radically redefine image-based species classification by using a molecular sequence in place of the species label and changing the learning objective from species identification to molecular sequence inference. Technically, this project should be relatively straightforward, but it represents a fundamental reframing of what it means to 'classify' an organism. The third project (D) has two complementary goals. First, to determine which discriminative features (traits) a trained classification network has learned to use. Second, to develop a method for inferring those traits on new images using the network's activations, including the ability to interpolate missing traits for species with sparse trait data. Project D will be more mathematical and statistical in nature than the others in Part B. I have begun some preliminary work, but progress is less straightforward because this project depends more on theoretical insights ('good ideas') than on data collection or routine model training. For this reason, I have allocated a longer time span and allowed it to overlap more with the other projects. The final project (E) is a synthesis effort. Here I plan to integrate all the different types of organismal classification—taxonomic, molecular, and functional—directly into a standard image-classification model (instead of only linking them via a contrastive multi-modal approach). This project will combine the methods and insights from all prior projects (along with relevant approaches from the literature) to develop a holistic framework for biologically informed

image classification of organisms, insects in particular.

Overall, the goal of Part B is to develop more robust image classification models for biological applications (especially entomology) and to increase their explainability and controllability. Through these projects, I also aim to better understand what features or patterns image-based species classifiers learn internally, and to demonstrate the benefits of grounding machine learning methods in biological and ecological knowledge.

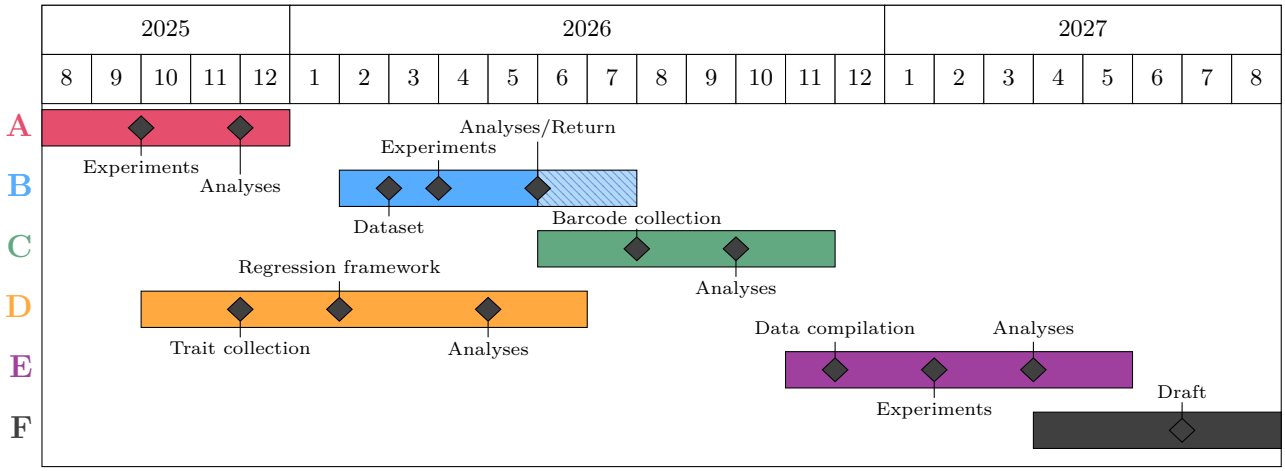


Figure 3: Planned timetable for the remaining projects in the second half of my PhD project.

A) Hierarchical Classification with Citizen Science Data
B) Research Environment Exchange: Inria/Pl@ntNet
C) Molecularly informed image classification of biological species
D) Inference of arthropod species and individual functional traits using classification neural networks
E) Multi-domain integration of biological species-level information for species image classification
F) Dissertation writing

4.1 Research Environment Exchange: Inria/Pl@ntNet

In the spring of 2026 I plan to conduct my Research Environment Exchange at University of Montpellier with Research Director Alexis Joly at the French Institute for Research in Computer Science and Automation (Inria), who works among other things as the scientific and technical director of Pl@ntNet[31]. During this exchange, I will to extend my approach described in Section 3.2 to the plant species identification problem.

4.2 Molecularly informed image classification of biological species

This project is motivated by several limitations of the traditional Linnaean taxonomic system in image species classification. First, taxonomy can be problematic due to issues like unresolved taxa and name resolution inconsistencies. Second, the fundamental problem that the biological species concept itself can be arbitrary and produce inconsistent species delineations[9],[10],[39]. These problems are exacerbated by the lack of alignment between the biological taxonomy and phylogeny, despite considerable efforts[8],[23],[25],[28],[43], and the sheer number of undescribed species[22],[29]. This project seeks to address these issues by developing a method for training and using molecular species-free image classification models. Specifically, the project will focus on the describing the effects of using the molecular representation which is both less rigid than the biological taxonomy and more quantitatively aligned with evolutionary relationships. In essence, this project is a direct continuation of the project on taxonomic hierarchical classification described in Section 3.2, aiming to take this a step further by using molecular sequences instead.

The project idea is to use available species annotated image data and replace the species labels with a fixed molecular subsequence suitable for species discrimination across the investigated taxon group—such as the COI barcode—and reframing species classification problem as a molecular sequence inference problem. Such a sequence can be trivially transformed into a binary sequence[10] suitable for binary classification. Admittedly, this reformulation may seem counterintuitive — after all, one cannot directly see a DNA sequence in an image. However, because related organisms share many portions of their DNA sequences in a hierarchical pattern, the model can exploit these correlations. In fact, the viability of this idea is supported by recent work (e.g., Gong

[10]The molecular alphabet consists of 4 characters, 2 bits, thus any fixed molecular segment can be encoded in a $2N$ binary sequence.

et al. [(17)]) that successfully link images with genetic information, giving merit that the task is both feasible and meaningful. A noteworthy technical benefit of using a fixed-length sequence as the output is scalability: a sequence of length N (bits) can uniquely encode exponentially more than N possible classes. By contrast, in a traditional classification model, the output dimensionality equals the number of classes, so adding classes linearly increases the model output size. In a typical classifier, the number of parameters in the final layer grows linearly with the number of classes. For instance, `TreeOfLife-200M` dataset[18] contains $952,257$ classes. Thus, for even a moderately sized network with a 1024-dimensional embedding space, would require just shy of one billion parameters just in its final classification layer to handle the number of classes! In theory, one million unique sequences (classes) could be encoded in fewer than 20 bits. In other words, a network following this approach might need on the order of only $20,000$ parameters in its final layer — a $\sim 50,000\times$ reduction compared to the traditional model![11]

Of course, a model like this would still need a way to map its predicted sequence back to an actual species. In other words, after the network outputs a sequence, we must perform an additional lookup using the sequence-to-species mapping that we initially used to label the data. I have not yet determined the exact method for this lookup[12], but it is a necessary step for using the model's predictions in practice.

The hypotheses are similar to those described in Section 3.2. Additionally, I hypothesize that training this sequence-based model will be far more computationally efficient than training a traditional classifier, owing to the dramatically smaller size of the final layer.

4.3 Inference of arthropod species and individual functional traits using classification neural networks

The projects and manuscripts in my PhD project have so far approached the computer vision-biodiversity monitoring problem from the species identification perspective. This is the most wide-spread direction within the field, operating under an implied assumption that any needed functional traits[5] (ecological attributes) can be appended later using species-trait databases[12]. However, this assumption is questionable for two reasons: (1) Species-trait data is extremely sparse, particularly in less charismatic and species-rich taxa such as insects, and (2) species-level trait data ignores intraspecific variation, which has been criticized as a limitation in ecological anlyses[4]. In this project I hypothesize that it is possible to derive a mapping from a model's latent space to organismal functional traits by analyzing the weights of the final layer of a trained species-classification network.

The aim of this project is both to interpolate sparse species-trait data and infer individual-level traits, by leveraging the learned latent traits encoded in the weights of trained species image classification models. Further mathematical derivations amount to solving the stated problem, but from the current progress on this project they amount to a low-rank multivariate multiple-regression analysis. A technical problem definition and sketch of the current progress can be found in Appendix A.

This project aligns with the overarching goal of my PhD by shifting from a strictly species-centric view of biodiversity to a functional view. Emphasizing functional traits in biodiversity assessments has long been advocated in ecology[27], and this work offers one path to integrate that perspective into computer vision models.

4.4 Multi-domain integration of biological species-level information for species image classification

In the previous projects I tackled species image classification from three new angles: incorporating biological taxonomy, reframing the task as molecular sequence inference, and linking model latent spaces to ecological trait data. A natural next step is to integrate all four domain—visual, taxonomic, molecular, and functional—into a single classification framework. This approach is similar in spirit to recent multi-modal frameworks like BioCLIP[18],[40]. However, it differs crucially in strategy: instead of aligning separate modalities in a joint embedding space (à la CLIP[33]), I plan to modify the classifier itself by redesigning its output head to incorporate multiple information types.

The core challenge for the project is figuring out how to integrate these disparate types of species information into one model. The aim is to construct an image classification model that can learn from heavily imbalanced species data both efficiently and robustly, while also remaining interpretable (i.e. we can understand what it learns about each biological domain). My plan is to synthesize the solutions from the prior projects (Section 3.2, Section 4.2, and Section 4.3) into a single, holistic method. By building on those existing methods instead of developing an entirely new approach from scratch, I can focus on exploring the benefits of integrating multiple views of organism identification within one framework. The major anticipated hurdles are: assembling a dataset that contains a large number of species with images plus associated molecular sequences and trait data for

[11] In practice the reduction would likely be somewhat less drastic, for example if the length of the barcode is around 500, the parameter reduction would be closer to $\sim 1000\times$.

[12] It may be possible to leverage existing bioinformatics algorithms, for example a form of generalized minimum-distance decoding, to match sequences to species.

each, and conducting training and experiments at a sufficiently large scale to demonstrate the benefits of this integrated approach. This project can be subdivided into several concrete subtasks: (1) identify a large image dataset of species (e.g., `TreeOfLife-200M`[18]); (2) curate a subset of those species that have known molecular sequences; (3) gather and interpolate functional or morphological trait data for these species (using sources like Gallagher *et al.* [(12)]); (4) integrate the methods from the previous projects (and relevant literature) into a unified model; and (5) execute experiments, analyze the results, and write up the findings.

The perspective of the project is to move the field of automated biodiversity monitoring toward an integrated, multifaceted view of biodiversity. By incorporating multiple biological data domains into our models, we aim not only to achieve methodological improvements in classification performance, but also to enable a more comprehensive and nuanced description of ecosystems.

5 Manuscripts

- Roy, D. B. *et al.* Towards a standardized framework for AI-assisted, image-based monitoring of nocturnal insects. *Philosophical Transactions of the Royal Society B: Biological Sciences* **379.** Publisher: Royal Society, 20230108. https://royalsocietypublishing.org/doi/10.1098/rstb.2023.0108 (2025) (May 2024)†
  Contribution: Writing—review and editing, particularly on section 3: *"3. Machine learning workflow"*.

- Svenning, A. *et al.* A General Method for Detection and Segmentation of Terrestrial Arthropods in Images. *bioRxiv,* 2025.04.08.647223. http://biorxiv.org/content/early/2025/04/14/2025.04.08.647223.abstract (Jan. 2025)⋆
  Contribution: First author, conceptualization, data annotation, development, experiments, writing, editing, and submission.

- Geckeler, C. *et al.* Field Deployment of BiodivX Drones in the Amazon Rainforest for Biodiversity Monitoring. *IEEE Transactions on Field Robotics* **2,** 336–352. ISSN: 2997-1101. https://ieeexplore.ieee.org/document/11018361/authors (2025) (2025)†
  Contribution: Writing—review and editing, particularly section X.C: *"C. SENSING PAYLOAD"* and section XI, XII, and XIII: *"Results"*, *"Discussion"*, and *"Conclusion"*.

†Published manuscript. ⋆First author manuscript.

6 References

[1]. Akyon, F. C., Onur Altinuc, S. & Temizel, A. *Slicing Aided Hyper Inference and Fine-Tuning for Small Object Detection* in *2022 IEEE International Conference on Image Processing (ICIP)* ISSN: 2381-8549 (2022), 966–970.

[2]. Bjerge, K. *et al.* Hierarchical classification of insects with multitask learning and anomaly detection. *Ecological Informatics* **77,** 102278. ISSN: 1574-9541. https://www.sciencedirect.com/science/article/pii/S1574954123003072 (2025) (Nov. 2023).

[3]. Bocking, S. Science and conservation: A history of natural and political landscapes. *Environmental Science & Policy. Into the fray. Strategic perspectives on biodiversity sciences and politics* **113,** 1–6. ISSN: 1462-9011. https://www.sciencedirect.com/science/article/pii/S1462901116308279 (2025) (Nov. 2020).

[4]. Bolnick, D. I. *et al.* Why intraspecific trait variation matters in community ecology. English. *Trends in Ecology & Evolution* **26.** Publisher: Elsevier, 183–192. ISSN: 0169-5347. https://www.cell.com/trends/ecology-evolution/abstract/S0169-5347(11)00024-3 (2025) (Apr. 2011).

[5]. Brousseau, P.-M., Gravel, D. & Handa, I. T. On the development of a predictive functional trait approach for studying terrestrial arthropods. en. *Journal of Animal Ecology* **87.** _eprint: https://besjournals.onlinelibrary.wiley.com/doi/pdf/10.1111/1365-2656.12834, 1209–1220. ISSN: 1365-2656. https://onlinelibrary.wiley.com/doi/abs/10.1111/1365-2656.12834 (2025) (2018).

[6]. Cardinale, B. J. *et al.* Biodiversity loss and its impact on humanity. en. *Nature* **486.** Publisher: Nature Publishing Group, 59–67. ISSN: 1476-4687. https://www.nature.com/articles/nature11148 (2025) (June 2012).

[7]. Chao, A. *et al.* Rarefaction and extrapolation with Hill numbers: a framework for sampling and estimation in species diversity studies. *Ecological Monographs* **84,** 45–67 (2014).

[8]. Chesters, D. The phylogeny of insects in the data-driven era. en. *Systematic Entomology* **45.** _eprint: https://resjournals.onlinelibrary.wiley.com/doi/pdf/10.1111/syen.12414, 540–551. ISSN: 1365-3113. https://onlinelibrary.wiley.com/doi/abs/10.1111/syen.12414 (2025) (2020).

[9]. De Queiroz, K. Species concepts and species delimitation. eng. *Systematic Biology* **56,** 879–886. ISSN: 1063-5157 (Dec. 2007).

[10]. Donoghue, M. J. A Critique of the Biological Species Concept and Recommendations for a Phylogenetic Alternative. *The Bryologist* **88.** Publisher: American Bryological and Lichenological Society, 172–181. ISSN: 0007-2745. https://www.jstor.org/stable/3243026 (2025) (1985).

[11]. Ellerbrok, J. S. *et al.* Most habitat's and species' assessments in German Natura 2000 sites reflect unfavourable conservation states. *Basic and Applied Ecology.* ISSN: 1439-1791. https://www.sciencedirect.com/science/article/pii/S1439179125000581 (2025) (July 2025).

[12]. Gallagher, R. V. *et al.* Open Science principles for accelerating trait-based science across the Tree of Life. en. *Nature Ecology & Evolution* **4.** Publisher: Nature Publishing Group, 294–303. ISSN: 2397-334X. https://www.nature.com/articles/s41559-020-1109-6 (2025) (Mar. 2020).

[13]. Geckeler, C. *et al.* Field Deployment of BiodivX Drones in the Amazon Rainforest for Biodiversity Monitoring. *IEEE Transactions on Field Robotics* **2,** 336–352. ISSN: 2997-1101. https://ieeexplore.ieee.org/document/11018361/authors (2025) (2025).

[14]. Geissmann, Q., Abram, P. K., Wu, D., Haney, C. H. & Carrillo, J. Sticky Pi is a high-frequency smart trap that enables the study of insect circadian activity under natural conditions. en. *PLOS Biology* **20.** Publisher: Public Library of Science, e3001689. ISSN: 1545-7885. https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.3001689 (2025) (July 2022).

[15]. Gharaee, Z. *et al. A Step Towards Worldwide Biodiversity Assessment: The BIOSCAN-1M Insect Dataset* arXiv:2307.10455. Nov. 2023. http://arxiv.org/abs/2307.10455 (2024).

[16]. Gharaee, Z. *et al. BIOSCAN-5M: A Multimodal Dataset for Insect Biodiversity* arXiv:2406.12723 [cs]. Mar. 2025. http://arxiv.org/abs/2406.12723 (2025).

[17]. Gong, Z. *et al. CLIBD: Bridging Vision and Genomics for Biodiversity Monitoring at Scale* arXiv:2405.17537 [cs]. Apr. 2025. http://arxiv.org/abs/2405.17537 (2025).

[18]. Gu, J. *et al. BioCLIP 2: Emergent Properties from Scaling Hierarchical Contrastive Learning* arXiv:2505.23883 [cs]. May 2025. http://arxiv.org/abs/2505.23883 (2025).

[19]. Hsieh, T. C., Ma, K. H. & Chao, A. *iNEXT: Interpolation and extrapolation for species diversity* manual (2024). http://chao.stat.nthu.edu.tw/wordpress/software_download/.

[20]. Jain, A. *et al. Insect Identification in the Wild: The AMI Dataset* en. in *Computer Vision – ECCV 2024* (eds Leonardis, A. *et al.*) (Springer Nature Switzerland, Cham, 2025), 55–73. ISBN: 978-3-031-72913-3.

[21]. Jia Deng *et al.* ImageNet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition.* MAG ID: 2108598243 S2ID: d2c733e34d48784a37d717fe43d9e93277a8c53e, 248–255 (June 2009).

[22]. Joppa, L. N., Roberts, D. L. & Pimm, S. L. How many species of flowering plants are there? eng. *Proceedings. Biological Sciences* **278,** 554–559. ISSN: 1471-2954 (Feb. 2011).

[23]. Kjer, K. M., Simon, C., Yavorskaya, M. & Beutel, R. G. Progress, pitfalls and parallel universes: a history of insect phylogenetics. *Journal of The Royal Society Interface* **13.** Publisher: Royal Society, 20160363. https://royalsocietypublishing.org/doi/full/10.1098/rsif.2016.0363 (2025) (Aug. 2016).

[24]. Knapp, S. *et al.* A Research Agenda for Urban Biodiversity in the Global Extinction Crisis. *BioScience* **71,** 268–279. ISSN: 0006-3568. https://doi.org/10.1093/biosci/biaa141 (2025) (Mar. 2021).

[25]. Liu, G.-Q., Lian, L. & Wang, W. The Molecular Phylogeny of Land Plants: Progress and Future Prospects. en. *Diversity* **14.** Number: 10 Publisher: Multidisciplinary Digital Publishing Institute, 782. ISSN: 1424-2818. https://www.mdpi.com/1424-2818/14/10/782 (2025) (Oct. 2022).

[26]. Luca Bertinetto *et al.* Making Better Mistakes: Leveraging Class Hierarchies With Deep Networks. *Computer Vision and Pattern Recognition.* ARXIV_ID: 1912.09393 MAG ID: 3035406632 S2ID: c2efec31d421aca7bf68200ffd656cf31d1ab977, 12506–12515 (June 2020).

[27]. McGill, B. J., Enquist, B. J., Weiher, E. & Westoby, M. Rebuilding community ecology from functional traits. *Trends in Ecology & Evolution* **21,** 178–185. ISSN: 0169-5347. https://www.sciencedirect.com/science/article/pii/S0169534706000334 (2025) (Apr. 2006).

[28]. Misof, B. *et al.* Phylogenomics resolves the timing and pattern of insect evolution. *Science* **346.** Publisher: American Association for the Advancement of Science, 763–767. https://www.science.org/doi/10.1126/science.1257570 (2025) (Nov. 2014).

[29]. Mora, C., Tittensor, D. P., Adl, S., Simpson, A. G. B. & Worm, B. How Many Species Are There on Earth and in the Ocean? en. *PLOS Biology* **9.** Publisher: Public Library of Science, e1001127. ISSN: 1545-7885. https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.1001127 (2025) (Aug. 2011).

[30]. Papyan, V., Han, X. Y. & Donoho, D. L. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences* **117.** Publisher: Proceedings of the National Academy of Sciences, 24652–24663. https://www.pnas.org/doi/10.1073/pnas.2015509117 (2025) (Oct. 2020).

[31]. *Pl@ntNet* en-US. https://plantnet.org/en/ (2024).

[32]. Pressey, R. L. *et al.* The mismeasure of conservation. English. *Trends in Ecology & Evolution* **36.** Publisher: Elsevier, 808–821. ISSN: 0169-5347. https://www.cell.com/trends/ecology-evolution/abstract/S0169-5347(21)00181-6 (2025) (Sept. 2021).

[33]. Radford, A. *et al. Learning Transferable Visual Models From Natural Language Supervision* arXiv:2103.00020 [cs]. Feb. 2021. http://arxiv.org/abs/2103.00020 (2025).

[34]. *Regulation - EU - 2024/1991 - EN - EUR-Lex* en. Doc ID: 32024R1991 Doc Sector: 3 Doc Title: Regulation (EU) 2024/1991 of the European Parliament and of the Council of 24 June 2024 on nature restoration and amending Regulation (EU) 2022/869 (Text with EEA relevance) Doc Type: R Usr_lan: en. https://eur-lex.europa.eu/eli/reg/2024/1991/oj/eng (2025).

[35]. Robin, L. The rise of the idea of biodiversity: crises, responses and expertise. en. *Quaderni. Communication, technologies, pouvoir.* Number: 76 Publisher: Les éditions de la Maison des sciences de l'Homme, 25–37. ISSN: 2105-2956. https://journals.openedition.org/quaderni/92 (2025) (Sept. 2011).

[36]. Roy, D. B. *et al.* Towards a standardized framework for AI-assisted, image-based monitoring of nocturnal insects. *Philosophical Transactions of the Royal Society B: Biological Sciences* **379.** Publisher: Royal Society, 20230108. https://royalsocietypublishing.org/doi/10.1098/rstb.2023.0108 (2025) (May 2024).

[37]. Schmeller, D. S. *et al.* Building capacity in biodiversity monitoring at the global scale. en. *Biodiversity and Conservation* **26,** 2765–2790. ISSN: 1572-9710. https://doi.org/10.1007/s10531-017-1388-7 (2025) (Nov. 2017).

[38]. Shkodrani, S., Wang, Y., Manfredi, M. & Baka, N. *United We Learn Better: Harvesting Learning Improvements From Class Hierarchies Across Tasks* arXiv:2107.13627 [cs]. July 2021. http://arxiv.org/abs/2107.13627 (2025).

[39]. Sokal, R. R. & Crovello, T. J. The Biological Species Concept: A Critical Evaluation. *The American Naturalist* **104.** Publisher: The University of Chicago Press, 127–153. ISSN: 0003-0147. https://www.journals.uchicago.edu/doi/abs/10.1086/282646 (2025) (Mar. 1970).

[40]. Stevens, S. *et al. BioCLIP: A Vision Foundation Model for the Tree of Life* arXiv:2311.18803 [cs]. May 2024. http://arxiv.org/abs/2311.18803 (2025).

[41]. Sutherland, W. J., Pullin, A. S., Dolman, P. M. & Knight, T. M. The need for evidence-based conservation. English. *Trends in Ecology & Evolution* **19.** Publisher: Elsevier, 305–308. ISSN: 0169-5347. https://www.cell.com/trends/ecology-evolution/abstract/S0169-5347(04)00073-4 (2025) (June 2004).

[42]. Svenning, A. *et al.* A General Method for Detection and Segmentation of Terrestrial Arthropods in Images. *bioRxiv,* 2025.04.08.647223. http://biorxiv.org/content/early/2025/04/14/2025.04.08.647223.abstract (Jan. 2025).

[43]. Timmermans, M. J. T. N., Lees, D. C. & Simonsen, T. J. Towards a mitogenomic phylogeny of Lepidoptera. eng. *Molecular Phylogenetics and Evolution* **79,** 169–178. ISSN: 1095-9513 (Oct. 2014).

[44]. *United Nations Treaty Collection* EN. https://treaties.un.org/pages/ViewDetails.aspx?chapter=27&mtdsg_no=XXVII-8&src=TREATY (2025).

[45]. Valan, M., Makonyi, K., Maki, A., Vondráček, D. & Ronquist, F. Automated Taxonomic Identification of Insects with Expert-Level Accuracy Using Effective Feature Transfer from Convolutional Networks. *Systematic Biology* **68,** 876–895. ISSN: 1063-5157. https://doi.org/10.1093/sysbio/syz014 (2024) (Nov. 2019).

[46]. Van Horn, G. *et al. The iNaturalist Species Classification and Detection Dataset* en. in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (IEEE, Salt Lake City, UT, June 2018), 8769–8778. ISBN: 978-1-5386-6420-9. https://ieeexplore.ieee.org/document/8579012/ (2025).

[47]. Zoology, G. B. Catálogo Taxonômico da Fauna do Brasil. por. https://www.gbif.org/dataset/811d48a6-fd04-4a34-81f2-7605492e54b8 (2025).

7 Manuscript A

A General Method for Detection and Segmentation of Terrestrial Arthropods in Images

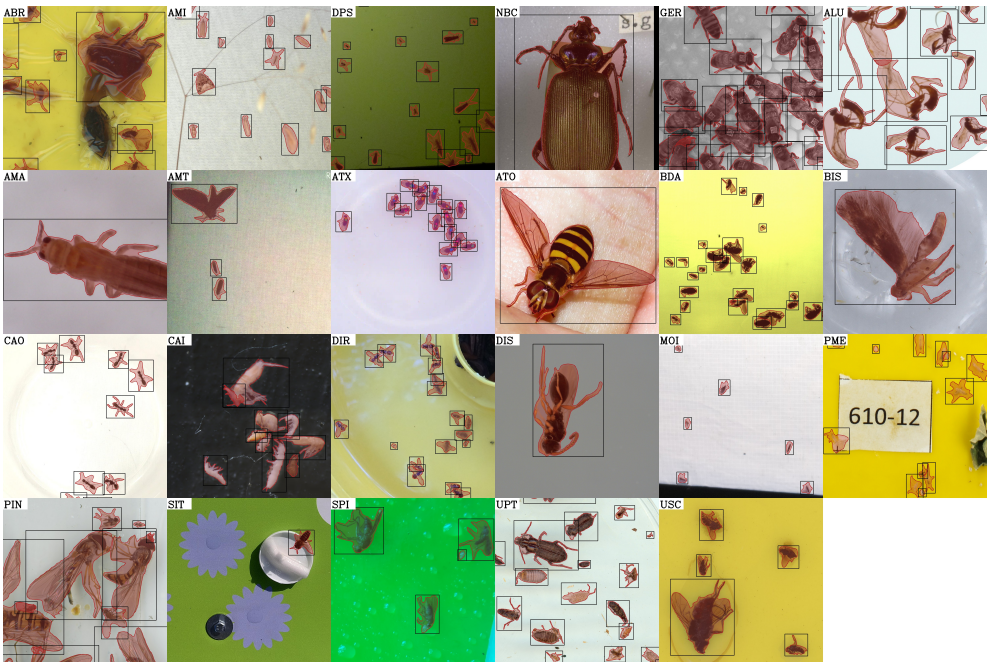
Asger Svenning[1], Guillaume Mougeot[1], Jamie Alison[1], Daphne Chevalier[4], Nisa Chavez Molina[4], Song-Quan Ong[1], Kim Bjerge[3], Juli Carrillo[4], Toke Thomas Høye[1], and Quentin Geissmann[2]

1 *Department of Ecoscience, Faculty of Technical Sciences, Aarhus University*

2 *Center for Quantitative Genetics and Genomics, Faculty of Technical Sciences, Aarhus University*

3 *Department of Electrical and Computer Engineering, Faculty of Technical Sciences, Aarhus University*

4 *Centre for Sustainable Food Systems, Faculty of Land and Food Systems, University of British Columbia, Located on Traditional xʷməθkʷəy̓əm Musqueam Territory*



Abstract

To better understand the status and trends of insects and other arthropods, emerging technologies like image recognition are developing rapidly. This is creating a strong demand for efficient and accurate algorithms for detection and localization of arthropods in images. Existing models have modest performance and do not generalise well to variation in scale, appearance and density of specimens, or imaging conditions. Consequently, each new application often requires manual labeling of training data and model training, which limits the uptake of image-based tools and technologies.

Here, we introduce `flatbug`, which is a powerful and general model to count and outline insects and other terrestrial arthropods in images. The training dataset is large and diverse and represent 23 different lab- and field-based imaging systems. The best `flatbug` model achieves an average $F1 = 94.2\%$ on our validation dataset. Crucially, we show that `flatbug` has great out-of-the-box performance and generalises well to novel contexts. When images from a given dataset are left out of model training, the performance of `flatbug` is only reduced by on average 7.1% for the dataset in question.

By using truly stratified cross-validation, we set a precedent for robust evaluation of deep learning model performance and generalization. We also take steps towards scale- and size-agnostic arthropod detection, by developing an integrated tiling framework for inference and training. Additionally, `flatbug` 's implementation of `YOLOv8` for instance segmentation enables downstream background removal and body size estimation.

The generaliseability of `flatbug` stems from the diversity of contexts represented in the `flatbug` dataset, including 113550 arthropods annotated across 6131 images. Alongside a fully documented Python package with tutorials for integration and analysis via https://github.com/darsa-group/flat-bug/, the `flatbug` dataset is available from https://www.doi.org/10.5281/zenodo.14761447. By providing performant models and the accompanying dataset, `flatbug` offers both a ready-to-use tool and a benchmark for the future. Overall, `flatbug` represents a significant methodological advance within arthropod image detection, with user-friendly integration for monitoring and research.

Keywords: Automated monitoring; insects; arthropods; entomology; deep learning; computer vision; multiple object detection and localization; instance segmentation

1 Introduction

Arthropods are an ecologically important group of organisms under pressure from multiple global change drivers (Wagner et al., 2021). Therefore, it is critically important to understand which arthropods are affected the most and by which drivers, where and when. However, for most arthropod taxa this is currently not possible due to the acute need for better and more standardised arthropod monitoring data, which consequently delays the production of ecological knowledge and conservation actions (Thomas et al., 2019). An important constraint is that traditional arthropod monitoring relies heavily on manual counting and identification of specimens, which requires highly specialised expertise, is very time-consuming and thus costly.

Recently, novel image-based tools for entomology have begun to emerge (Alison & Høye, 2024; Gal, Saragosti, & Kronauer, 2020; Geissmann, 2022; Gharaee, Gong, Pellegrino, Zarubiieva, Haurum, Lowe, McKeown, et al., 2023; Høye et al., 2021; van Klink et al., 2022). However, most of these tools depend on accurate localisation of individual insects in images, and it has been recognised that such tools need to be improved (Høye et al., 2025; van Klink et al., 2024). Therefore, general, powerful and accurate machine learning-based tools are needed for automatically detecting and extracting individual arthropods on a broad spectrum of image types, backgrounds, and resolutions. This task is particularly difficult due to the diversity and complexity of possible images in arthropod monitoring contexts, encompassing multiple orders of magnitude in both image and organism size, image-to-organism size ratio, and organism density, as well as variable lighting, imaging conditions and background compositions.

Generalized machine learning models have already begun to transform monitoring of larger fauna and flora, with tools such as MegaDetector (Beery et al., n.d.) and iNaturalist ("A Community for Naturalists · iNaturalist", n.d.), among others, enhancing scalability and reducing costs for species detection and identification across mammals (Aodha et al., 2018; Shepley et al., 2021; Willi et al., 2019; Z. Wu et al., 2023), amphibians (Kimura & Sota, 2023), birds (Stowell et al., 2019), insects (Hong et al., 2021), plants (Mäder et al., 2021; "Pl@ntNet", n.d.) and more (Kirillov et al., 2023; Kloster et al., 2023). However, existing methods for arthropod detection are typically developed for specialized use cases and lack the ability to generalize across diverse taxa and contexts (Hong et al., 2021; Schneider et al., 2023; Sys et al., 2022). Some previous works, such as Mazen (2023), have used citizen data to create a general model; however, despite the diversity of citizen science images, they rarely contain more than one individual and are not standardized. The lack of open and standardized datasets adhering to the FAIR (Wilkinson et al., 2016) principles (Findable, Accessible, Interoperable, Reusable) has slowed progress in data annotation and decreased its efficient usage (Schneider et al., 2023; van Klink et al., 2022).

Although many models, such as the YOLO (Redmon et al., 2016) series and Mask R-CNN (He et al., 2018), are widely applied for generalized multiple object detection (and segmentation), these models are typically

limited to fixed-size inference at relatively low resolutions, making them unsuitable for processing large images containing small instances. To overcome this challenge, the hyper-inference framework SAHI (Akyon et al., 2022) is often applied on top of models such as YOLOv8 (Jocher et al., 2023). SAHI performs hyper-inference—inference with a base model using a meta strategy—by tiling the image in fixed size tiles at the native scale, with some specified overlap, and then combines the tile predictions with a full-image prediction. This enables the usage of fixed-size inference models such as the YOLO series and Mask R-CNN for large images with instances at the native and full scale, but falls short for intermediate-size instances, especially for very large images.

In addition to addressing fixed-size inference on large images, many detection models also contend with duplicate predictions, a problem which is only exacerbated by SAHI-like hyper-inference. In practice, SAHI, the YOLO series(Redmon et al., 2016), and others tackle this challenge by applying prediction deduplication (or merging) through non-maximum suppression (or its variants), which computes the Intersection over Union (IoU) on bounding boxes rather than on instance segmentation polygons or masks—a strategy frequently used even in segmentation models because calculating intersections for concave polygons is considerably more computationally expensive than for axis-aligned rectangles.

Most current arthropod detection models rely on bounding boxes—either exclusively for both training and inference or at least for processes like non-maximum suppression—primarily due to the lower annotation cost and reduced computational demands compared to segmentation. However, while instance segmentation requires more labor-intensive annotations and involves more complex IoU computations (e.g., between polygons), it offers several crucial benefits: it facilitates more accurate deduplication by leveraging precise object boundaries, enables effective background removal, and improves instance size estimation. These advantages are key to advancing automated arthropod monitoring.

1.1 Contribution

In this manuscript, we aim to address the above challenges in arthropod detection, and provide a practical tool for monitoring schemes, lab experiments, and related applications. Additionally, we aim to set a precedent for more reproducible, effective, robust, and general models and metrics within the machine learning and arthropod research communities, with emphasis on data sharing and evaluating model generalization.

To that end, we developed `flatbug` (https://github.com/darsa-group/flat-bug/), a general multiple-object detection and segmentation model for terrestrial arthropods recorded by diverse imaging systems. `flatbug` is an open-source model, but also a unified framework for training, evaluation, and inference, built on YOLOv8 (Jocher et al., 2023) and trained on a combination of entirely novel data and re-annotated datasets from the literature (Geissmann & Svenning, 2025). `flatbug` integrates a custom-built, scale-agnostic hyper-inference algorithm inspired by SAHI (Akyon et al., 2022), with an appropriate training domain and an evaluation pipeline, specifically tailored for training YOLOv8 models intended for use with `flatbug`. Crucially, in contrast to YOLOv8 and SAHI, `flatbug` principally employs a segmentation-first approach: using segmentation polygon IoU (Intersection over Union) for deduplication, which compares the overlap between the actual shapes of detected objects, rather than their bounding boxes, and is therefore better able to accurately distinguish closely spaced individuals, enhancing `flatbug`'s ability to detect tightly packed arthropods.

As well as fully accessible, open code for our model and hyper-inference algorithm, we provide the accompanying training and validation dataset following the FAIR principles (Wilkinson et al., 2016). The dataset for `flatbug` is unprecedented in its diversity, utilizing 23 subdatasets from distinct arthropod detection applications. Crucially, the dataset is explicitly stratified, allowing our approach to reach unprecedented performance by deploying a true novel-data evaluation regime. Rarely deployed in previous work, this mitigates data leakage and provides more accurate assessments of generalization. Exploiting the diversity and explicit stratification of the `flatbug` dataset, we take a step beyond previous efforts to perform a set of robust cross-validation experiments. First, we explore how removing a subdataset from training affects performance for all datasets in a typical leave-one-out cross-validation procedure. Furthermore, we consider second-order interactions between individual subdatasets (strata) using an original leave-two-out cross-validation approach, allowing us to investigate the relationships and redundancy between subdatasets using a novel inter-strata redundancy metric.

2 Methods & Materials

All experiments, models and software are built in `Python 3.11`, while experiment statistics and figures are created in `R 4.4.1` using the `tidyverse` extensively (R Core Team, 2024; Van Rossum & Drake, 2009; Wickham et al., 2019). Deep learning modeling and post-processing is based on `PyTorch`, `NumPy` and the YOLOv8 segmentation model architecture, while some polygon calculations are executed with `shapely` (Gillies et al., 2025; Harris et al., 2020; Jocher et al., 2023; Paszke et al., 2017). Manual annotations were performed using the online annotation platform `CVAT` (Sekachev et al., 2020). All training experiments were run on UCloud

through DeIC on NVIDIA H100 hardware in Linux Ubuntu 24.04 using `submitit` and SLURM for process management("Facebookincubator/Submitit", 2025; Yoo et al., 2003).

2.1 Dataset

We constructed a diverse dataset by compiling a set of 23 subdatasets (either publicly available images or constructed for this article) covering a wide domain in terms of arthropod taxonomy, cameras, pose and background. In most cases, since most original images were not annotated for instance segmentation, we annotated these images manually, or semi-automatically. In addition to re-annotating available images for 15 subdatasets derived from existing data, we produced 8 previously undescribed subdatasets. The resulting dataset, containing 6131 annotated images from the 23 subdatasets, is available under the Creative Commons license (10.5281/zenodo.14761447) (Geissmann & Svenning, 2025). Table 1, lists, references and briefly describes the individual subdatasets. Each subdataset has its own DOI and is deposited as a separate Zenodo entry which can be found in Table S1.

Table 1: Dataset description and characteristics. The three-letter abbreviation and full names of each subdataset can be found in the first two columns, while image charactistics and content descriptions can be found in the remaining six columns.

		Images						
	Name	Count	Context	Taxonomic Coverage	Instance count	Crowding	Sensor (Standardized)	
ABR	abram2023	26	yellow sticky cards	mixed flying insects	multiple	sometimes touching	flatbed scanner	(✗)
ALU	ALUS[43]; [44]	352	specimens in ethanol	mixed flying insects	several	often adjacent	camera	(✓)
AMA	amarathunga2022[3]; [4]	550	specimens in ethanol	thrips	single	none	microscope	(✓)
AMI	AMI-traps	153	live insect on white-lit screen	mixed flying insects	multiple	often adjacent	camera	(✓)
AMT	AMT[35]; [36]	110	live insect on white-lit screen	mixed flying insects	multiple	often adjacent	camera	(✓)
ATX	anTraX[12]; [13]	212	live insect in laboratory setup	ants	multiple	very often touching	camera	(✓)
ATO	ArTaxOr[34]	1050	live specimen in natural environment	varied	one or few	occasional	camera	(✗)
BDA	biodiscover-arm	53	specimens in ethanol	mixed crawling arthropods	multiple	often adjacent	camera	(✓)
BIS	BIOSCAN[17]; [18]	501	specimens in ethanol	mixed flying insects	single	none	camera	(✓)
CAO	cao2022[8]; [61]	60	live insect in laboratory setup	ants	few	often touching	camera	(✓)
CAI	CollembolAI[50]	43	specimens in ethanol	Collembola	high density	often touching	camera (tiled)	(✓)
DPS	Diopsis[22]	63	live insect on flat screen	mixed flying insects	multiple	occasional	camera	(✓)
DIR	DIRT[28]	291	mostly specimens captured in liquid trap	mixed flying insects	multiple	often touching	camera	(✗)
DIS	DiversityScanner[63]; [64]	529	specimens in ethanol	mixed flying insects	single	none	camera	(✓)
GER	gernat2018	38	live image of hive drawer	honey bees	very high density	very often touching	camera	(✓)
MOI	Mothitor	12	live insect on white-lit screen	moths	few	rarely touching	camera	(✓)
NBC	NHM-beetles-crops[26]	729	museum drawers	beetles	multiple	rarely touching	camera	(✓)
PME	PeMaToEuroPep	284	yellow sticky cards	mixed flying insects	multiple	rarely touching	camera	(✓)
PIN	pinoy2023	76	transparent sticky cards	mixed flying insects	multiple	often overlapping	flatbed scanner	(✗)
SIT	sittinger2023[48]	400	flat surface (simplified artificial flowers)	pollinators	few	rarely touching	camera	(✓)
SPI	sticky-pi[14]	476	yellow sticky cards	mixed flying insects	multiple	often touching	camera	(✓)
UPT	ubc-pitfall-traps	150	specimens in ethanol	mixed crawling arthropods	multiple	sometimes touching	flatbed scanner	(✗)
USC	ubc-scanned-sticky-cards	50	yellow sticky cards	mixed flying insects	multiple	often touching	flatbed scanner	(✗)

2.2 Training

`flatbug` training employs a modified YOLOv8 segmentation training regime, with changes restricted to the sampling and image augmentation stages. Training images are sampled in a manner that mirrors the hyper-inference prediction regime, which the models are intended for. This is achieved by sampling 'tile'-like crops from the full images in the dataset using the following strategy: First, we sample a relevant zoom level, Z:

$$\begin{cases} Z^2 \sim \mathcal{U}\left(1, {}^{T}/_{S}\right), & {}^{T}/_{S} \geq 1 \\ \sqrt{Z} \sim \mathcal{U}\left({}^{T}/_{S}, 1\right), & {}^{T}/_{S} < 1 \end{cases}$$

where T is the desired tile size (default=1024) and $S = \max\{|D_x|, |D_y|\}$ is the maximum dimension size (height; width) of the image. The square root or power of two is added to adjust for the number of possible unique tiles at each zoom-level. Given a specific zoom level z realized from Z, we then either sample or calculate a coordinate offset for the tile o (if the tile size is larger than the dimension, we center the tile):

$$\begin{cases} o \sim \left\lfloor \mathcal{U}\left(\frac{T}{2}, \frac{z|D|-T}{2}\right) \right\rfloor, & z|D| > T \\ o = \left\lfloor \frac{z|D|-T}{2} \right\rfloor, & z|D| \leq T \end{cases}$$

for each dimension. After extracting a random tile from an image, we perform the following standard augmentation operations; rotation (uniform degree probability), color jitter, color inversion (25% probability) and horizontal/vertical flip. The color jitter, horizontal, and vertical flip parameters are controlled through the standard YOLOv8 hyperparameter interface, while the translate and scale parameters from YOLOv8 have been disabled (since they conflict with our tile sampling strategy). We then remove instances with less than 97.5% or 32px of their (scaled) area within the extracted image region (tile)—either because they are fully/partially outside the region, or because they are too small at the sampled zoom level. The removed instances are inpainted using the OpenCV(Bradski, 2000) 'telea' inpainting method(Telea, 2004) if they overlap with the image region to avoid obvious inpainting artifacts. Secondly, we deterministically oversample images to account for the above process: larger images have many more potentially unique 'tiles' approximately proportional to the pixel area of the image. This step ensures that the training process is not biased by differences in how each subdataset is stored; some subdatasets consist of 'precomputed' tiles, where the original images have been gridded and cut before annotation, leading to these subdatasets containing many more individual images than other subdatasets where the full images are annotated instead. See Appendix K for the technical details.

2.3 Inference

`flatbug` employs a custom-built hyper-inference pyramid-tiling algorithm—visualized in Fig. 1 and Fig. 2—comprising pre-processing, tiling, and inference with post-processing. Before inference, the image is zero-padded by 32px on all sides to avoid edge artifacts and ensure consistent inference coverage, particularly near boundaries. Tiling begins by defining N_S zoom levels, Z, as a geometric series running from

$$Z_0 = \frac{T}{S}$$

to

$$Z_{N_S} = 1$$

with a common ratio of $r_Z = 1.5$ by default. For efficiency we truncate the zoom levels at 0.9 and always include 1: $Z_{N_S-1} \leq 0.9, Z_{N_S} = 1$. For each zoom level, a minimal list of tiles is generated based on a specified tile size and minimum overlap, ensuring overlaps are almost equal ($\pm$ 1 pixel). The top (y) or left (x) corner coordinate of each tile, p, is calculated independently for each dimension using:

$$\begin{aligned} p_0 &= 0 \\ p_i &= p_{i-1} + L + \mathbb{1}_{[i<(O \bmod (N-1))]} \\ N = \left\lceil \frac{S-O}{T-O} \right\rceil, \quad L &= \left\lfloor \frac{O}{N-1} \right\rfloor, \quad O = NT - S \end{aligned}$$

where L is the step size, O represents the total overlap, and N is the number of tiles required to cover a dimension of size S with tiles of size T. This formula ensures that non-divisible overlaps are allocated 1 pixel at a time to initial tiles, creating a consistent layout. The tile top-left corner coordinates, P, are then given by all pairs of corner top and left coordinates for each dimension:

$$P = \{(x, y) \mid x \in p(x), y \in p(y)\}$$

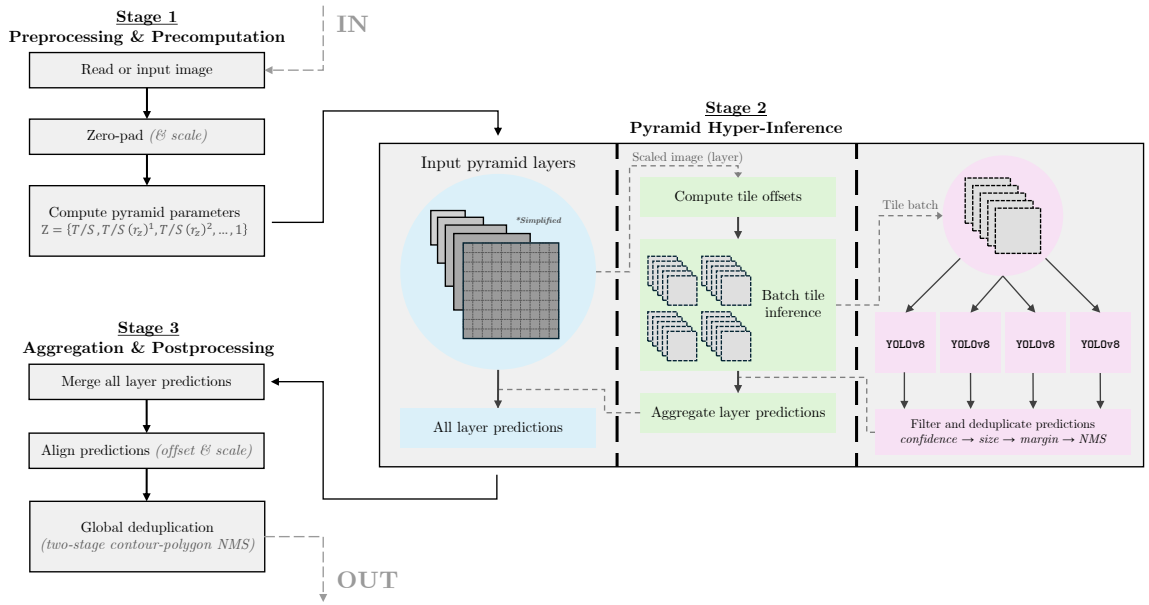


Figure 1: Simplified visual diagram of the `flatbug` inference pipeline. The `flatbug` inference pipeline is fully GPU-accelerated for all steps between reading the input image and global deduplication with CUDA and half-precision compatibility. Global deduplication is also mostly GPU-accelerated, however it proved simpler to use `shapely`(Gillies et al., 2025) a Python wrapper around `GEOS`(GEOS contributors, 2024) for polygon intersection calculation specifically.
Only steps we deemed semantically or technically important in highlighting the defining characteristics of the pipeline are included. The code is available on the `flatbug` repository (https://github.com/darsa-group/flat-bug/).
See Fig. 2 for an accurate depiction of the tile pyramid.

If the scaled and padded image is smaller than the tile size in a dimension, $T < S$, then this dimension is zero-padded symmetrically such that a single tile exactly covers this dimension, $T = S$.

Then, given a zoom level, Z_i, and the corresponding tile positions, P, tiles are processed in batches (default=16) using YOLOv8 inference with two key post-processing modifications: First, bounding box IoU is replaced with contour polygon IoU for deduplication, improving performance on irregular shapes. Second, predictions near tile edges (default: $\leq$16px) are removed to reduce detections of instances that are artifically clipped. The predictions are then offset and scaled to align with the tile positions and zoom levels before combining the results across all zoom levels. Finally, positions are adjusted to account for the initial image padding, and global deduplication is performed using a two-stage NMS (non-max suppression) approach similar to Tan and Wang (2024). The first stage reduces the number of polygon intersection calculations by identifying connected components via pairwise bounding box IoU with a reduced threshold (default=0.05). The second stage applies standard polygon-based NMS within each component, reducing the number of polygon intersection calculations from $O(N^2)$, where N is the number of predicted instances, to

$$O(|C| \max_{S \in C} |S|)$$

where C is the set of connected components. This approach drastically improves efficiency in images with low instance density, since it is significantly faster to calculate intersections between axis-aligned rectangles (bounding boxes) than arbitrary (simple) 2D polygons.

2.4 Evaluation

For comprehensive model evaluation, we used a simple custom-built end-to-end evaluation scheme on the fixed validation split. The end-to-end evaluation starts by running full hyper-inference on all images in the validation split. For each image, we then attempt to match ground-truth instances with predicted instances using a modified NMS algorithm (for tie-breaking) and a polygon-IoU threshold (default=0.2). Then we calculate bootstrap summary statistics recall, precision, and F1 with percentile confidence intervals, denoted as $Q_{x\%}(F^T)$ where $x\%$ is the percentile, F is the statistic, and T is the treatment or stratum, for each subdataset included

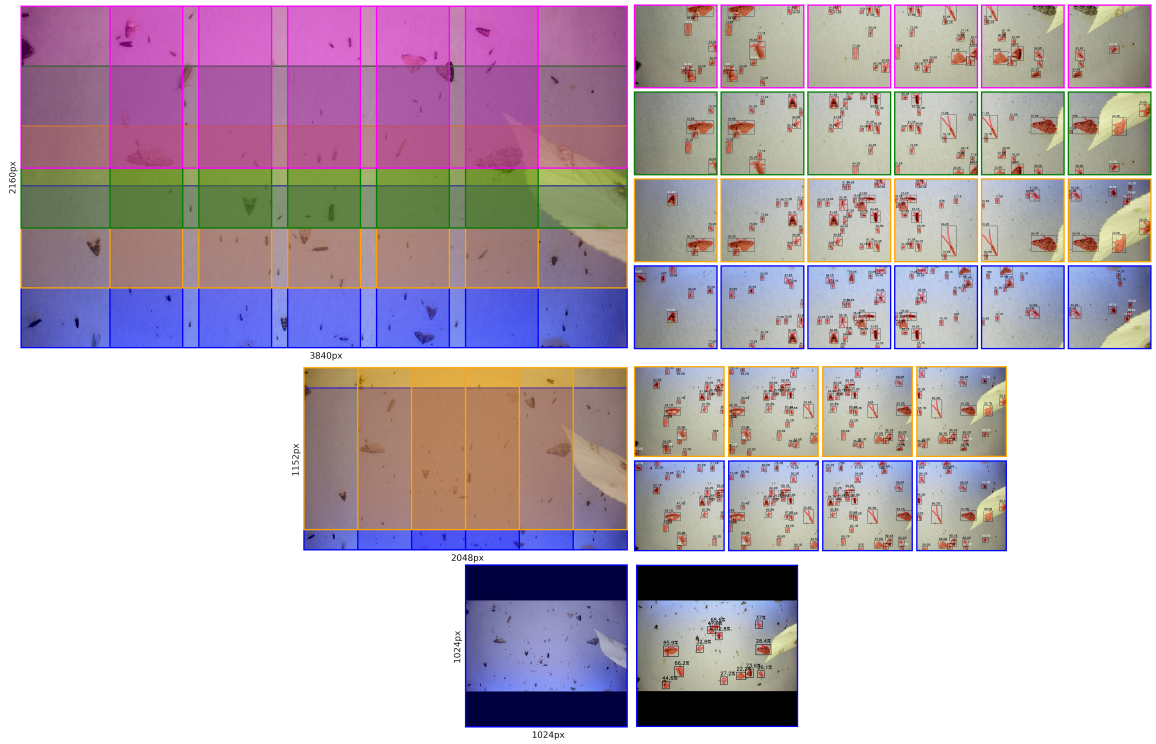


Figure 2: Example visualization of our pyramid-tiling strategy on an image from the AMI subdataset. Each row corresponds to a 'layer' in the pyramid. Left: Tiles overlayed on the scaled image with scaled dimension sizes on the left/bottom. Right: Extracted tiles with predictions; all tiles are 1024px × 1024px.

in the stratum given by T. The bootstrapping procedure simply resamples instances—either inferred/FP (false positives) or labels/FN (false negatives) or matched/TP (true positives)—stratified by subdataset. Changes in performance of a metric are either denoted by Δ_F^T for the absolute change in statistic F or δ_F^T which gives the normalized relative change in statistic F, on only the images in the stratum given by T:

$$\Delta_F^T = F^T - F^*$$
$$\delta_F^T = \frac{\Delta_F^T}{\eta_F^T}$$
$$\eta_F^T = \begin{cases} 1 - F^* & , \quad \Delta_F^T \geq 0 \\ F^* & , \quad \Delta_F^T < 0 \end{cases}$$

where F^* is the baseline (control) under the specified treatment and/or stratum, T. The normalization factor η for δ simply ensures that $\delta \in [-1, 1]$ in contrary to the absolute change with the range $\Delta \in [-F^*, 1 - F^*]$. Intuitively, this means that if $\delta = 1/2$ (50%) then the treated model has gained 50% of the potential performance gain, and if $\delta = -1/2$ (−50%) then it has lost 50% of the potential performance loss.

For all evaluation metrics we also make sure to remove all ground-truth and predictions with below 32px, as measured by the square root of the pixel are of the contour polygon, as these would also have been removed during training. See Appendix N for further details.

For in-training diagnostics, we utilize the native YOLOv8 segmentation evaluation scheme with the previously described modified training augmentation and oversampling techniques on a fixed validation set five times, since our modified augmentation (potentially) samples a 'tile' which covers a smaller region of each image. Validation is conducted after the first epoch, then every 5 epochs by default, and then once at the end of training.

2.5 Experiments

All experiments comprised two sequential steps: (1) training and (2) end-to-end evaluation. The following training hyperparameters are shared for all experiments: initial learning rate (lr0 = 10^{-3}), final learning rate

proportion (lrf = 10^{-5}), stochastic gradient descent (optim = SGD), and batch size (batch = 32), as well as all other YOLOv8 segmentation training hyperparameters not mentioned here. The three 'experiments' were as follows: (1) training YOLOv8 segmentation models at the four available sizes nano (N), small (S), medium (M) and large (L), (2) leave-one-out cross-validation training with fine tuning, and (3) leave-two-out cross-validation training. These experiments were designed to quantify and evaluate the performance of `flatbug` using the four available YOLOv8 backbone sizes, the generalizability of `flatbug` to new images both similar and dissimilar to the images in the subdatasets of the `flatbug` dataset, and the relationships between the subdatasets across the domain of possible subdatasets. Experiments 2 and 3 were performed using only the medium (M) YOLOv8 segmentation model variant for 50 and 20 epochs, respectively, while models in experiment 1 were trained for 300 epochs. For efficiency, we only performed in-training evaluation every 25 epochs during our experiments.

Further details for each experiment are presented below.

2.5.1 Experiment 1: Backbone size comparison

Four models, one at each of the available model sizes in YOLOv8 (L/M/S/N), were trained for 300 epochs on our entire training dataset and evaluated on the entire validation dataset.

2.5.2 Experiment 2: Leave-one-out cross-validation

24 medium (M) size models were trained for 50 epochs, 23 out of 24 models (corresponding to the 23 subdatasets) were trained on our training dataset with a particular subdataset left out, while the last model was trained on the full dataset as a baseline (control); these are the OOB (out-of-box) models. Each of these models was then trained for a further 10 epochs only on the subdataset that was left out (the control model was again trained on the full dataset); these are the FT (fine-tuned) models. For this experiment (Section 3.2) we only present changes in performance metrics at the level of individual subdatasets. In other words, we assess how excluding a given subdataset from training affects the end-to-end evaluation metrics for that same subdataset. The baseline (control) metrics were taken from the two full dataset models (OOB and FT) evaluated on each subdataset.

2.5.3 Experiment 3: Leave-two-out cross-validation

A subset of 16 out of the 23 subdatasets were chosen (ALU, AMA, AMI, AMT, ATO, BDA, BIS, CAI, DIS, DPS, GER, PIN, SIT, SPI, UPT, USC) for leave-two-out cross-validation. The subset of 16 subdatasets were chosen ad-hoc, aiming to be as representative of the full 23 subdatasets as possible, while reducing the number of sub-experiments (i.e. models) to a manageable number. For each unique pairwise combination (including pairs of the same dataset) we trained a medium (M) model for 20 epochs on all 23 subdatasets except the particular pair of subdatasets. A control model was also trained for 20 epochs on all 23 training datasets. Subsequently, all resulting models (N=136 + 1) were evaluated on all datasets leading to a data-cube for each validation metric, $\mathbf{F}$, where element $\mathbf{F}_{(k)i,j}$ corresponds to the value for metric F only on subdataset k for the model which is trained on all subdatasets except i and j. The data cube was then transformed to the normalized relative change:

$$\delta_F^{(k)} = \frac{\mathbf{F}^{(k)} - \mathbf{F}^{(k)*}}{\eta_F^{(k)}}$$

where $\mathbf{F}^{(k)*}$ is the performance of the control model on the subdataset k and $\eta_F^{(k)}$ is again a normalization factor such that $\delta_F \in [-1, 1]$. To quantify the similarity of our subdatasets we defined a redundancy metric which we call one-way redundancy ρ^1:

$$\rho_{ij}^1 = \delta_{F1}^{(i)i,j} - \delta_{F1}^{(i)i,i}$$

where F1 is taken from a model with subdatasets i, j removed from training and evaluated on subdataset (i), corresponding to the extra decrease in performance when omitting the second (j) on top of the first (i) subdataset. Since we want a symmetric redundancy metric in order to use it as a dissimilarity/distance metric, we then define the two-way redundancy ρ^2 as the average of ρ^1 across the diagonal:

$$\rho^2 = \frac{\rho^1 + (\rho^1)^T}{2}$$

Both ρ^1 and $\rho^2 \in [-1, 1]$ should be interpreted such that negative values correspond to synergism (positive performance cross-effect), while 0 corresponds to redundancy (neutral performance cross-effect) and positive values corresponding to antagonism (negative performance cross-effect).

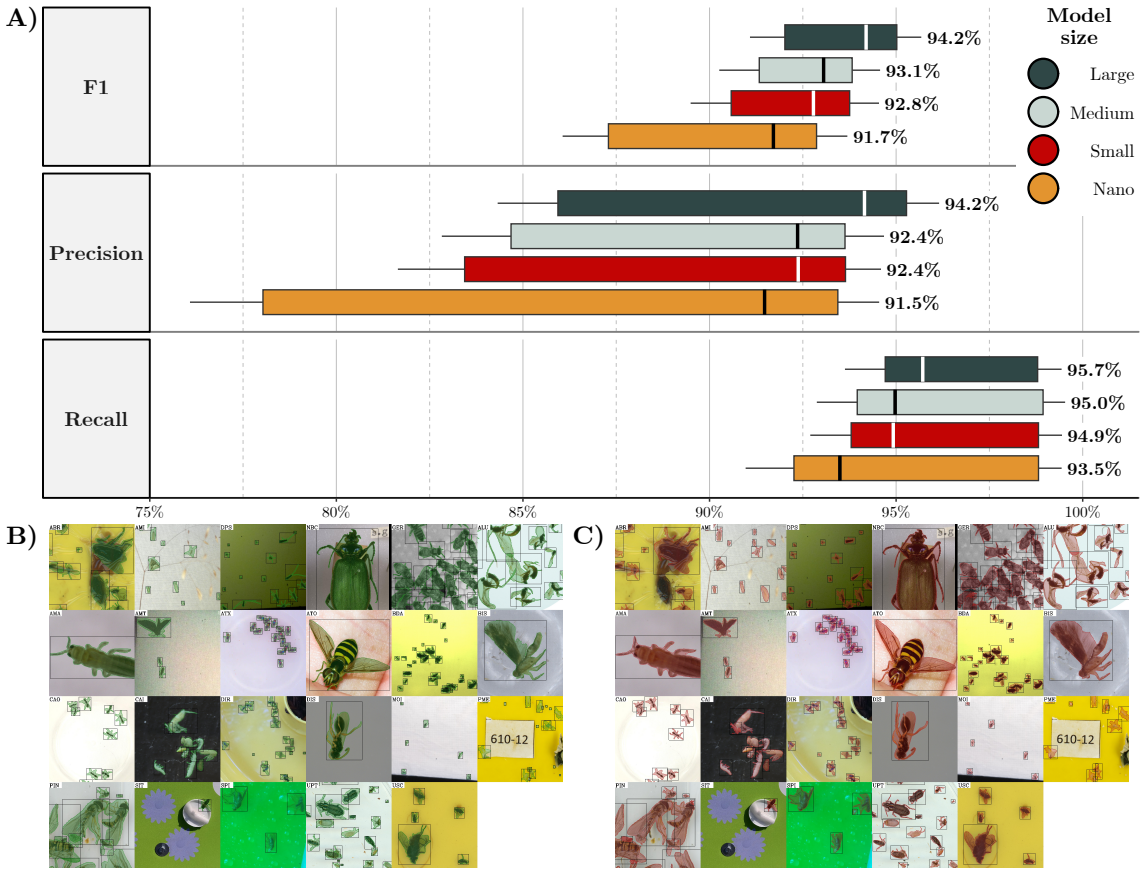


Figure 3: Validation performance of the deployment `flatbug` models at the four available model sizes (**A**) with true label (**B**) and predicted instance segmentation polygons and bounding boxes (**C**) on a hand-picked ROI (in the validation split) from each subdataset in the `flatbug` dataset. Box plots are based on a subdataset stratified bootstrap of model instance predictions (TP/FP/FN) on the validation split and show the 2.5%, 25%, 50% (median; also written next to each box plot), 75% and 97.5% bootstrap quantiles.

3 Results

3.1 Experiment 1: Backbone size comparison

To provide a reasonable baseline for the performance of `flatbug` and a quantifiable justification when choosing a particular YOLOv8 backbone size with `flatbug`, we conducted a standard train-validation experiment with the four available backbone sizes: **L**arge, **M**edium, **S**mall, and **N**ano. All models in this experiment were trained for 300 epochs using our training dataset consisting of 23 subdatasets and 5153 images and evaluated on our evaluation dataset with the same subdatasets and 978 images.

Here, `flatbug` achieves cutting-edge performance of $\widehat{\mathrm{F1}}_L \approx 94.2\%[91.1\%, 95.7\%]$ for the large `flatbug` variant and $\widehat{\mathrm{F1}}_M \approx 93.1\%[90.3\%, 94.6\%], \widehat{\mathrm{F1}}_S \approx 92.8\%[89.5\%, 94.5\%], \widehat{\mathrm{F1}}_N \approx 91.7\%[86.1\%, 93.7\%]$ for the medium, small and nano `flatbug` variants respectively (numbers in square brackets are 95% confidence intervals). These metrics are concordant with the traditional machine-learning precision-recall curve and AP50%/AP50-95% metrics (Appendix L). As expected, smaller variants performed slightly worse on average, with more substantial decreases in worst-case performance of particularly the smallest of the variants; $Q_{2.5\%}(\Delta_{F1}^{\mathrm{M-L}}) \approx -0.73\%[-1.76\%, -0.08\%]$, $Q_{2.5\%}(\Delta_{F1}^{\mathrm{S-L}}) \approx -2.03\%[-3.66\%, -0.97\%]$ and $Q_{2.5\%}(\Delta_{F1}^{\mathrm{N-L}}) \approx -4.64\%[-7.48\%, -2.61\%]$ (Fig. 3). Upon close inspection (Fig. 3A), we found that differences between model sizes were primarily driven by a few outlier datasets, particularly ATO, BIS, and AMA in precision and AMI and BDA in recall (Fig. S6).

From a qualitative comparison of the true label instance polygons (Fig. 3B) the inferred instance segmentation polygons produced by the large (L) `flatbug` variant (Fig. 3C) also reveal high-fidelity contours and a low rate of false positives and negatives. The differences between the predicted and labeled detections can broadly be grouped into three categories; (1) the labeled instance is smaller than model size detection-threshold [AMI=1; DPS=1; BDA=1; DIR=1; PME=5], (2) incorrect separation of instances [CAI=1; PIN=3] and (3)

instance not detected or non-instance incorrectly detected [ABR=1; AMI=6; SPI=1] (Fig. 3B-C).

These results show that `flatbug` is a strong multiple-object detection and segmentation model for a broad spectrum of terrestrial arthropod images and instance sizes, and that particularly the three largest versions (L/M/S) are robust across the subdataset domains.

3.2 Experiment 2: Leave-one-out cross-validation

To assess the potential generalizability of `flatbug`, we investigated the out-of-box performance penalty in a leave-one-out cross-validation experiment. As explained in Section 2.5.2, leave-one-out models were trained on all subdatasets except one, for 50 epochs, subsequently the models were then evaluated only on the particular subdataset which was excluded (left-out) giving the leave-one-out performance. The leave-one-out models were then compared to a control model, which was trained on all subdatasets for 50 epochs. Then, each of the leave-one-out models was trained (fine tuned) for a further 10 epochs, only on the particular dataset which was excluded during the initial training. The control model was again trained (fine tuned) on all subdatasets, and compared with each fine-tuned leave-one-out model based on performance for only the relevant subdataset. Crucially, for both the leave-one-out and fine-tuned models, our comparisons include only the performance on the relevant subdataset for both the treatment and control model, not their overall performance on the whole dataset.

We identified a small, but significant, decrease in F1 performance of $\delta_{F1} \approx -7.1\%[-12.1\%, -0.4\%]$ when leaving individual subdatasets out. This was driven by a larger significant decrease in recall $\delta_{R} \approx -10.3\%[-13.9\%, -6.6\%]$, while precision remained unchanged $\delta_{P} \approx 4.5\%[-4.8\%, 13.7\%]$ (see Section 2.4 for details on δ). Although precision was not significantly affected in the leave-one-out scenario on average, there was significant variation among subdatasets, with a strong negative impact on subdatasets BIS ($\delta_{P}^{BIS} \approx -47.4\%$) and SIT ($\delta_{P}^{SIT} \approx -76.4\%$), while conversely AMT ($\delta_{P}^{AMT} \approx 73.1\%$) and CAO ($\delta_{P}^{CAO} \approx 100.0\%$) reached higher precision when excluded from training. We also found that fine-tuning these models for a modest number of epochs (10) on only the previously left-out dataset consistently removed the performance penalty $\delta_{F1} \approx 0.9\%[-1.3\%, 6.7\%]$, $\delta_{R} \approx 0.0\%[-2.4\%, 4.5\%]$, while marginally improving precision $\delta_{P} \approx 5.1\%[1.7\%, 9.9\%]$ (Fig. 4).

This demonstrates that `flatbug` is potentially able to generalize to novel subdomains across most terrestrial arthropod images, even without any retraining, and that even a small amount of retraining—only including images from the novel domain—will lead to high performance, similar to that expected from a domain-specific model.

3.3 Experiment 3: Leave-two-out cross-validation

To explore our coverage and robustness across the potential data domain, we defined a pairwise subdataset redundancy metric—two-way redundancy, ρ^2, (see Section 2.5.3 for details on ρ^2)—based on the leave-two-out cross-validation experiment using 16 out of the 23 subdatasets in the complete `flatbug` dataset. We found that most (11 out of 16) of the investigated subdatasets hierarchically cluster in a single central synergistic cluster (measured by F1) interactions, while three subdatasets (ATO, DPS, CAI) are not positively affected by the other investigated subdatasets and the remaining two subdatasets (BIS, PIN) appear to form an outlier cluster with mutual benefit (Fig. 5). On the other hand, we found that the pair AMI-AMT is strongly synergistic with the lowest two-way redundancy, while being firmly placed in the major cluster, which should be expected as these two subdatasets originate from the same camera trap and imaging system (although from different iterations).

We corroborated the overall pattern by an ordination of the pairwise redundancy matrix, this ordination has a similar, but not identical structure, to the hierarchical clustering tree, where subdatasets BIS, CAI and DPS qualitatively appear as outliers. By constructing a minimum spanning tree on the ordinated subdatasets, we also noted that most (10 out of 15) edges in the minimum spanning tree have negative two-way redundancies, with most (3/5) of the positive edges connect to ATO, DPS and CAI (Fig. S4), further validating the existence of a major central synergistic cluster, and the identity of the outlying subdatasets.

These patterns, particularly the existence of a major central synergistic cluster, show that the `flatbug` dataset (Geissmann & Svenning, 2025) has high coverage across its central domain (top-down images of terrestrial insects with low- to semi-complex backgrounds), while some of the peripheral domains (such as citizen science images) would benefit from the addition of similar subdatasets.

4 Discussion

We have presented `flatbug`, a powerful, general, scale- and size-agnostic model for precise detection and segmentation of individual arthropods from image-based approaches in entomology. It provides a simple and flexible out-of-the-box method for precisely extracting individual organisms in a very broad range of images.

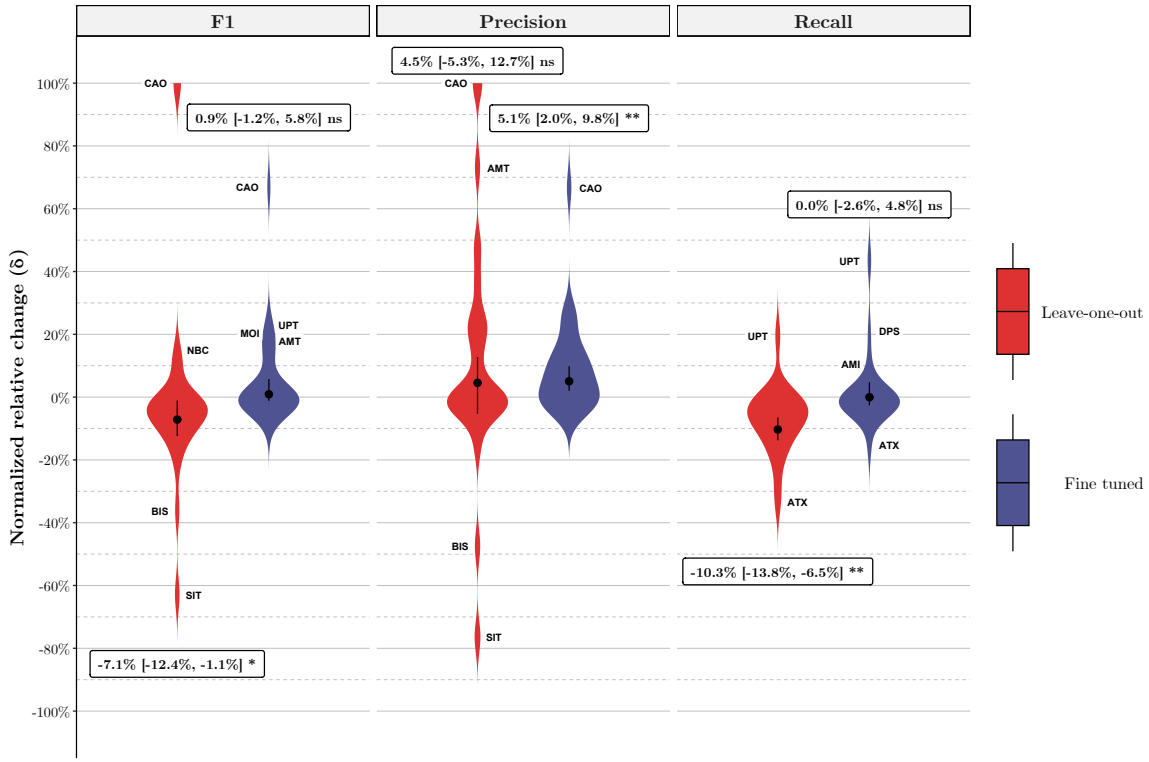


Figure 4: Density estimates (violins) and bias-corrected and accelerated (BCa)(DiCiccio & Efron, 1996) confidence intervals with empirical p-values(North et al., 2002) based on a weighted bootstrap for the normalized relative change, δ, (the proportion of the potential performance degradation/gain achieved under the treatment; see Section 2.4 for details on δ) between a leave-one-out and a control `flatbug` model (M) (leave-one-out) and a fine-tuned and a second control `flatbug` model (fine-tuned). Confidence interval and density for δ are weighted proportionally to the logarithm of the number of instances (true or false) in each subdataset as a compromise between micro- and macro-averaging. Three-letter annotations (e.g. CAO) are 'outlier' subdatasets for each sub-experiment.
By definition, the normalized relative change, δ, of the control models is 0 in all cases, meaning that confidence intervals that overlap with zero, represent scenarios in which the performance of the alternate models are not significantly different from the control models.

As image-based approaches in entomology continue to gain traction (Høye et al., 2021; Kitzes et al., 2021; Schneider et al., 2023; van Klink et al., 2022), `flatbug` is solving the critical challenge of accurate and efficient arthropod detection. In combination with the `flatbug` accompanying dataset (Geissmann & Svenning, 2025) and robust performance evaluation, we pave the way for developments in individual detection within the automated arthropod monitoring community.

A major methodological milestone in computer vision for entomology is moving detection from specialized, bounding-box-based approaches to general, segmentation-based approaches. Previous methods for individual detection within ecological monitoring, particularly within terrestrial arthropod or entomological monitoring, and research have had limited taxonomic scope and little or no ability to generalize, whether to different taxonomic or image domains (Gal, Saragosti, & Kronauer, 2020; Geissmann, 2022; Hong et al., 2021; Kimura & Sota, 2023; Kloster et al., 2023; Sys et al., 2022; Z. Wu et al., 2023). Although some studies have attempted to segment arthropods in images (Geissmann, 2022), arthropod detection studies rarely attempt to demonstrate generality (Gal, Saragosti, & Kronauer, 2020; Hong et al., 2021; Sys et al., 2022), and tests of generalization, such as the leave-one-out experiment presented here, are extremely rare (Mazen, 2023; Willi et al., 2019).

We believe this discrepancy is caused by a lack of standardization between datasets and a lack of FAIR approaches to data management. Furthermore, significantly reduced annotation effort for segmentation masks compared to bounding boxes has meant that the potential benefits of segmentation is underexplored and likely understated.

We argue that segmentation offers four major benefits: (1) the background of detected individuals can be removed, reducing potential bias in downstream classification, (2) inference deduplication with non-max

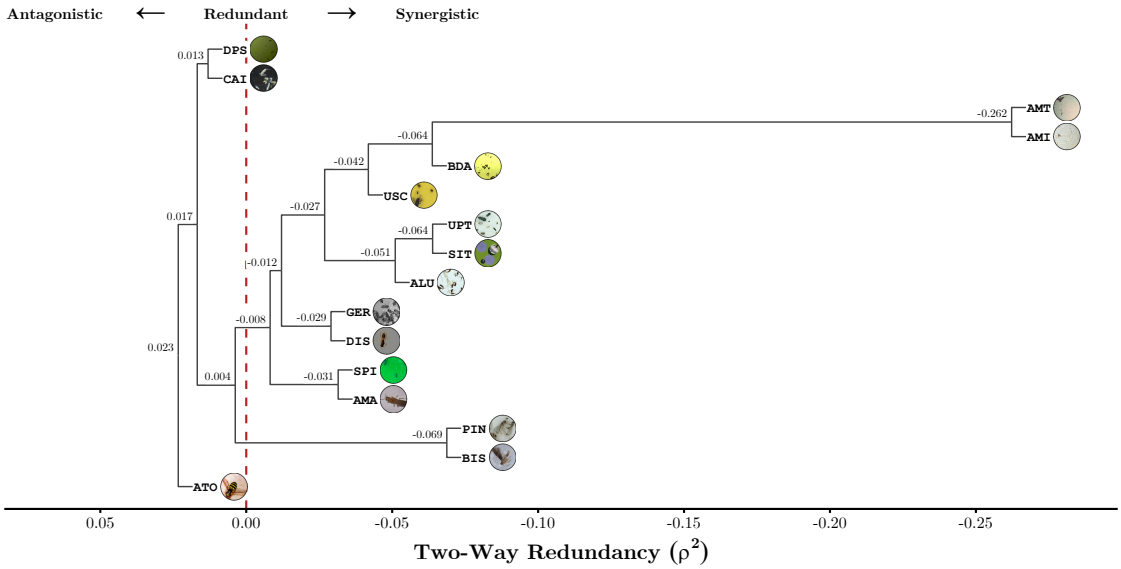


Figure 5: Hierarchical clustering of the pairwise two-way redundancy, ρ^2, between 16 out of the 23 subdatasets based on F1 performance on the validation dataset. Negative values of ρ^2 can be interpreted as synergism with a positive performance interaction between subdatasets, zero corresponds to perfectly redundant subdatasets with no performance interaction, while positive values correspond to antagonistic subdatasets with a negative performance interaction (see Section 2.5.3 for further details on ρ^2). The `flatbug` performance on all subdatasets which are joined by an internal node below zero should be considered 'resilient', meaning that the model should be able to perform relatively well on other similar datasets.

suppression of segmentation polygon IoU offers a more accurate estimate for instance overlap compared with bounding box IoU, (3) individual size estimation is much more accurate using segmentation areas than bounding box areas, especially for long-legged arthropods and (4) it opens up the possibility for semantic segmentation of individual body parts as in Mráz et al. (2024).

`flatbug` contributes to overcoming these methodological problems in multiple ways, where the three main contributions to improving the localization of arthropods in images can be summarized as instance representation, input size independence, and dataset diversity and structure.

The first key feature of `flatbug`, is the polygon-based instance representation through our strict adherence to a segmentation-first approach for the entire inference pipeline. Whereas YOLOv8 utilizes bounding boxes for deduplication (Jocher et al., 2023), `flatbug` uses the segmentation contour polygons for deduplication. To overcome and essentially trivialize the significant additional overhead associated with computing the intersection of arbitrary simple 2D polygons (segmentation contours), compared to the intersection of axis-aligned rectangles (bounding boxes), `flatbug` introduces a novel two-stage NMS (non-max suppression) algorithm.

The second key feature of `flatbug` is our solution to the scale-dependence of fixed-size inference detection models like YOLOv8 (Jocher et al., 2023), and even the hyper-inference framework SAHI (Akyon et al., 2022), that is often used to alleviate this issue. To solve this problem our hyper-inference algorithm employs a 'tile-pyramid' approach, tiling the image at geometrically spaced scales—from the native scale to the full image—using fixed-size tiles with specified overlap (Fig. 2). This differs from SAHI's single-scale tiling by allowing detection at multiple scales, improving the identification of objects of varying sizes. Our method enables scale-agnostic detection on very large images, accurately identifying all instance sizes—from the smallest visible arthropods to the largest—including intermediate sizes, rather than being limited to only the extremes.

The third key feature of `flatbug`, is the uniquely large and diverse, in terms of taxonomic coverage and image types, segmentation training dataset (Geissmann & Svenning, 2025) that forms the foundation of the `flatbug` model(s). The unique structure, size, and diversity of our dataset allows us to investigate critically needed generalized performance in an unprecedented manner, enabling robust claims about the out-of-box and out-of-domain performance. Previous methods have not been able to effectively quantify performance on novel data captured using different imaging systems and/or with different taxa, due to more narrow test datasets, often simply consisting of a subset of their specialized dataset. We show that `flatbug` can be directly applied in new contexts (without retraining) by utilizing the explicit subdomain stratification of our dataset in a leave-one-out cross-validation experiment, with only a minor performance degradation of -7.1% [-12.1%, -0.4%] relative decrease in F1 (δ_{F1}) on most datasets (Section 3.2).

The variation in performance degradation between subdatasets in the leave-one-out experiment (Fig. 4) suggests an underlying structure and similarity between subdatasets. To investigate this similarity, we expanded our analysis with a pairwise leave-two-out experiment to define a subdataset dissimilarity metric to map the domain space. In order to explicitly quantify the similarity structure, we design the novel dissimilarity metric 'two-way redundancy' (ρ^2), which describes how the performance of `flatbug` on one dataset changes, when another dataset is removed along with it. Of the 16 subdatasets we investigated in this analysis, we found that most datasets form a major synergistic cluster, with the subdatasets AMI and AMT being particularly synergistic, as is expected since these subdatasets originate from the same imaging system (Fig. 5). On the other hand, the subdatasets ATO, DPS and CAI appeared to be redundant to antagonistic with the remaining subdatasets, suggesting a considerable domain shift. This is likely explained by ATO being the only subdataset consisting of diverse images of insects on natural backgrounds (citizen science images), while CAI is unique in containing images of white insects (Collembola) on a black background (soil). It was somewhat surprising that DPS emerged as an outlier, given that its images are captured by a light-attracting moth camera trap, from which we expected it to cluster with AMI and AMT. We speculate that slightly lower-quality images on a yellowish screen (instead of white), combined with lower labelling consistency, might explain this discrepancy.

The observed generalizability and its boundaries offer a direct and simple path to improving `flatbug`, and set the stage for future research within machine learning in automated arthropod monitoring. Through community inclusion and efforts, additional datasets and annotations will likely be able to further improve `flatbug`, especially as more of the vast array of citizen science images from public platforms can be integrated. While `flatbug` is built on YOLOv8, much of our work and code is model-agnostic and can be adapted to emerging architectures. For example, YOLOv11—the successor to YOLOv8—has already been released, a trend we expect to continue. We have also demonstrated that although hyper-inference frameworks such as SAHI already exist, there is considerable scope for improvements in this area. For the method we have presented, this is particularly evident in the `flatbug` pyramid-tiling algorithm produces a very high degree of tile redundancy (Fig. 2), which could be reduced by further research, possibly leading to a significant decrease in the computational cost of inference.

`flatbug` is an open-source Python package (https://github.com/darsa-group/flat-bug/), which we believe offers cutting-edge performance and user-friendliness in its current state, due to its ease of installation, use and integration through our Python and CLI API and online demo (https://colab.research.google.com/github/darsa-group/flat-bug/blob/master/docs/flat-bug.ipynb). Crucially, unlike previous approaches `flatbug` is scale- and size-agnostic, has a broad taxonomic domains and can tackle a wide diversity of image types, although we also outline several key paths for improvement. As our experiments show, through multiple novel technological developments, `flatbug` as a tool opens up several new research avenues in entomology, conservation, monitoring, agriculture, and beyond.

5 Acknowledgements

We are thankful to a number of contributors who collected and provided previously unpublished images (and annotations) for specific sub-datasets (denoted by their three-letter code), including Nathan Pinoy; Paul Abram(ABR); Hans Jørgen Skydt Andersen and Jeppe Fogh Rasmussen(BDA); Mariana Abarca(MOI); and Tim Gernat and Gene Robinson(GER). We would also like to acknowledge testers of flatbug for their valuable feedback and help in improving the initial versions, these include: Graham Smith, Sara Nawoya, Tom August and his group, and David Rolnick and his group. The work was supported by the Global Innovation Network Program grant no. 2084-00048 (Danish Ministry of Higher Education and Science). Part of the computation done for this project was performed on the UCloud interactive HPC system, which is managed by the eScience Center at the University of Southern Denmark. The work was supported by the European Union's Horizon Europe Research and Innovation programme, under Grant Agreement No. 101060639 (MAMBO) and is partly based upon work from COST Action InsectAI CA22129, supported by COST (European Cooperation in Science and Technology). Work in the Plant Insect Ecology and Evolution Lab (J.C., D.C., N.C.) was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC), ALLRP 570736-2021, and the AgriScience Program under Agriculture and Agri-Food Canada's Sustainable Canadian Agricultural Partnership as part of Organic Science Cluster 4. Q.G. was supported by the Novo Nordisk Foundation Start Package grant (NNF22OC0077040) and the 2022-2023 BIODIVERSA+ joint call for research proposals, with the funding organisation Innovation Fund Denmark (grant no. 2128-00003).

6 Conflicts of Interest

Authors declare no conflicts of interest.

7 Author Contributions

A. Svenning, Q. Geissmann and T. T. Høye conceived the ideas. A. Svenning, G. Mougeot and Q. Geissmann contributed to the implementations. A. Svenning, G. Mougeot, Q. Geissmann and T. T. Høye led the writing. A. Svenning, D. Chevalier, G. Mougeot, J. Alison, J. Carrillo, K. Bjerge, Q. Geissmann, S. -Q. Ong and T. T. Høye contributed significantly to review and editing of the drafts. A. Svenning and G. Mougeot analysed the results. A. Svenning, N. C. Molina, N. Pinoy, Q. Geissmann and S. -Q. Ong contributed to training and evaluation data annotation. D. Chevalier, J. Alison, J. Carrillo, N. C. Molina, Q. Geissmann and S. -Q. Ong collected the data. J. Carrillo, Q. Geissmann and T. T. Høye were instrumental in project management and supervision. All authors contributed critically to the drafts and gave final approval for publication.

8 Data Availability

Data available on Zenodo 10.5281/zenodo.14761447 (Geissmann & Svenning, 2025). Code is available on GitHub https://github.com/darsa-group/flat-bug/.

9 References

[1] Akyon, F. C., Onur Altinuc, S., & Temizel, A. (2022). Slicing Aided Hyper Inference and Fine-Tuning for Small Object Detection. *2022 IEEE International Conference on Image Processing (ICIP)*, 966–970. https://doi.org/10.1109/ICIP46576.2022.9897990

[2] Alison, J., & Høye, T. T. (2024). Automated visual systems for insect monitoring and conservation. In *Routledge Handbook of Insect Conservation*. Routledge.

[3] Amarathunga, D., Ratnayake, M. N., Grundy, J., & Dorin, A. (2022). Image dataset of two morphologically close thrip species: Western Flower Thrips and Plague Thrips. https://doi.org/10.26180/21547650.v3

[4] Amarathunga, D. C., Ratnayake, M. N., Grundy, J., & Dorin, A. (2022). Fine-grained image classification of microscopic insect pest species: Western Flower thrips and Plague thrips. *Computers and Electronics in Agriculture*, *203*, 107462. https://doi.org/10.1016/j.compag.2022.107462

[5] Aodha, O. M., Gibb, R., Barlow, K. E., Browning, E., Firman, M., Freeman, R., Harder, B., Kinsey, L., Mead, G. R., Newson, S. E., Pandourski, I., Parsons, S., Russ, J., Szodoray-Paradi, A., Szodoray-Paradi, F., Tilova, E., Girolami, M., Brostow, G., & Jones, K. E. (2018). Bat detective—Deep learning tools for bat acoustic signal detection. *PLOS Computational Biology*, *14*(3), e1005995. https://doi.org/10.1371/journal.pcbi.1005995

[6] Beery, S., Morris, D., & Yang, S. (n.d.). Efficient pipeline for camera trap image review.

[7] Bradski, G. (2000). The OpenCV library. *Dr. Dobb's Journal of Software Tools*.

[8] Cao, X. (2021). ANTS–ant detection and tracking. *3*. https://doi.org/10.17632/9ws98g4npw.3

[9] A Community for Naturalists · iNaturalist. (n.d.).

[10] DiCiccio, T. J., & Efron, B. (1996). Bootstrap confidence intervals. *Statistical Science*, *11*(3), 189–228. https://doi.org/10.1214/ss/1032280214

[11] Facebookincubator/submitit. (2025, January).

[12] Gal, A., Saragosti, J., & Kronauer, D. (2020, April). anTraX: High throughput video tracking of color-tagged insects (benchmark datasets). https://doi.org/10.5281/zenodo.3740546

[13] Gal, A., Saragosti, J., & Kronauer, D. J. (2020). anTraX, a software package for high-throughput video tracking of color-tagged insects (G. J. Berman, C. Dulac, J. W. Shaevitz, & A. Perez-Escudero, Eds.). *eLife*, *9*, e58145. https://doi.org/10.7554/eLife.58145

[14] Geissmann, Q. (2022, March). Sticky Pi – Machine Learning Data, Configuration and Models. https://doi.org/10.5281/zenodo.6382496

[15] Geissmann, Q., & Svenning, A. (2025, January). Flatbug-dataset a compilation of dataset of terrestrial arthropodes on various surfaces. https://doi.org/10.5281/zenodo.14761447

[16] GEOS contributors. (2024). *GEOS computational geometry library*. Manual. https://doi.org/10.5281/zenodo.11396894

[17] 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.-Y. C., Agda, J., Ratnasingham, S., Steinke, D., Chang, A. X., Taylor, G. W., & Fieguth, P. (2023, November). A Step Towards Worldwide Biodiversity Assessment: The BIOSCAN-1M Insect Dataset. https://doi.org/10.48550/arXiv.2307.10455

[18] Gharaee, Z., Gong, Z., Pellegrino, N., Zarubiieva, I., Haurum, J. B., Lowe, S. C., T.A, M. J., Ho, C. C., McLeod, J., Wei, Y.-Y. C., Agda, J., Ratnasingham, S., Steinke, D., Chang, A. X., Taylor, G. W., & Fieguth, P. (2023, June). BIOSCAN-1M Insect Dataset. https://doi.org/10.5281/zenodo.8015206

[19] Gillies, S., van der Wel, C., Van den Bossche, J., Taves, M. W., Arnott, J., Ward, B. C., et al. (2025, January). Shapely. https://doi.org/10.5281/zenodo.5597138

[20] Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., ... Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, *585*(7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2

[21] He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2018, January). Mask R-CNN. https://doi.org/10.48550/arXiv.1703.06870

[22] Hogeweg, L., Huijbers, C., Zeegers, T., van Leeuwen, J., & Buesink, R. (2024, March). DIOPSIS camera annotated detection and classification dataset 2021. https://doi.org/10.5281/zenodo.10853097

[23] Hong, S.-J., Nam, I., Kim, S.-Y., Kim, E., Lee, C.-H., Ahn, S., Park, I.-K., & Kim, G. (2021). Automatic Pest Counting from Pheromone Trap Images Using Deep Learning Object Detectors for Matsucoccus thunbergianae Monitoring. *Insects*, *12*(4), 342. https://doi.org/10.3390/insects12040342

[24] Høye, T. T., Ärje, J., Bjerge, K., Hansen, O. L. P., Iosifidis, A., Leese, F., Mann, H. M. R., Meissner, K., Melvad, C., & Raitoharju, J. (2021). Deep learning and computer vision will transform entomology. *Proceedings of the National Academy of Sciences*, *118*(2), e2002545117. https://doi.org/10.1073/pnas.2002545117

[25] Høye, T. T., Montagna, M., Oteman, B., & Roy, D. B. (2025). Emerging technologies for pollinator monitoring. *Current Opinion in Insect Science*, *69*, 101367. https://doi.org/10.1016/j.cois.2025.101367

[26] Humphries, J. (2020). Beetle Drawer Scans. https://doi.org/10.5519/0018095

[27] Jocher, G., Qiu, J., & Chaurasia, A. (2023, January). Ultralytics YOLO.

[28] Kalamatianos, R., Karydis, I., Doukakis, D., & Avlonitis, M. (2018). DIRT: The Dacus Image Recognition Toolkit. *Journal of Imaging*, *4*(11), 129. https://doi.org/10.3390/jimaging4110129

[29] Kimura, K., & Sota, T. (2023). Evaluation of Deep Learning-Based Monitoring of Frog Reproductive Phenology. *Ichthyology & Herpetology*, *111*(4), 563–570. https://doi.org/10.1643/h2023018

[30] Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., Dollár, P., & Girshick, R. (2023, April). Segment Anything. https://doi.org/10.48550/arXiv.2304.02643

[31] Kitzes, J., Blake, R., Bombaci, S., Chapman, M., Duran, S. M., Huang, T., Joseph, M. B., Lapp, S., Marconi, S., Oestreich, W. K., Rhinehart, T. A., Schweiger, A. K., Song, Y., Surasinghe, T., Yang, D., & Yule, K. (2021). Expanding NEON biodiversity surveys with new instrumentation and machine learning approaches. *Ecosphere*, *12*(11), e03795. https://doi.org/10.1002/ecs2.3795

[32] Kloster, M., Burfeid-Castellanos, A. M., Langenkämper, D., Nattkemper, T. W., & Beszteri, B. (2023). Improving deep learning-based segmentation of diatoms in gigapixel-sized virtual slides by object-based tile positioning and object integrity constraint. *PloS One*, *18*(2), e0272103. https://doi.org/10.1371/journal.pone.0272103

[33] Mäder, P., Boho, D., Rzanny, M., Seeland, M., Wittich, H. C., Deggelmann, A., & Wäldchen, J. (2021). The Flora Incognita app – Interactive plant species identification. *Methods in Ecology and Evolution*, *12*(7), 1335–1342. https://doi.org/10.1111/2041-210X.13611

[34] Mazen, F. M. A. (2023). Arthropod Taxonomy Orders Object Detection in ArTaxOr dataset using YOLOX. *Journal of Engineering and Applied Science*, *70*(1), 113. https://doi.org/10.1186/s44147-023-00284-8

[35] Mielke Möglich, J. (2023). Example pictures AMT. https://doi.org/10.17192/fdr/194

[36] Möglich, J. M., Lampe, P., Fickus, M., Younis, S., Gottwald, J., Nauss, T., Brandl, R., Brändle, M., Friess, N., Freisleben, B., & Heidrich, L. (2023). Towards reliable estimates of abundance trends using automated non-lethal moth traps. *Insect Conservation and Diversity*, *16*(5), 539–549. https://doi.org/10.1111/icad.12662

[37] Mráz, R., Štěpka, K., Pekár, M., Matula, P., & Pekár, S. (2024). MAPHIS—Measuring arthropod phenotypes using hierarchical image segmentations. *Methods in Ecology and Evolution*, *15*(1), 36–42. https://doi.org/10.1111/2041-210X.14250

[38] North, B. V., Curtis, D., & Sham, P. C. (2002). A Note on the Calculation of Empirical P Values from Monte Carlo Procedures. *The American Journal of Human Genetics*, *71*(2), 439–441. https://doi.org/10.1086/341527

[39] Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., & Lerer, A. (2017). Automatic differentiation in PyTorch.

[40] Pl@ntNet. (n.d.).

[41] R Core Team. (2024). *R: A language and environment for statistical computing*. Manual. Vienna, Austria.

[42] Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You Only Look Once: Unified, Real-Time Object Detection. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 779–788. https://doi.org/10.1109/CVPR.2016.91

[43] Schneider, S., Fryxell, J., & ALUS (Alternative Land Use Services). (2021, October). Subsamples of ALUS Southern Ontario Insects. https://doi.org/10.5683/SP2/LMRVFN

[44] Schneider, S., Taylor, G. W., Kremer, S. C., Burgess, P., McGroarty, J., Mitsui, K., Zhuang, A., deWaard, J. R., & Fryxell, J. M. (2022). Bulk arthropod abundance, biomass and diversity estimation using deep learning for computer vision. *Methods in Ecology and Evolution*, *13*(2), 346–357. https://doi.org/10.1111/2041-210X.13769

[45] Schneider, S., Taylor, G. W., Kremer, S. C., & Fryxell, J. M. (2023). Getting the bugs out of AI: Advancing ecological research on arthropods through computer vision. *Ecology Letters*, *26*(7), 1247–1258. https://doi.org/10.1111/ele.14239

[46] Sekachev, B., Manovich, N., Zhiltsov, M., Zhavoronkov, A., Kalinin, D., Hoff, B., TOsmanov, Kruchinin, D., Zankevich, A., DmitriySidnev, Markelov, M., Johannes222, Chenuet, M., a-andre, telenachos, Melnikov, A., Kim, J., Ilouz, L., Glazov, N., ... Truong, T. (2020, August). Opencv/cvat: V1.1.0. https://doi.org/10.5281/zenodo.4009388

[47] Shepley, A., Falzon, G., Meek, P., & Kwan, P. (2021). Automated location invariant animal detection in camera trap images using publicly available data sources. *Ecology and Evolution*, *11*(9), 4494–4506. https://doi.org/10.1002/ece3.7344

[48] Sittinger, M. (2023, March). Image dataset for training of an insect detection model for the Insect Detect DIY camera trap. https://doi.org/10.5281/zenodo.7725940

[49] Stowell, D., Wood, M. D., Pamuła, H., Stylianou, Y., & Glotin, H. (2019). Automatic acoustic detection of birds through deep learning: The first Bird Audio Detection challenge. *Methods in Ecology and Evolution*, *10*(3), 368–380. https://doi.org/10.1111/2041-210X.13103

[50] Sys, S., Weißbach, S., Jakob, L., Gerber, S., & Schneider, C. (2022). CollembolAI, a macrophotography and computer vision workflow to digitize and characterize samples of soil invertebrate communities preserved in fluid. *Methods in Ecology and Evolution*, *13*(12), 2729–2742. https://doi.org/10.1111/2041-210X.14001

[51] Tan, S. K., & Wang, X. (2024). A novel two-stage omni-supervised face clustering algorithm. *Pattern Analysis and Applications*, *27*(3), 83. https://doi.org/10.1007/s10044-024-01298-5

[52] Telea, A. (2004). An Image Inpainting Technique Based on the Fast Marching Method. *Journal of Graphics Tools*, *9*(1), 23–34. https://doi.org/10.1080/10867651.2004.10487596

[53] Thomas, C. D., Jones, T. H., & Hartley, S. E. (2019). "Insectageddon": A call for more robust data and rigorous analyses. *Global Change Biology*, *25*(6), 1891–1892. https://doi.org/10.1111/gcb.14608

[54] van Klink, R., August, T., Bas, Y., Bodesheim, P., Bonn, A., Fossøy, F., Høye, T. T., Jongejans, E., Menz, M. H. M., Miraldo, A., Roslin, T., Roy, H. E., Ruczyński, I., Schigel, D., Schäffler, L., Sheard, J. K., Svenningsen, C., Tschan, G. F., Wäldchen, J., ... Bowler, D. E. (2022). Emerging technologies revolutionise insect ecology and monitoring. *Trends in Ecology & Evolution*, *37*(10), 872–885. https://doi.org/10.1016/j.tree.2022.06.001

[55] van Klink, R., Sheard, J. K., Høye, T. T., Roslin, T., Do Nascimento, L. A., & Bauer, S. (2024). Towards a toolkit for global insect biodiversity monitoring. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *379*(1904), 20230101. https://doi.org/10.1098/rstb.2023.0101

[56] Van Rossum, G., & Drake, F. L. (2009). *Python 3 reference manual*. CreateSpace.

[57] Wagner, D. L., Grames, E. M., Forister, M. L., Berenbaum, M. R., & Stopak, D. (2021). Insect decline in the Anthropocene: Death by a thousand cuts. *Proceedings of the National Academy of Sciences*, *118*(2), e2023989118. https://doi.org/10.1073/pnas.2023989118

[58] Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., ... Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686

[59] Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., ... Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*(1), 160018. https://doi.org/10.1038/sdata.2016.18

[60] Willi, M., Pitman, R. T., Cardoso, A. W., Locke, C., Swanson, A., Boyer, A., Veldthuis, M., & Fortson, L. (2019). Identifying animal species in camera trap images using deep learning and citizen science. *Methods in Ecology and Evolution*, *10*(1), 80–91. https://doi.org/10.1111/2041-210X.13099

[61] Wu, M., Cao, X., Yang, M., Cao, X., & Guo, S. (2022). A dataset of ant colonies' motion trajectories in indoor and outdoor scenes to study clustering behavior. *GigaScience*, *11*, giac096. https://doi.org/10.1093/gigascience/giac096

[62] Wu, Z., Zhang, C., Gu, X., Duporge, I., Hughey, L. F., Stabach, J. A., Skidmore, A. K., Hopcraft, J. G. C., Lee, S. J., Atkinson, P. M., McCauley, D. J., Lamprey, R., Ngene, S., & Wang, T. (2023). Deep learning enables satellite-based monitoring of large populations of terrestrial mammals across heterogeneous landscape. *Nature Communications*, *14*(1), 3072. https://doi.org/10.1038/s41467-023-38901-y

[63] Wührl, L., & Pylatiuk, C. (2021). DiversityScanner. https://doi.org/10.17605/OSF.IO/EN594

Appendix

A Technical problem definition for Section 4.3

The idea, problem statement, and mathematical framework are exclusively self conceived.

The framework for this project is based on a recontextualization of the final layer of a standard CNN classification model:

$$\log\left(P(\mathbf{C} \mid \mathbf{X}, \Theta)\right) \propto \mathbf{W}\mathbf{e} + \mathbf{b}$$

Where $\mathbf{C}$ is a column vector with the probabilities for the n classes in repertoire of a CNN classification model with parameters Θ. Here $\mathbf{X}$ is the input data, $\mathbf{W} = \Theta_{\text{last,weight}}$ is the weight matrix of the final layer, $\mathbf{b} = \Theta_{\text{last,bias}}$ is the bias vector of the final layer and $\mathbf{e}$ is a vector with the "embeddings" of the CNN backbone, corresponding to the activations of the penultimate layer.

Then, if we consider the log-probability of a single class, i, we see that:

$$\log\left(P(\mathbf{C} = i \mid \mathbf{X}, \Theta)\right) \propto \mathbf{W}_i\mathbf{e} + \mathbf{b}_i \tag{A:1}$$

$$P(\mathbf{C} = i \mid \mathbf{X}, \Theta) = \exp\left(\mathbf{W}_i\mathbf{e} + \mathbf{b}_i - \eta\right) \tag{A:2}$$

Where η is a normalization term such that the model outputs valid probabilities; $\sum P(\mathbf{C} \mid \mathbf{X}, \Theta) = 1$[13].

Thus the embeddings, $\mathbf{e}$, can be interpreted as proxy-morphological traits, and the weights, $\mathbf{W}$, can be interpreted as coefficients for the 'logistic regression models' based on these proxy-morphological traits. Similarly, if we know the prototypical species (class) traits, $\mathbf{K}$, and we assume these are sufficiently discriminatory, then it is also possible to define an a priori 'trait classification' model:

$$P(C = i \mid \mathbf{t}, \mathbf{K}_i) = \mathcal{F}(\mathbf{t}, \mathbf{K}_i) \tag{A:3}$$

The key is then: Given a function mapping from the embeddings to the actual trait space exists, $f : \mathbf{e} \to \mathbf{t}$—that we want to discover—Equation A:2 and Equation A:3 should be equal, which could be exploited under the right assumptions.[14]

Problem: Find $f : \mathbf{e} \to \mathbf{t}$, given $\mathbf{W}$, $\mathbf{b}$ and $\mathbf{K}$.

A.1 Example: Linear estimation of functional traits from embeddings

Assume that there exists a linear transformation of the embeddings, $\mathbf{e}$, (the activations of the penultimate layer) of a CNN classification model, to the trait values, $\mathbf{t}$:

$$\mathbf{t} = \mathbf{P}\mathbf{e} + \mathbf{c} \;\Leftrightarrow\; f(\mathbf{e}) = \mathbf{P}\mathbf{e} + \mathbf{c} \tag{A:4}$$

Where $\mathbf{P} \in \mathbb{R}^{|\mathbf{t}|,|\mathbf{e}|}$ is a transformation matrix and $\mathbf{c} \in \mathbb{R}^{|\mathbf{t}|}$ is an offset. Then:

$$\mathbf{P}^+(\mathbf{t} - \mathbf{c}) = \mathbf{P}^+\mathbf{P}\mathbf{e} \tag{A:5}$$

$$\mathbf{e} = \mathbf{P}^+(\mathbf{t} - \mathbf{c}) \;\Leftrightarrow\; f^{-1}(\mathbf{t}) = \mathbf{P}^+(\mathbf{t} - \mathbf{c}), \quad |\mathbf{e}| > |\mathbf{t}| \tag{A:6}$$

Where $\mathbf{P}^+$ is the pseudo-inverse of $\mathbf{P}$. Then for a converged model given Equation A:4 holds:

$$\arg\max_{\mathbf{t}}(\mathcal{F}(\mathbf{t}, \mathbf{K}_i)) = \mathbf{K}_i \quad \text{and} \quad \mathbf{W}_i\mathbf{e} + \mathbf{b}_i \propto \log \mathcal{F}(t, \mathbf{K}_i) \tag{A:7}$$

$$\Rightarrow \arg\max_{\mathbf{e}}\left(\mathbf{W}_i\mathbf{e} + \mathbf{b}_i\right) = f^{-1}(\mathbf{K}_i) = \mathbf{P}^+(\mathbf{K}_i - \mathbf{c}) \tag{A:8}$$

If $\mathbf{E}$ is constrained or sufficiently regularized, such as by batch-normalization or normalization to a unit vector, then a solution $\mathbf{e}'_i = \arg\max_{\mathbf{e}}\left(\mathbf{W}_i\mathbf{e} + \mathbf{b}_i\right) = \alpha\frac{(\mathbf{W}_i)^T}{||\mathbf{W}_i||_2}$ can be found[15], such that $\mathbf{W}' = \text{diag}\left(\mathbf{W}\mathbf{W}^T\right)^{-1/2}\mathbf{W}$, giving the equations:

$$\mathbf{P}^+(\mathbf{K}_i - \mathbf{c}) = \alpha\mathbf{W}'_i \tag{A:9}$$

$$\mathbf{K}_i = \alpha\mathbf{P}\mathbf{W}'_i + \mathbf{c} \;\Leftrightarrow\; \mathbf{K}_i^T = \alpha(\mathbf{W}'_i)^T\mathbf{P}^T + \mathbf{c}^T \tag{A:10}$$

$$\mathbf{K} = \alpha\mathbf{W}'\mathbf{P}^T + \mathbf{1}\mathbf{c}^T \tag{A:11}$$

This establishes a direct linear relationship, allowing the transformation, $\mathbf{P}$, and offsets, $\mathbf{c}$, to be estimated by solving a multivariate multiple regression problem, with the class traits, $\mathbf{K}$, as the response variables and the prototypical embeddings, $\mathbf{W}'$, as the predictors.

[13] i.e. `softmax`.

[14] I have experimented with various different approaches, but the one described here is the simplest.

[15] This fact is related to the behavior known as 'Neural Collapse' described in Papyan *et al.* [(30)].