A multi-modal dataset for insect biodiversity with imagery and DNA at the trap and individual level

John Quinto[1,2,*] **Johanna Orsholm**[3,*] **Brendan Furneaux**[4] **Tommi Mononen**[5]
Ellen Nein[3,6] **Tomas Roslin**[3,5] **Iuliia Zarubiieva**[1,2] **Akshita Gupta**[7] **Prajakta Darade**[8]
Scott C. Lowe[2] **Graham W. Taylor**[1,2]

[1]University of Guelph [2]Vector Institute [3]Swedish University of Agricultural Sciences
[4]University of Jyväskylä [5]University of Helsinki [6]Swedish Museum of Natural History
[7]Technische Universität Darmstadt [8]Indian Institute of Technology Indore

Abstract

Insects comprise millions of species, many experiencing severe population declines under environmental and habitat changes. High-throughput approaches are crucial for accelerating our understanding of insect diversity, with DNA barcoding and high-resolution imaging showing strong potential for automatic taxonomic classification. However, most image-based approaches rely on individual specimen data, unlike the unsorted bulk samples collected in large-scale ecological surveys. We present the Mixed Arthropod Sample Segmentation and Identification (MassID45) dataset for training automatic classifiers of bulk insect samples. It uniquely combines molecular and imaging data at both the unsorted sample level and the full set of individual specimens. Human annotators, supported by an AI-assisted tool, performed two tasks on bulk images: creating segmentation masks around each individual arthropod and assigning taxonomic labels to over 17,000 specimens. Combining the taxonomic resolution of DNA barcodes with precise abundance estimates of bulk images holds great potential for rapid, large-scale characterization of insect communities. This dataset pushes the boundaries of tiny object detection and instance segmentation, fostering innovation in both ecological and machine learning research.

1 Introduction and Background

Anthropogenic climate change is contributing to rapid population declines in arthropods, the most diverse group of organisms on Earth [5, 46, 7]. Unfortunately, efforts to study this crisis are hampered by a severe lack of taxonomic expertise [49, 33]. These limitations point to the need for high-throughput methods of monitoring insect communities, particularly through machine learning-based classification. Current insect datasets, which primarily contain single-specimen images [13, 14, 43], do not align with the unsorted bulk samples produced by large-scale ecological studies. Thus, there is a need for taxonomically annotated bulk-level training images.

While obtaining images from bulk samples is straightforward, subsequent analyses of these *bulk images* are challenging. First, bulk images contain many small, densely packed insects with limited morphological details, which are difficult to detect with standard computer vision approaches. Second, classifying these insects requires training data that accounts for their high taxonomic diversity.

To address these challenges, we present the Mixed Arthropod Sample Segmentation and Identification (MassID45) dataset, collected from flight interception traps deployed in Sweden and Finland in

[*]Joint first author.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: The 3rd Workshop on Imageomics: Discovering Biological Knowledge from Images Using AI.

2021. MassID45 features 45 bulk arthropod samples with corresponding high-resolution bulk images, which were then sorted into individual specimen images. For each sample, we also provide DNA metabarcoding data, which can be combined with the images to enable absolute, taxon-specific abundance estimates, while enhancing classifier performance [3, 17].

In this work, we leverage AI-assisted annotation workflows to provide detailed segmentation masks and taxonomic classification labels for over 17,000 arthropod specimens across 49 bulk images. We then demonstrate how MassID45 can be used to benchmark a less-explored challenge for standard instance segmentation models: the detection of tiny, dense, and sometimes overlapping insects. This dataset is poised to support a wide range of ecological applications, such as training automated classifiers for bulk samples, accurately counting small specimens in large collections, and enabling large-scale morphological analyses. Thus, MassID45 provides a valuable resource for both ecologists and machine learning researchers to advance automated biodiversity monitoring.

2 Dataset

MassID45 consists of 45 bulk arthropod samples collected from 19 sites across Sweden and Finland in 2021 using Townes-style Malaise traps [4]. As seen in Appendix D, the dataset is available on Zenodo. We describe the relevant bulk imaging and annotation details below.

2.1 Bulk and individual imaging protocols

The samples were first analyzed through a bulk workflow [9] without prior sorting, including DNA metabarcoding and imaging. Each sample, containing hundreds to thousands of individual specimens, was submerged in a shallow layer of ethanol and imaged as a whole in a translucent sorting tray (see Appendix A.1 for details). 41 bulk samples yielded one bulk image each. For the remaining four bulk samples, which weighed more than 10 g, the insect specimens were divided into two sorting trays, yielding two bulk images each. This process yielded 49 high-resolution bulk images, each sized 8192 $\times$ 5464 pixels. Following bulk imaging, each sample was sorted into individual specimens, which were then imaged and barcoded (see Appendix A.2 for details).

2.2 Annotation workflow

Step 1: Create segmentation masks. An AI-assisted workflow was used to annotate the numerous arthropod specimens in each bulk image, with specimen counts ranging from 36 to 3,228 per image. Similar to Schneider et al. [38], the watershed algorithm was used to obtain initial coarse masks for all potential objects in a bulk image. Due to the large number of specimens, each bulk image was split into 4 $\times$ 4 equally-sized sub-images, with overlapping borders to ensure complete arthropod coverage. Human annotators used the Toronto Annotation Suite (TORAS) [22], a web-based tool incorporating the Segment Anything Model (SAM) [23], to verify and refine the initial masks. Annotators then classified each mask as arthropod (`b` for "bug"), debris (`d`), edge artifact (`e`), or unknown (`u`), with only arthropod classifications refined and retained for analysis. The annotated sub-images were exported in MS-COCO format [28] for training detection models.

Step 2: Assign taxonomic labels. Using the taxonomy described in Appendix B.1, an expert annotator with experience in arthropod identification then assigned taxonomic labels to each insect mask. For each bulk image, a list of taxa present in the corresponding sample was compiled from the individual specimen DNA barcoding. Using these available taxa, the expert then assigned the lowest (i.e., most specific) taxonomic rank possible for each arthropod mask. To indicate uncertainty, the expert was allowed to assign several labels to each arthropod. The high-confidence (HC) label belonged to the highest taxonomic rank, while all lower-ranking labels were treated as low confidence (LC). The expert was also asked to perform a quality check of the segmentation masks, correcting annotations where insects/insect parts were missed or debris was mistaken for insects. A breakdown of the taxonomic labels, as well as the completeness of the annotations are discussed in Appendix B.2.

2.3 Machine learning dataset

In this work, we frame MassID45 as a benchmark for single-class instance segmentation on tiny, densely packed objects. Specifically, we focus on the task of localizing arthropod instances, excluding

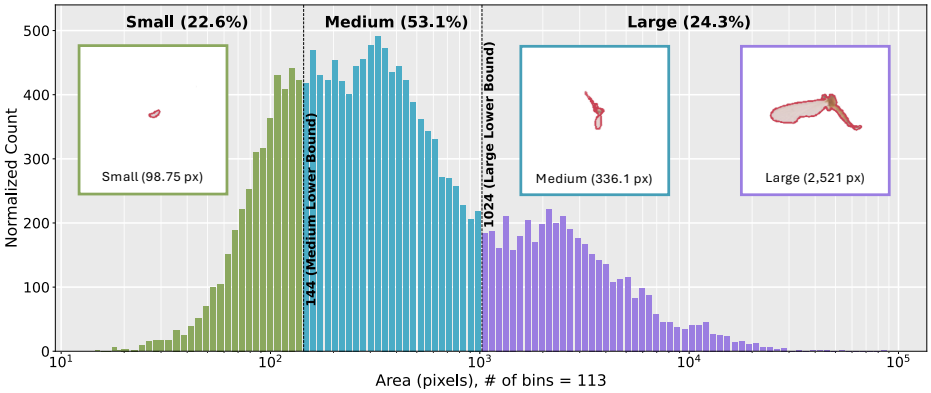

Figure 1: Distribution of insect mask areas for "small" (<144 pixels), "medium" (≥ 144 but <1024 pixels), and "large" (≥ 1024 pixels) insects. Counts are adjusted such that the area of a bar is proportional to the count in that bin. The three images show the median masks for small, medium, and large insects, all at the same magnification.

debris and other artifacts. While we follow MS-COCO evaluation conventions [28], its object size definitions are ill-suited for our data, where 76.5% of arthropod masks are "small." We therefore adopt the more granular area thresholds from the iSAID dataset [48], a remote sensing dataset for small objects: "small" (<144 pixels), "medium" (≥ 144 but <1024 pixels), and "large" (≥ 1024 pixels).

In total, the 49 fully annotated bulk images contain segmentation masks for 17,937 arthropods. Mask areas range between 15.1 and 83,182.4 pixels, with a mean of 1152.2 and and median 343.4 pixels (see Figure 1).

2.4 Preprocessing bulk images

We first merged the annotations from the sub-images back together to match the original bulk images, allowing us to divide the images into tiles as needed for model training. We then used the Shapely library [16] in Python to merge segmentation masks with multiple polygons and to correct invalid (i.e., self-intersecting) polygon masks. The insect masks were processed as concave hulls, filling in holes (e.g., areas between legs) in the edited segmentation masks to create single polygons. 63 of the 17,937 insect masks (0.351%) contained unconnected polygons that could not be merged via a unary union. In such cases, we took the polygon with the largest area as the final mask. After cleaning the segmentation masks, we manually cropped the bulk images to only contain the areas in which insects were present. This resulted in finalized cropped bulk images of different dimensions.

3 Experiments

In this work, we benchmark instance segmentation performance on MassID45 using two paradigms: zero-shot learning and supervised learning. By comparing the performance of zero-shot and supervised approaches, we can assess whether the expert annotations are valuable enough to justify the annotation effort, or whether existing generalist models achieve adequate detection performance "out-of-the-box' on the MassID45 data. We describe implementation details for our training and inference pipelines below.

3.1 Dividing bulk images into tiles

Due to GPU memory constraints and the high resolution of the images, we could not process entire bulk images during training or inference. Thus, we split the bulk images into tiles, similar to previous work [45, 10]. While the images can also be down-scaled to produce a resolution which fits within GPU memory, tiling preserves the pixel density of the original images, and maximizes the visual features available for detecting small insects.

Based on our analysis in Appendix C.2, we split the bulk images into 512×512 pixel tiles, treating each tile as a separate image during model training and inference. To ensure "cut-off" insects along the boundary of one tile were shown intact in adjacent tiles, we used an overlap of 60% between tiles, similar to previous work on small object detection [10].

During inference, the same insect can appear in multiple overlapping tiles. Treating each tile as an independent image would lead to duplicate detections and inaccurate performance estimates. To address this, we used slicing-aided hyper-inference (SAHI), to merge predictions across overlapping tiles and accurately reconstruct detections in the full bulk image [2]. The SAHI algorithm has previously been used for small object detection problems in remote sensing [31, 27, 15] and pest monitoring [11]. Overlapping predictions were considered duplicates and merged if their intersection over union (IoU) was at least 50%.

3.2 Data partitioning

We randomly partitioned the bulk images into training (40 images, 81.6%), validation (3 images, 6.1%), and testing (6 images, 12.2%) sets. After dividing the bulk images into 512×512 tiles, this resulted in 17,062 training tiles, 1244 validation tiles, and 1586 testing tiles. To prevent data leakage, all tiles from a given bulk image were assigned to the same dataset split. Including insects that were duplicated and/or partially cut between tiles, the tiled training set contained 110,520 insects, the tiled validation set 5867, and the tiled test set 6241.

Based on prior work in remote sensing and underwater imagery [10, 12], we also applied geometric and colour-based augmentations to the bulk images to improve model generalization (see Appendix C.1 for details).

3.3 Evaluation metrics

Using the predictions merged with SAHI, we calculated evaluation metrics following the MS-COCO evaluation scheme [28], which relies on IoU, precision, and recall. For a given instance mask prediction, IoU quantifies the overlap between the predicted instance masks and ground truth annotations:

$$\text{IoU} = \frac{TP_p}{TP_p + FP_p + FN_p},$$

where TP_p represents the number of predicted pixels that matched the ground truth (true positives), FN_p denotes the number of ground truth pixels missed by the prediction (false negatives), and FP_p represents the number of background pixels incorrectly labelled as part of the instance (false positives). When calculating the evaluation metrics, we used a confidence threshold ($conf_{eval}$) to filter out uncertain predictions and an IoU threshold (IoU_{eval}) to define how strictly the predicted masks must overlap with the ground truth annotations to be considered correct. That is, we categorized each predicted instance as a true positive (TP_i), false positive (FP_i), or false negative (FN_i) based on whether its IoU with the ground truth masks exceeded IoU_{eval}. Based on the instance-level categorizations, we then calculated precision and recall.

Precision was calculated as

$$\text{Precision} = \frac{TP_i}{TP_i + FP_i}.$$

It quantifies how many of the insects detected by the model were actually correct. Conversely, recall was calculated as

$$\text{Recall} = \frac{TP_i}{TP_i + FN_i}.$$

It reflects how many of the actual insect specimens were detected by the model. We calculated precision-recall curves by keeping IoU_{eval} fixed and varying $conf_{eval}$. Following the MS-COCO evaluation scheme [28], we then calculated the average precision (AP), defined as the area under the precision-recall curve, for several IoU_{eval} thresholds. Here, we report the following aggregate metrics:

- $\text{AP}_{50:5:95}$: mean of the AP values calculated across IoU_{eval} thresholds ranging from 50% to 95% in 5% increments.
- AP_{50}: AP at a fixed IoU_{eval} of 50%.

Table 1: Instance segmentation results on the MassID45 test set for the zero-shot and supervised baselines. For each mask AP metric, the top result per paradigm is **bolded**.

Paradigm	Detector	$AP_{50:5:95}$	AP_{50}	AP_{75}	$AP^{S}_{50:5:95}$	$AP^{M}_{50:5:95}$	$AP^{L}_{50:5:95}$
Zero-shot	CutLER	22.7	40.0	22.1	0.80	18.1	59.0
	Grounding DINO + SAM 2.1	**27.1**	47.6	**27.0**	1.30	**22.6**	**66.3**
	Florence-2 + SAM 2.1	16.5	28.8	16.7	3.00	12.1	41.7
	Gemini 2.0 Flash + SAM 2.1	26.2	**50.0**	23.8	**3.30**	18.3	64.2
Supervised	Mask R-CNN	42.5	**83.1**	36.6	20.0	41.6	70.4
	Mask2Former	41.4	78.7	37.4	20.5	40.0	71.1
	Mask DINO	**43.5**	80.9	**40.1**	**21.1**	**43.5**	**73.1**

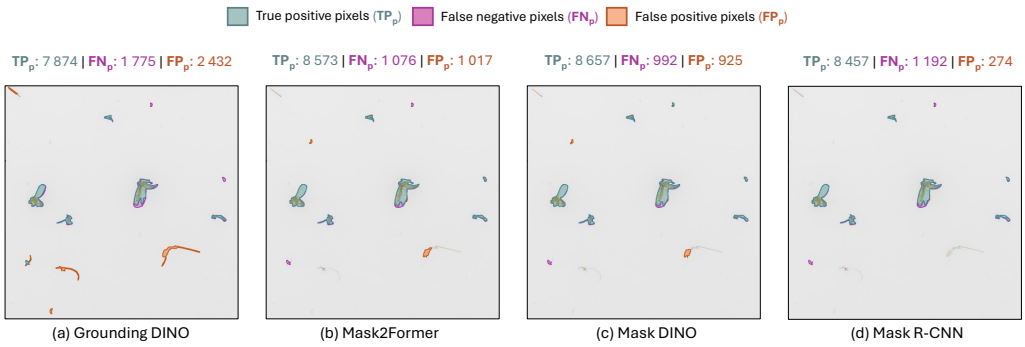

Figure 2: Visual instance segmentation results for one example patch from the MassID45 test set. For each detector, we selected distinct confidence thresholds that maximized that model's F1-score on the validation set (see Appendix C.5 for details). Predicted masks are compared for (a) the top-performing zero-shot model, Grounding DINO; and (b) – (d) the 3 supervised baselines: Mask2Former, Mask DINO, and Mask R-CNN. Above each panel, we show the areas occupied by TPs, FPs, and FNs in pixels. Best viewed on a colour display with zoom.

- AP_{75}: AP at a fixed IoU_{eval} of 75%.

We also measured $AP_{50:5:95}$ for the "small", "medium", and "large" object categories (see Section 2.3), denoting them as $AP^{S}_{50:5:95}$, $AP^{M}_{50:5:95}$, and $AP^{L}_{50:5:95}$, respectively. We report the final evaluation metrics for each baseline on the test set of six bulk images.

In practical settings, metrics like AP_{50} and AP_{75} are more informative than an aggregate metric like $AP_{50:5:95}$, as they dictate how closely the predicted masks must adhere to the true insect contours. For analyses involving specimen counts, the coarse detection of insects is sufficient; hence AP_{50} is more indicative of model performance. Conversely, AP_{75} is more important for biomass estimation tasks where tight, precise insect mask predictions are required. In practice, a model should also have a fixed operating point where only predictions above a certain confidence threshold are accepted. We derive these operating points in Appendix C.5.

3.4 Benchmarking models

Zero-Shot Detectors. We benchmark the generalization capabilities of several zero-shot detectors by applying them to a challenging new domain: small arthropods from the MassID45 data, applying the same SAHI approach for inference. These included unsupervised (CutLER [47]), open-vocabulary (Grounding DINO [29], Florence-2 [51]), and multi-modal models (Gemini 2.0 Flash [18]). For the latter three detectors, bounding box predictions were used to prompt the Segment Anything Model (SAM 2.1) [35] to produce instance masks. Full implementation details are in Appendix C.3.

Supervised Detectors. We fine-tuned three instance segmentation architectures originally developed for standard computer vision datasets like MS-COCO [28]. These models include a popular baseline for instance segmentation, Mask R-CNN [20, 50], and two recent transformer-based methods, Mask2Former [8] and Mask DINO [24]. The latter two use transformer-based architectures, an approach that has driven recent advances in computer vision [40, 36, 25], to achieve state-of-the-art results on MS-COCO. To leverage transfer learning [6], we initialized all supervised models with a ResNet-50 backbone pretrained on MS-COCO [28]. Detailed training details can be found in Appendix C.4.

4 Results

After fine-tuning on the MassID45 annotations, the supervised models significantly outperformed all zero-shot baselines (see Table 1). The top-performing supervised model, Mask DINO, achieved a mask $AP_{50:5:95}$ of 43.5%, a substantial improvement over the 27.1% achieved by the top-performing zero-shot method, Grounding DINO.

Using tailored confidence thresholds that we derived in Appendix C.5, we then visualized predictions from each model on an exemplar patch from the test set (see Figure 2). Qualitatively, Grounding DINO could successfully localize and segment larger arthropods, but missed most small insects. It also misidentified QR codes as insects (Figure 2a). In contrast, the supervised models produced instance masks that align well with the ground truth. However, the supervised models tended to confuse small, loose debris with insects and vice-versa (Figure 2b-d). For this exemplar patch, we also reported the number of true positive (TP), false positive (FP), and false negative (FN) pixels to illustrate the differences between each model's predictions. The zero-shot Grounding DINO model predicted significantly more FPs and FNs than the supervised models. Conversely, the three supervised models predicted similar numbers of FPs and FNs, with Mask DINO predicting the fewest FNs, and Mask R-CNN detecting the fewest FPs. Similar trends can be seen when aggregating the TP, FP, and FN pixels across the six bulk images in the MassID45 test set (see Table 4 in Appendix C.5).

The relatively poor performance of the zero-shot baselines suggests fine-tuning is still needed for specialized tasks like detecting arthropods from the MassID45 dataset. More importantly, this finding underscores the importance of expert annotations for bulk image analyses, where the complexities of the detection task are caused by the small size of the arthropods, as well as their high similarity to surrounding debris. While not explored in this work, fine-tuning these zero-shot methods on the MassID45 dataset may prove beneficial. Thus, this analysis frames MassID45 as a challenging benchmark dataset for custom supervised models, vision foundation models, and other zero-shot detectors, as it assesses their ability to recognize tiny, ambiguous objects rather than larger common objects that are typically considered in the literature.

5 Acknowledgments

The generation of samples were funded by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 856506; ERC-synergy project LIFEPLAN). Barcoding, metabarcoding and imaging of the samples were funded by the Swedish Environmental Protection Agency (agreement 225-20-002 with TR). This work was supported by the AI and Biodiversity Change (ABC) Global Center, which is funded by the US National Science Foundation under Award No. 2330423 and Natural Sciences and Engineering Research Council of Canada under Award No. 585136. We also acknowledge the support of the Government of Canada's New Frontiers in Research Fund Award No. NFRFT-2020-00073. PD was supported by Mitacs through the Mitacs Globalink Research Internship program. GWT was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) Award No. RGPIN-2019-04737, the Canada Research Chairs program Award No. CRC-2021-00561, and the Canada CIFAR AI Chairs program. Resources used in preparing this research were provided, in part, by the Province of Ontario, the Government of Canada through CIFAR, and companies sponsoring the Vector Institute.

References

[1] *European Nucleotide Archive*, http://identifiers.org/ena.embl:PRJEB86111. 2025.

[2] F. C. Akyon, S. O. Altinuc, and A. Temizel. "Slicing Aided Hyper Inference and Fine-tuning for Small Object Detection". In: *2022 IEEE International Conference on Image Processing (ICIP)* (2022), pp. 966–970. DOI: 10.1109/ICIP46576.2022.9897990.

[3] S. Badirli, C. Picard, G. O. Mohler, F. Richert, Z. Akata, and M. Dundar. "Classifying the unknown: Insect identification with deep hierarchical Bayesian learning". In: *Methods in Ecology and Evolution* 14 (2023), pp. 1515–1530. URL: https://api.semanticscholar.org/CorpusID:258269725.

[4] G. G. Banelyte, A. M. Farrell, H. M. K. Rogers, et al. "Global Malaise Trap Project and LIFEPLAN Malaise Sampling". In: *protocols.io* (Dec. 2023). DOI: 10.17504/protocols.io.kqdg3xkdqg25/v2. URL: https://dx.doi.org/10.17504/protocols.io.kqdg3xkdqg25/v2 (visited on 11/13/2024).

[5] O. Bánki, Y. Roskov, M. Döring, et al. *Catalogue of Life*. Version 2024-11-18. Amsterdam, Netherlands, Nov. 2024. DOI: 10.48580/dgjy9. URL: https://www.checklistbank.org/dataset/305232.

[6] S. Bozinovski. "Reminder of the First Paper on Transfer Learning in Neural Networks, 1976". In: *Informatica (Slovenia)* 44 (2020). URL: https://api.semanticscholar.org/CorpusID:227241910.

[7] P. Cardoso, P. S. Barton, K. Birkhofer, et al. "Scientists' warning to humanity on insect extinctions". In: *Biological conservation* 242 (2020), p. 108426. DOI: 10.1016/j.biocon.2020.108426.

[8] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar. "Masked-attention Mask Transformer for Universal Image Segmentation". In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022, pp. 1280–1289. DOI: 10.1109/CVPR52688.2022.00135.

[9] J. R. deWaard, S. deWaard, M. Kuzmina, et al. "LIFEPLAN Malaise Sample Metabarcoding". In: *protocols.io* (Oct. 2024). DOI: 10.17504/protocols.io.5qpvokn3xl4o/v1. URL: https://dx.doi.org/10.17504/protocols.io.5qpvokn3xl4o/v1 (visited on 11/13/2024).

[10] J. Ding, N. Xue, G.-S. Xia, et al. "Object Detection in Aerial Images: A Large-Scale Benchmark and Challenges". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44.11 (2022), pp. 7778–7796. DOI: 10.1109/TPAMI.2021.3117983.

[11] F. Fotouhi, K. Menke, A. Prestholt, et al. "Persistent monitoring of insect-pests on sticky traps through hierarchical transfer learning and slicing-aided hyper inference". In: *Frontiers in Plant Science* Volume 15 - 2024 (2024). ISSN: 1664-462X. DOI: 10.3389/fpls.2024.1484587. URL: https://www.frontiersin.org/journals/plant-science/articles/10.3389/fpls.2024.1484587.

[12] A. Galloway, D. Brunet, R. Valipour, et al. "Predicting dreissenid mussel abundance in nearshore waters using underwater imagery and deep learning". In: *Limnology and Oceanography: Methods* 20.4 (2022), pp. 233–248. DOI: 10.1002/lom3.10483. URL: https://aslopubs.onlinelibrary.wiley.com/doi/abs/10.1002/lom3.10483.

[13] Z. Gharaee, Z. Gong, N. Pellegrino, et al. "A step towards worldwide biodiversity assessment: the BIOSCAN-1M insect dataset". In: *Proceedings of the 37th International Conference on Neural Information Processing Systems*. Ed. by A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine. Vol. 36. NIPS '23. New Orleans, LA, USA: Curran Associates Inc., 2023, pp. 43593–43619. URL: https://proceedings.neurips.cc/paper_files/paper/2023/file/87dbbdc3a685a97ad28489a1d57c45c1-Paper-Datasets_and_Benchmarks.pdf.

[14] Z. Gharaee, S. C. Lowe, Z. Gong, et al. "BIOSCAN-5M: A Multimodal Dataset for Insect Biodiversity". In: *Advances in Neural Information Processing Systems*. Ed. by A. Globerson, L. Mackey, D. Belgrave, et al. Vol. 37. Curran Associates, Inc., 2024, pp. 36285–36313. URL: https://proceedings.neurips.cc/paper_files/paper/2024/file/3fdbb472813041c9ecef04c20c2b1e5a-Paper-Datasets_and_Benchmarks_Track.pdf.

[15] B. T. Gia, T. Bui Cong Khanh, H. H. Trong, et al. “Enhancing Road Object Detection in Fisheye Cameras: An Effective Framework Integrating SAHI and Hybrid Inference”. In: *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 2024, pp. 7227–7235. DOI: 10.1109/CVPRW63382.2024.00718.

[16] S. Gillies, C. van der Wel, J. Van den Bossche, M. W. Taves, J. Arnott, B. C. Ward, et al. *Shapely*. Version 2.0.7. Jan. 2025. DOI: 10.5281/zenodo.5597138. URL: https://github.com/shapely/shapely.

[17] Z. Gong, A. T. Wang, X. Huo, et al. “CLIBD: Bridging vision and genomics for biodiversity monitoring at scale”. In: *International Conference on Learning Representations*. 2025.

[18] Google DeepMind. *Introducing Gemini 2.0: our new AI model for the agentic era*. en. https://blog.google/technology/google-deepmind/google-gemini-ai-update-december-2024/#ceo-message/. [Online; accessed 23-April-2025]. Dec. 2024.

[19] B. Hardwick, D. Kerdraon, H. M. K. Rogers, et al. “LIFEPLAN: A worldwide biodiversity sampling design”. In: *PLOS ONE* 19.12 (Dec. 2025), pp. 1–15. DOI: 10.1371/journal.pone.0313353. URL: 10.1371/journal.pone.0313353.

[20] K. He, G. Gkioxari, P. Dollár, and R. Girshick. “Mask R-CNN”. In: *2017 IEEE International Conference on Computer Vision (ICCV)*. 2017, pp. 2980–2988. DOI: 10.1109/ICCV.2017.322.

[21] P. D. N. Hebert, T. W. Braukmann, S. W. Prosser, et al. “A Sequel to Sanger: amplicon sequencing that scales”. In: *BMC genomics* 19 (2018), pp. 1–14. DOI: 10.1186/s12864-018-4611-3.

[22] A. Kar, S. W. Kim, M. Boben, et al. *Toronto Annotation Suite*. https://aidemos.cs.toronto.edu/toras. 2021.

[23] A. Kirillov, E. Mintun, N. Ravi, et al. “Segment Anything”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. Oct. 2023, pp. 4015–4026. DOI: 10.1109/ICCV51070.2023.00371.

[24] F. Li, H. Zhang, H. Xu, et al. “Mask DINO: Towards A Unified Transformer-based Framework for Object Detection and Segmentation”. In: *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2023, pp. 3041–3050. DOI: 10.1109/CVPR52729.2023.00297.

[25] X. Li, H. Ding, H. Yuan, et al. “Transformer-Based Visual Segmentation: A Survey”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 46.12 (2024), pp. 10138–10163. DOI: 10.1109/TPAMI.2024.3434373.

[26] *LIFEPLAN EPA*. *BOLD Systems*. DOI: https://dx.doi.org/10.5883/DS-LPEPA22.

[27] J. Lin, H. Lin, and F. Wang. “STPM_SAHI: A Small-Target Forest Fire Detection Model Based on Swin Transformer and Slicing Aided Hyper Inference”. In: *Forests* 13.10 (2022). ISSN: 1999-4907. DOI: 10.3390/f13101603. URL: https://www.mdpi.com/1999-4907/13/10/1603.

[28] T.-Y. Lin, M. Maire, S. Belongie, et al. “Microsoft COCO: Common Objects in Context”. In: *Computer Vision – ECCV 2014*. Ed. by D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars. Cham: Springer International Publishing, 2014, pp. 740–755. ISBN: 978-3-319-10602-1.

[29] S. Liu, Z. Zeng, T. Ren, et al. “Grounding DINO: Marrying DINO with Grounded Pre-training for Open-Set Object Detection”. In: *Computer Vision – ECCV 2024*. Ed. by A. Leonardis, E. Ricci, S. Roth, O. Russakovsky, T. Sattler, and G. Varol. Cham: Springer Nature Switzerland, 2025, pp. 38–55. ISBN: 978-3-031-72970-6.

[30] I. Loshchilov and F. Hutter. “Decoupled Weight Decay Regularization”. In: *ICLR*. 2017. URL: https://openreview.net/forum?id=Bkg6RiCqY7.

[31] L. Nguyen and K. Nguyen. “YOSCA: Confidence Adjustment for Better Object Detection in Aerial Images”. In: *Vietnam Journal of Computer Science* (2025), pp. 1–20. DOI: 10.1142/S219688882450026X.

[32] J. Orsholm, J. Quinto, H. Autto, et al. *MassID45: Mixed Arthropod Sample Segmentation and Identification*. *Zenodo*. DOI: 10.5281/zenodo.15479861.

[33] D. L. Pearson, A. L. Hamilton, and T. L. Erwin. “Recovery Plan for the Endangered Taxonomy Profession”. In: *BioScience* 61.1 (Jan. 2011), pp. 58–63. ISSN: 0006-3568. DOI: 10.1525/bio.2011.61.1.11.

[34] S. Ratnasingham, C. Wei, D. Chan, et al. "BOLD v4: A Centralized Bioinformatics Platform for DNA-Based Biodiversity Data". In: *DNA Barcoding: Methods and Protocols*. Ed. by R. DeSalle. New York, NY: Springer US, 2024, pp. 403–441. ISBN: 978-1-0716-3581-0. DOI: 10.1007/978-1-0716-3581-0_26.

[35] N. Ravi, V. Gabeur, Y.-T. Hu, et al. "SAM 2: Segment Anything in Images and Videos". In: *The Thirteenth International Conference on Learning Representations*. 2025. URL: https://openreview.net/forum?id=Ha6RTeWMd0.

[36] A. M. Rekavandi, S. Rashidi, F. Boussaid, S. Hoefs, E. Akbas, and M. bennamoun. "Transformers in Small Object Detection: A Benchmark and Survey of State-of-the-Art". In: *arXiv preprint arXiv:2309.04902* (2023). DOI: 10.48550/arXiv.2309.04902.

[37] T. Ren, S. Liu, A. Zeng, et al. "Grounded SAM: Assembling Open-World Models for Diverse Visual Tasks". In: *arXiv preprint arXiv:2401.14159* (2024). DOI: 10.48550/arXiv.2401.14159.

[38] S. Schneider, G. W. Taylor, S. C. Kremer, et al. "Bulk arthropod abundance, biomass and diversity estimation using deep learning for computer vision". In: *Methods in Ecology and Evolution* 13.2 (2022), pp. 346–357. DOI: 10.1111/2041-210X.13769.

[39] C. L. Schoch, S. Ciufo, M. Domrachev, et al. "NCBI Taxonomy: a comprehensive update on curation, resources and tools". In: *Database* 2020 (2020), baaa062. DOI: 10.1093/database/baaa062.

[40] T. Shehzadi, K. A. Hashmi, D. Stricker, and M. Z. Afzal. "Object Detection with Transformers: A Review". In: *arXiv preprint arXiv:2306.04670* (2023). DOI: 10.48550/arXiv.2306.04670.

[41] L. N. Smith and N. Topin. "Super-Convergence: Very Fast Training of Residual Networks Using Large Learning Rates". In: *arXiv preprint arXiv:1708.07120* (2017). DOI: 10.48550/arxiv.1708.07120.

[42] D. Steinke, J. T. A. McKeown, A. Zyba, J. McLeod, C. Feng, and P. D. N. Hebert. "Low-cost, high-volume imaging for entomological digitization". In: *ZooKeys* 1206 (2024), pp. 315–326. ISSN: 1313-2989. DOI: 10.3897/zookeys.1206.123670.

[43] D. Steinke, S. Ratnasingham, J. Agda, et al. "Towards a Taxonomy Machine: A Training Set of 5.6 Million Arthropod Images". In: *Data* 9.11 (2024). ISSN: 2306-5729. DOI: 10.3390/data9110122. URL: https://www.mdpi.com/2306-5729/9/11/122.

[44] M. Tan and Q. Le. "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks". In: *Proceedings of the 36th International Conference on Machine Learning*. Ed. by K. Chaudhuri and R. Salakhutdinov. Vol. 97. Proceedings of Machine Learning Research. PMLR, Sept. 2019, pp. 6105–6114. URL: https://proceedings.mlr.press/v97/tan19a.html.

[45] F. Ö. Ünel, B. O. Özkalayci, and C. Çiğla. "The Power of Tiling for Small Object Detection". In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 2019, pp. 582–591. DOI: 10.1109/CVPRW.2019.00084.

[46] D. L. Wagner. "Insect Declines in the Anthropocene". In: *Annual Review of Entomology* 65.Volume 65, 2020 (2020), pp. 457–480. ISSN: 1545-4487. DOI: 10.1146/annurev-ento-011019-025151.

[47] X. Wang, R. Girdhar, S. X. Yu, and I. Misra. "Cut and Learn for Unsupervised Object Detection and Instance Segmentation". In: *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2023, pp. 3124–3134. DOI: 10.1109/CVPR52729.2023.00305.

[48] S. Waqas Zamir, A. Arora, A. Gupta, et al. "iSAID: A Large-scale Dataset for Instance Segmentation in Aerial Images". In: *IEEE/CVF Computer Vision and Pattern Recognition Workshops (CVPRW)*. 2019.

[49] Q. D. Wheeler, P. H. Raven, and E. O. Wilson. "Taxonomy: Impediment or Expedient?" In: *Science* 303.5656 (2004), pp. 285–285. DOI: 10.1126/science.303.5656.285.

[50] Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo, and R. Girshick. *Detectron2*. https://github.com/facebookresearch/detectron2. 2019.

[51] B. Xiao, H. Wu, W. Xu, et al. "Florence-2: Advancing a Unified Representation for a Variety of Vision Tasks". In: *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2024, pp. 4818–4829. DOI: 10.1109/CVPR52733.2024.00461.

Appendices

In these appendices, we provide additional details about the imaging protocols, taxonomic labeling and annotations, machine learning experiments, and usage of the MassID45 dataset. The appendices are summarized below.

- Appendix A. Additional details about the bulk and individual specimen imaging protocols.
- Appendix B. Additional details about the taxonomic annotations.
- Appendix C. Additional details about the experiments for small instance segmentation.
- Appendix D. Dataset details and usage notes.

A Imaging Protocol — Additional Details

Here we provide additional information about the image acquisition and DNA metabarcoding protocols described in Section 2.1, particularly for the bulk and individual specimen images.

A.1 Bulk imaging details

We sampled arthropod communities at 19 sites in Sweden and northern Finland using Townes-style Malaise traps [4]. We deployed the traps continuously during 2021 and emptied them once per week. The MassID45 dataset we present here constitutes a subset of 45 of these samples, collected between March 31 and October 25, 2021 (Figure 3).

After collection, samples were shipped to the Centre for Biodiversity Genomics, Guelph, Canada, where they were preserved in fresh 96% ethanol and stored at –20 °C until analysis. We weighed the arthropods from the bulk sample after filtering out the ethanol to obtain the wet biomass. We then performed non-destructive lysis for DNA extraction and collected three technical replicates from each sample. After extraction, we amplified a short (418 bp) fragment within the standard barcoding region of COI, which we then sequenced on an Illumina NovaSeq 6000. After DNA extraction, we transferred each sample to a translucent sorting tray (44 $\times$ 39 cm) with a shallow layer of ethanol and carefully spread out the specimens to minimize overlap.

After DNA extraction, we transferred each sample to a translucent sorting tray (44 $\times$ 39 cm) with a shallow layer of ethanol and carefully spread out the specimens to minimize overlap. We placed the tray on an LED panel inside a modified light cube, where the front panel was removed and a hole was added in the ceiling to fit a camera (Figure 4a). To further improve light conditions, we used two ring lights placed on opposing sides of the light cube. The bulk images were taken with a Canon EOS R5 camera and an RF 24–240 mm F4-6.3 IS USM zoom lens mounted on a large copy stand. The following camera settings were used: a focal length of 27 mm, aperture f/20, shutter speed 1/6 seconds, and ISO 100.

We manually edited the full-resolution RAW images (45 megapixels; 8192×5464) in Adobe Lightroom Classic to improve contrast and ensure visibility of both light and dark insect body parts, using the following settings: we increased exposure by 1.3 stops, set whites and highlights to −100, and shadows to 50. To restore image contrast and colour we also adjusted clarity and saturation to 20 and increased the white balance from 4200K to 5050K. To reduce noise and purple fringing, we applied luminance noise reduction and defringe values of 20. Finally, we increased sharpening to 60 and saved the images in JPEG format.

A.2 Individual specimen imaging details

After bulk imaging was completed, we placed each specimen from the bulk samples in a separate well in a 96-well microplate for individual analysis. Specimens smaller than 5 mm were placed directly in the well and imaged using a Keyence VHX-7000 Digital Microscope system [43]. For larger specimens (approximately >5 mm), we removed a single leg for DNA extraction and pinned the main body of the arthropod for imaging using an automatic Imaging Rig [42]. We amplified and sequenced full 658-bp DNA barcodes for each specimen using single-molecule real-time (SMRT) sequencing [21] on a PacBio Sequel platform. The success rate of amplification and sequencing of DNA barcodes from the individual specimens was 97.5%, though only 89.6% passed quality and

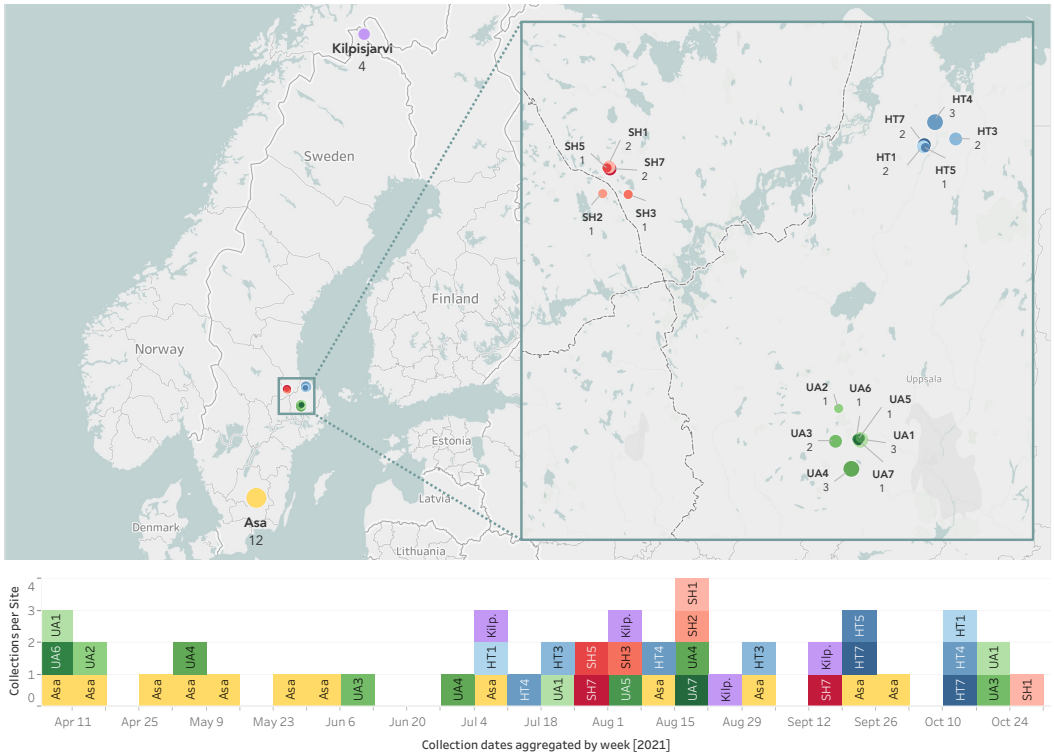

Figure 3: Geographical distribution and collection dates of samples. Each sample is uniquely named with a six-character alphanumeric code and has associated geographic and temporal information, including the latitude and longitude of the sampling site, as well as placement and collection dates. Top: Locations of the 19 sampling sites across Sweden and northern Finland. SH, HT, and UA are part of a hierarchical sampling design [19], each including 5–7 trap locations. The size of each circle is proportional to the number of samples collected at that site, which is also indicated by an integer below the trap name. Bottom: Temporal distribution of the MassID45 samples collected between March 31 and October 25, 2021. Collection dates have been aggregated by week so that samples collected during the same week are displayed in the same column, regardless of what day of the week they were collected. Hierarchically organized sites (SH, HT, and UA) are coloured with shades of the same main colours to emphasize their geographical proximity.

contamination checks. In combination with factors affecting metabarcoding success, such as primer bias and amplification of non-target DNA (e.g., gut contents), we therefore expect some discrepancies between the individual and bulk-level DNA barcoding data. We uploaded images and DNA barcodes to BOLD and assigned taxonomic classifications based on both image and molecular information using the BOLD ID engine. We retained all specimens for future morphological reference in the natural history collection of the Centre for Biodiversity Genomics (BIOUG).

B Taxonomic Labeling — Additional Details

We expand on the taxonomic labels described earlier in Section 2.2, including the sample-specific taxonomies provided to the expert annotator, and an analysis on annotation completeness.

B.1 Sample-specific taxonomies

Using individual-level DNA barcodes, we constructed sample-specific taxonomies to guide the annotations of the corresponding bulk images. First, we compiled a base taxonomy containing the ranks kingdom, phylum, subphylum, class, order, suborder, infraorder, superfamily, family, subfamily, genus, and species, starting with the taxonomy provided by BOLD [34]. We then supplemented the taxonomy with the ranks suborder, infraorder and superfamily from Dyntaxa, the Swedish taxonomy

Figure 4: (**a**) Imaging setup used to capture bulk images of the MassID45 dataset, including the positioning of the camera, light cube and ring light sources. (**b**) A representative image captured using the described imaging setup, with the sides trimmed.

database (downloaded from https://artfakta.se/), which covers all arthropods recorded from Sweden. We subset the Dyntaxa taxonomy to Hexapoda and Arachnida and combined it with the BOLD taxonomy by matching genus names within phyla and classes. Any taxonomic discordances between the taxonomies were resolved by giving the BOLD taxonomy precedence as follows. We used family names as they occurred in the BOLD taxonomy, and manually checked by comparison with the NCBI taxonomy database [39] that the suborder, infraorder, and superfamily from Dyntaxa were correct for all cases where Dyntaxa used a different family name. If NCBI listed another suborder, infraorder, or superfamily for the BOLD family name, we changed the discordant rank in our taxonomy to match NCBI. However, we did not add any information from NCBI to ranks that were empty in our taxonomy. If there was no taxonomic information for the BOLD family, we kept the information from Dyntaxa for suborder, infraorder, and superfamily. Diplura, Collembola, and Protura occurred as classes in BOLD but as orders in Dyntaxa. We kept them as classes in our taxonomy and removed all sub- and infraorders, as the same taxa appeared as orders in BOLD. The orders Phthiraptera and Psocoptera in the Dyntaxa taxonomy were combined into order Psocodea in BOLD. We therefore used the latter in our taxonomy. We also added "microlepidoptera" as an informal taxonomic group between the ranks of infraorder and superfamily. This group included 14 superfamilies within the order Lepidoptera (Adeloidea, Choreutoidea, Gelechioidea, Gracillarioidea, Micropterigoidea, Nepticuloidea, Pterophoroidea, Pyraloidea, Schreckensteinioidea, Tineoidea, Tischerioidea, Tortricoidea, Urodoidea, and Yponomeutoidea). While microlepidoptera is not a true taxonomic group, it is a useful classification when working with insect images with low resolution. Finally we generated sample-specific taxonomies by using the taxonomic classifications obtained from DNA barcoding of individual specimens in each of the 45 samples.

B.2 Annotation completeness

To evaluate how accurately the bulk image annotations reflect the true number of arthropods in the samples, we compared the number of segmentation masks annotated as arthropods with the actual number of specimens isolated from each sample (Figure 5a). We found that in samples containing more than approximately 250 arthropods, the number of arthropods based on the bulk image annotations was substantially lower than the true count. Some of these discrepancies occurred in samples with a high abundance of springtails (Collembola), which are often small, pale, and difficult to separate from debris in the bulk samples. Restricting the comparison to individual specimens classified as Insecta or Arachnida (both of which are typically larger and darker than springtails)

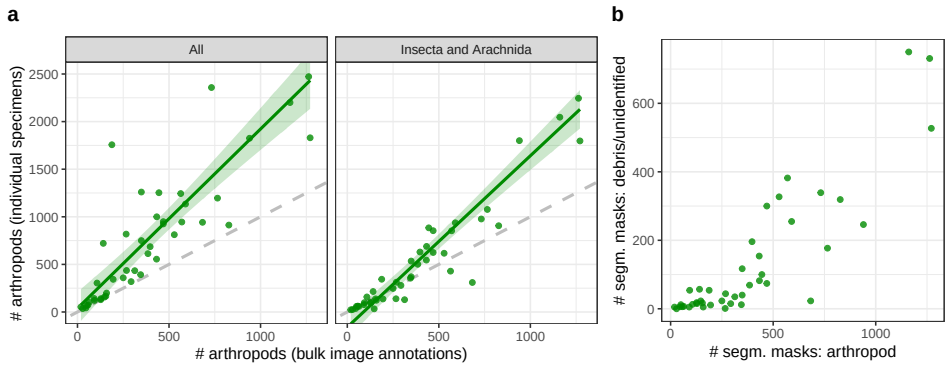

Figure 5: (**a**) For each sample, a comparison was made between the number of arthropods annotated in the bulk images and the number of individual specimens isolated from the corresponding samples, here shown for all taxa (left) and restricted to Insecta and Arachnida (right). The green line represents a linear regression fit with a 95% confidence interval between the two quantities, and the dashed grey line indicates a 1:1 relationship. (**b**) For each sample, a comparison was made between the number of segmentation masks tagged as arthropods and the number tagged as debris or unidentifiable across all bulk images.

Table 2: Number of unique taxa and specimens annotated at each taxonomic rank. Values are showed separately for high-confidence annotations (HC) and low-confidence annotations (LC).

Rank	# Taxa	$\text{Labelled}_{\text{HC}}$	$\text{Labelled}_{\text{LC}}$	$\text{Labelled}_{\text{HC}}$ (%)	$\text{Labelled}_{\text{LC}}$ (%)
Phylum	1 (1)	17,892	17,892	100.0%	100.0%
Class	4 (4)	17,570	17,834	98.2%	99.7%
Order	23 (25)	15,042	16,463	84.1%	92.0%
Suborder	8 (8)	11,218	12,745	62.7%	71.2%
Infraorder	5 (6)	7792	10,778	43.5%	60.2%
Superfamily	34 (47)	6261	8730	35.0%	48.8%
Family	92 (129)	4546	6358	25.4%	35.5%
Subfamily	27 (36)	993	1174	5.5%	6.6%
Genus	35 (55)	584	694	3.3%	3.9%
Species	17 (23)	63	74	0.4%	0.4%

reduced the difference between annotated counts and true specimen counts. Overall, the absolute discrepancy in counts tended to increase with larger samples, suggesting two possibilities. First, it is inherently challenging for human annotators to detect all insects in images where the total number of individuals is very high. These samples also often contained substantial debris, which can obscure smaller insects (Figure 5b). Additionally, because each insect occupies only a small proportion of the image, especially tiny insects may appear visually indistinct or blurry, making them difficult to annotate correctly. Second, annotator fatigue may set in for these large samples, leading to fewer corrections for arthropods missing a segmentation mask once the total count is already high.

We were able to annotate the majority of specimens in the bulk images at rank suborder or above with high confidence (Table 2). Including low-confidence annotations increased the number of specimens annotated at lower ranks, with almost half of the specimens annotated at superfamily level, and more than a third at the family level (Table 2).

C Experiments — Additional Details

Here we describe implementation details and additional analyses for the experiments described in Section 3. We delve into our data augmentation pipeline, experiments to determine the optimal upsampling factor and tile size, the technical details of our zero-shot and supervised baselines, as well as our tailored confidence thresholds from Figure 2.

Table 3: Geometric and colour-based data augmentations used for the training data, where p denotes the probability of applying each transformation.

Category	Augmentation	Parameters
Geometric	Random horizontal flip	$p = 0.5$
	Random rotation	$\{0°, 90°, 180°, 270°\}$, $p = 0.25$ each
Colour	Random brightness	Uniform in range $[-15\%, +15\%]$
	Random contrast	Uniform in range $[-10\%, +10\%]$
	Random saturation	Uniform in range $[-15\%, +15\%]$

C.1 Data augmentations

To artificially increase the number of training samples and improve generalization, we applied data augmentations to the tiled images from the training partition, drawing on prior work focused on small object detection in remote sensing and underwater imagery [10, 12]. It is important to note that our tiling process also acted as a form of data augmentation, as the arthropods could be present in multiple adjacent tiles. We employed both geometric and colour-based augmentations (Table 3), which introduced variations to the bulk images while ensuring the insects could still be identified. For example, random rotations and horizontal flips mimicked the possible orientations that arthropods can assume when placed in the sorting trays. Random adjustments to brightness, contrast, and saturation were intended to make the models more robust to small differences in lighting across bulk images, as well as natural colouration differences among arthropods (e.g., in different life stages). We applied these augmentations to the tiled bulk images, then resized each augmented tile to a fixed input size of 1024×1024 using bilinear interpolation before presenting them to the model during training.

C.2 Determining upsampling factor for tiles

If an entire bulk insect sample is downsampled to fit within a model's input size of 1024×1024 pixels, each insect is rendered at a lower resolution than in the original image, leading to blurred contours and fewer visible details — especially problematic for detecting small insects. An alternative is to divide the images into tiles to preserve visual details. Using smaller tiles than the required input size and instead upsampling the images to target resolution can affect model performance. For example, presenting images to models at higher resolutions allows the model to spend more compute in processing the full input image, potentially improving its performance [44]. Correspondingly, we investigated how much the model's performance could be increased if the original images were upsampled before presenting them to the model.

To determine the optimal upsampling factor for our instance segmentation models, we performed training and inference while varying the dimensions of the bulk image tiles. As we decreased the size of our tiles, we needed to increase the upsampling rate to reach our fixed input size of 1024×1024 pixels. We performed this analysis on the validation set using the SAHI approach to ensure this hyperparameter selection was not based on the test partition. For each trial, we maintained a fixed input size of 1024×1024 pixels, a common input size for pretraining instance segmentation models [20, 24, 8]. Tiles smaller than this input size were upsampled to 1024×1024 pixels using bilinear interpolation. Thus, a tile size of 1024×1024 pixels would require a zoom factor of one to reach our desired input size, 512×512 pixels would require a zoom factor of two, and so forth until 128×128 pixels, which would require a zoom factor of eight.

As the zoom factor increases, the relative size of the arthropods in each tile increases, although each tile includes less spatial context, and more arthropods are cut between tiles. We observed that all three models achieve the best mask AP when they use 512×512 pixel tiles or a zoom factor of two (Figure 6), which we consequently used for all further experiments.

While increasing the zoom factor from one to two improves instance segmentation performance, higher zoom factors gradually degrade performance. Very small tiles, with zoom factors of six and eight, showed the worst mask AP across all models, suggesting that the increase in relative size is offset by the lack of spatial context when small tiles are used. Such context may be important for distinguishing small insects from surrounding debris. For example, as the tile size decreases, more insects are split between tiles. These partial insects may be more difficult to distinguish from debris,

which includes loose insect legs and wings. This also complicates the inference stage as a) our models must correctly identify partial insects, and b) the SAHI algorithm must correctly merge fragmented insect predictions across tiles.

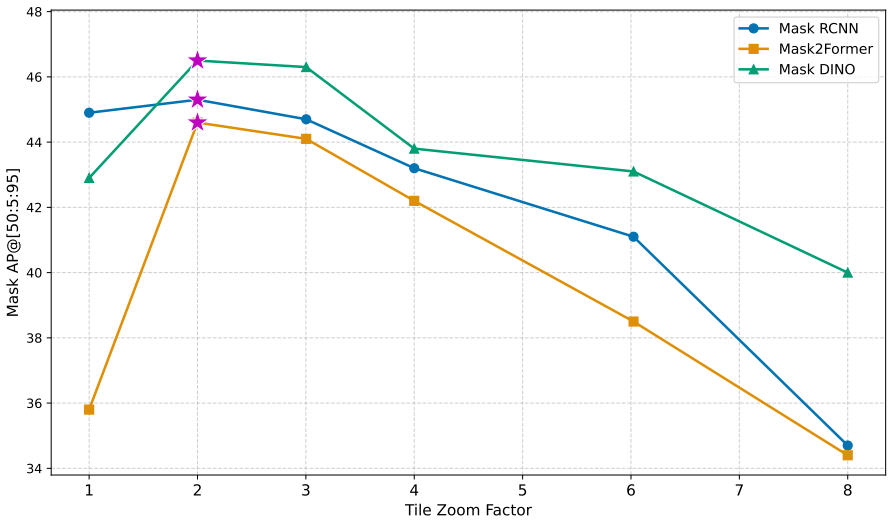

Figure 6: Validation mask AP versus tile zoom factor for our supervised baselines: Mask R-CNN, Mask2Former, and Mask DINO. For all three models, a zoom factor of two, corresponding to a tile size of 512×512 pixels, which is upsampled to a 1024×1024 pixels model input, gives the best instance segmentation performance.

C.3 Implementation details for zero-shot methods

First, we assessed the Cut and Learn (CutLER) model [47], an unsupervised instance segmentation method trained on a dataset without human annotations. CutLER leverages a self-training process where the model is iteratively trained on its own predictions to refine the quality of subsequent instance masks. For all CutLER experiments, we used a self-trained Cascade Mask R-CNN checkpoint (`cutler_cascade_final`) [47].

We then evaluated Grounding DINO [37, 29] and Florence-2 [51], which can localize objects of interest through text prompts. These text prompts can denote simple category names or referring expressions. For all Grounding DINO experiments, we provided the Grounding DINO-B model (`groundingdino_swinb_cogcoor`) [29] with the prompt "`insect.`", where "." is used as a delimiter for different object classes. We then used the default box and text thresholds of 0.35 and 0.25, respectively.

For our Florence-2 evaluations, we used the publicly available `Florence-2-large-ft` checkpoint [51]. In addition to a text prompt, Florence-2 requires a task prompt denoting whether to perform captioning, detection, or other vision-language tasks.

Thus, we provided the following prompt to Florence-2: "`<OPEN_VOCABULARY_DETECTION> small brown-yellow insects`". To suppress large bounding box predictions, we filtered out bounding boxes that occupy more than 40% of the area of a given 512×512 pixel tile (for comparison, the largest specimen in the dataset had a ground truth mask equal to 32% of a tile).

Lastly, we leveraged Gemini 2.0 Flash's spatial understanding capabilities to perform object detection [18]. With a temperature of 0.5, we provided the following system instructions: "`Return bounding boxes as a JSON array with labels. Never return masks or code fencing. Limit to 50 objects. Never repeat or duplicate bounding boxes. If an object is present multiple times, return the same label for each instance.`"

When performing detection, we used the following text prompt: "`Detect the 2d bounding boxes of the small brown insects, ants, flies, and/or gnats. Exclude loose wings, legs, and debris.`" As with Florence-2, bounding boxes occupying more than 40% of a tile

were filtered out before being used as prompts for SAM 2.1. We performed inference with the `sam2.1_hiera_large` checkpoint without any fine-tuning on the MassID45 training set [35].

C.4 Training details for supervised models

We trained all supervised models for 15,000 iterations with a batch size of 8 (2 images per GPU with 4 GPUs), using the AdamW [30] optimizer with a peak learning rate of 5×10^{-5} and weight decay of 0.05. All training runs used a one-cycle cosine annealed learning rate schedule [41] with a warm-up period of 4500 iterations. Training was performed using four NVIDIA RTX6000 GPUs. For inference, we applied the SAHI approach as described in Section 3.1, dividing the bulk images from the test partition into 512×512 pixel tiles with 60% overlap, and merging duplicate predictions with an IoU $> 50\%$.

C.5 Model evaluation with tailored confidence thresholds

To determine appropriate confidence thresholds for each model, we plotted precision-recall (PR) curves using their predictions on the MassID45 validation set (see Figure 7). We fixed the IoU threshold to 50%, indicating that we consider predicted masks as correct if they overlap by more than 50% in area with the ground truth. Each point corresponds to the precision and recall at a particular confidence threshold. Thus, for each model we selected the confidence threshold with the highest F1-score — the harmonic mean between precision and recall. This optimal confidence threshold is generally the point closest to the top-right corner of the PR curve, which represents perfect precision and perfect recall. These confidence thresholds can be interpreted as suggested operating points for each model when used on bulk images in a real-world setting.

Using these tuned confidence thresholds, we performed inference on the MassID45 test set, then filtered out any predictions below each model's confidence threshold. We then measured the number of TP, FP, and FN pixels predicted by each model on the test set (see Table 4). Consistent with our exemplar patch in Figure 2, Mask DINO predicts the highest number of TP pixels and lowest number of FN pixels, while Grounding DINO has the highest proportions of FPs and FNs. Mask R-CNN predicts the fewest FPs, while Mask2Former generally achieves a balance between Mask DINO and Mask R-CNN.

Table 4: Proportion of true positive (TP), false positive (FP), and false negative (FN) pixels for each model on the MassID45 test set after tuning confidence thresholds.

Model	**TP Area**		**FP Area**		**FN Area**	
Grounded SAM 2.1	712,478	(63.9%)	253,701	(22.8 %)	148,456	(13.3 %)
Mask2Former	783,319	**(80.7%)**	110,204	(11.4 %)	77,615	(7.99%)
Mask DINO	**787,067**	**(80.7%)**	114,215	(11.7 %)	**73,867**	**(7.57%)**
Mask R-CNN	777,120	(81.4%)	**93,473**	**(9.79%)**	83,814	(8.78%)

D Dataset availability and notes

We provide guidance for accessing and using the MassID45 dataset, detailing the included data types, the location of model checkpoints and code, as well as information for practioners using the MassID45 in their own research.

D.1 Data records

The MassID45 dataset is organized into two resolution levels (Table 5): bulk samples containing bulk images, metabarcoding data, and taxonomic image annotations and individual specimens containing individual images and DNA barcoding data. Sample metadata, bulk sample images, bulk image annotations, and models described here are all available from Zenodo [32]. Sample metadata is provided in a CSV file with one row per sample, uniquely identified by a six-character alphanumeric code. The same sample code is used as the file name of the corresponding bulk image, followed by suffix `_{image}`, where image is 1 or 2, in cases where there is more than one image per sample. We provide each image raw in CR3 format and edited in JPEG format. Bulk image

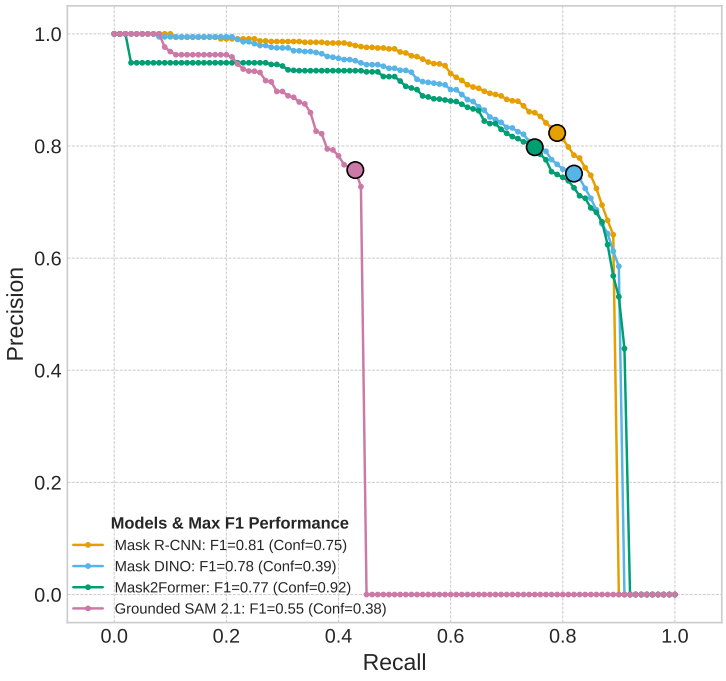

Figure 7: Precision-recall (PR) curves for the strongest zero-shot model (Grounded SAM 2.1) and the three supervised models (Mask R-CNN, Mask DINO, Mask2Former). We selected the confidence threshold for each model by finding the point on the PR curve with the highest F1-score.

annotations are available from step 1 and 2 in both COCO and TORAS format. The trained models are provided as PyTorch checkpoints and can be used for model inference with the code provided at https://github.com/uoguelph-mlrg/MassID45.

Raw sequencing reads for bulk samples are available from ENA [1] under project accession number PRJEB86111. The sequences for each sequencing replicate are represented by two gzipped FASTQ files, containing the R1 and R2 paired-end reads. Thus, for each physical sample there are a total of six files, with names of the form `{sample}_Rep{i}.R{read}.fastq.gz`, where `{sample}` is the six-character alphanumeric code uniquely identifying the sample, `{i}` is an integer between 1 and 3 indicating the replicate number, and `{read}` is 1 or 2. Accession numbers for individual samples and read files, along with a script to download all relevant files, are provided in `MassID45_ENA_accnos.tsv`, and `download_MassID45_ENA.sh`, respectively, both available at the above GitHub repository. Individual arthropod images and DNA barcode sequences are available as project ID `DS-LPEPA22` on BOLD [26]. On BOLD, the `field ID` variable corresponds to the sample code used in the sample metadata and bulk image names, while the `sample ID` is an identifier unique to each individual specimen.

D.2 Usage Notes

Our annotation workflow consisted of two separate steps, where only a subset of the annotations from step 1 (those categorized as arthropods, `b`) were annotated in step 2. In step 2, the main task of the annotator was to provide a taxonomic label for each specimen. However, if the second annotator disagreed with the first categorization, they could change it to one of the three other categories (`d`, `e`, or `u`). For the full set of annotations with both broad categories and taxonomic annotations, the output from step 1 and 2 must therefore be merged. For the taxonomic annotations, we used multiple labels as a way to express annotator uncertainty. If a single label is required, we therefore recommend careful selection of which taxon name to use.

While efforts were made to ensure the bulk images were fully annotated, some insects that were at the boundaries of the 4×4 annotator patches may have been missed. As mentioned above in Appendix B.2, insects may appear blurry in the images. This limitation to image quality can be

Table 5: Overview of data types included in the MassID45 dataset.

Resolution	Data type	Quantity	Description
Bulk samples $N = 45$	Bulk images	49 images (of 45 samples)	Images depicting unsorted insect samples, with 1–2 images per sample (41 samples 1:1, 4 samples 1:2).
	Metabarcoding data	45 samples	COI sequences from metabarcoding of unsorted insect samples. Each sample has three technical replicates.
	Taxonomic image annotations	17,940 annotations	Segmentation masks and expert taxonomic assignments for individual arthropods in the bulk images.
Individual specimens $N = 35{,}586$	Individual images	35,586 images	Images of each arthropod specimen from the 45 bulk samples.
	Barcoding data	35,586 sequences	COI sequences from DNA barcoding of individual insect specimens.

addressed by techniques like super-resolution, which reconstructs plausible high-quality details from low-resolution images. We leave this for future work.

While effective on the bulk images from the MassID45 data, the fine-tuned instance segmentation models provided in this work may not generalize to bulk images taken under different imaging protocols. Such a distributional shift would necessitate transfer learning on the user's own set of bulk images. Nevertheless, pre-trained weights from our instance segmentation models may prove beneficial for other detection tasks involving small objects. We encourage further experimentation on the MassID45 dataset, particularly with existing instance segmentation models and vision foundation models.