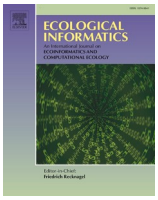
Ecological Informatics 80 (2024) 102521

Contents lists available at ScienceDirect

Ecological Informatics

journal homepage: www.elsevier.com/locate/ecolinf



STARdbi: A pipeline and database for insect monitoring based on automated image analysis

Tamar Keasar [a], Michael Yair [b], Daphna Gottlieb [c], Liraz Cabra-Leykin [d], Chen Keasar [e,*]

[a] Department of Biology, University of Haifa - Oranim, Tivon 36006, Israel
[b] Independent researcher, Harishonim 17, Ein Vered 4069600, Israel
[c] Department of Food Science, Institute of Postharvest and Food Science, The Volcani Center, ARO, Rishon-LeZion 7528809, Israel
[d] Department of Evolutionary and Environmental Biology, University of Haifa, Haifa 3498838, Israel
[e] Department of Computer Science, Ben-Gurion University of the Negev, Beer Sheva 84105, Israel

ARTICLE INFO

Keywords:
Classification
High-throughput screening
Insect monitoring
Machine learning
Object detection
Sticky trap

ABSTRACT

Insects are highly abundant and diverse, and play major roles in ecosystem functions. Monitoring of insect populations is key to their sustainable management. However, the labor and expertise needed to identify insects, and the challenges of archiving the wealth of data collected in monitoring programs, often limit these efforts. We describe a pipeline to reduce the barriers associated with curating and mining big data of insect biodiversity. The pipeline, STARdbi, includes capturing flying insects with sticky traps, scanning the traps, storing the trap-images in a public database with a web-based interface, and applying machine learning models to extract information from the images. To illustrate the insights that can be gained from STARdbi, we describe two case studies. One of them involves monitoring of circadian activity patterns of grain pests and of their natural enemies, and the other compares insect abundance, biomass and size distributions between agricultural and semi-natural habitats. We invite the community of insect ecologists to contribute to the STARdbi database, and to use its image analysis tools to address diverse ecological and evolutionary questions.

1. Introduction

Insects play a vital role in all ecosystems, occupying key positions in food webs as both herbivores and carnivores, and being major mediators of plant pollination. Thus, they also profoundly impact human lives. Blood-sucking insects transmit devastating diseases such as malaria, dengue fever, and plague, while insect pollinators are essential for our food supply. Pest insects consume our crops, but pest-feeding insects protect them. Human activity, in turn, dramatically influences insect demography, distribution and phenology. Thus, much research effort aims to detect, promote, or mitigate changes in insect populations.

Many important processes in insect ecology occur over large scales in space (e.g., long-term migrations) or time (e.g., multi-year population cycles), and hence are difficult to study with standard experimental approaches. Moreover, manipulative entomological experiments, which are frequently small-scale, often lack sufficient statistical power to detect important effects such as impacts of farming practices on insect populations. These limitations are increasingly addressed using ecoinformatics approaches (Rosenheim and Gratton, 2017), namely analyzing large datasets of observational data collected for diverse purposes. Yet, shortages in entomological field data still constrain our ability to answer key questions, such as: How do agricultural management practices impact insect pests and their natural enemies? How do they affect insect biodiversity? How do invasive insects spread? How does climate change affect the distribution of insect disease vectors? Furthermore, the scarcity of entomological big data resources often fuels controversies regarding its interpretation, such as around recent estimates of global insect declines (Sánchez-Bayo and Wyckhuys, 2019).

Three major aspects of current entomological practices limit the applicability of ecoinformatics approaches to insect studies: reliance on expert identification, discarding of bycatch, and data availability. The next three paragraphs discuss these limitations in the context of monitoring flying insects, presenting the problems and current technological approaches to their alleviation. To conclude the introduction, we present our STARdbi vision (acronym for '**S**ticky **T**raps of **AR**thropods, **d**ata**b**ase of **i**mages'), which takes a further step in embracing technology to cope with insect ecoinformatics challenges.

Insect monitoring relies heavily on visual identification of field-

* Corresponding author.
E-mail addresses: tkeasar@research.haifa.ac.il (T. Keasar), dafnag@volcani.agri.gov.il (D. Gottlieb), keasar@bgu.ac.il (C. Keasar).

https://doi.org/10.1016/j.ecoinf.2024.102521
Received 14 October 2023; Received in revised form 10 February 2024; Accepted 10 February 2024
Available online 12 February 2024

caught individuals. Several simple and effective traps exist, such as malaise traps, sticky traps, and pitfall traps. However, identifying and counting the trapped specimens is labor-intensive and requires taxonomic expertise, thus limiting the scale of monitoring. Several recent stud/ ies harnessed the power of deep-learning (*DL*) based image processing to alleviate the burden on app/ b/lied entomologists who monitor specific forestry, agricultural and medical pests (For recent reviews, see Schneider et al., 2023, Teixeira et al., 2023). Such efforts comprise four steps: (a) collection of insect images; (b) labeling of individual insects by experts, to generate two or more classes (e.g., 'pest', 'natural enemy', and 'other'); (c) building a statistical model of the classes (***training*** in the machine learning jargon); and (d) applying the model to new instances, such as field caught insects (***inference***). These projects span a range of approaches for image acquisition and analysis, as illustrated by the following examples: the insects are photographed (Ciampi et al., 2023), video-recorded (Wei and Zhan, 2024) or scanned (Júnior et al., 2022). The DL models employ sequential detection and classification stages (Rustia et al., 2021), or combine the detection and classification tasks into a single step (Wei and Zhan, 2024). Insects are grouped into broad classes in some studies (e.g., whiteflies vs. thrips, Rustia et al., 2022, wasps vs. flies, Kalfas et al., 2023). Other projects aim to distinguish between closely-related and similarly-looking species (e.g. Kittichai et al., 2021), or to identify a single focal species among all others (e.g. Salamut et al., 2023). Finally, some of the authors share their datasets of labeled insect images as resources for training of new models (e.g., Ciampi et al., 2023; Wang et al., 2020). Notwithstanding the importance of the above-mentioned studies, considering the widespread use of DL methods in other fields, the relatively scarce use of these tools in insect ecoinformatics hints at a high entrance barrier that needs to be lowered. Notably, none of the studies that we are aware of associate the trap images with meta-data (time, place etc.), and none of them offer a database to which users may add their own images.

Most insect capturing methods are not species-specific (pheromone traps are the exception). Whenever they are applied to monitor a few species of interest, most of the trapped specimens are unintended bycatch. Regardless of the identification method, manual or automated, non-focal species are typically ignored and are sometimes even discarded. Reuse of these samples for additional studies that focus on other species is complicated if possible at all. Yet, ethical as well as efficiency considerations call for maximizing the information extracted from the bycatch. Systematically studying the spatiotemporal spread of an invasive species, for example, requires numerous samples collected over much time and a wide geographical range. Such a study is close to impossible if the investigators have to actually do all the field work. Yet these specimens (or their absence) may be hiding in the bycatch of many other, apparently unrelated, studies, and even in routine agricultural or public health samples. While actual specimens are hard to share, sharing images thereof is easy.

Reducing the entrance barrier to the use of DL methods and allowing large- and even global-range surveys requires a unified framework with accessible software and datasets (Høye et al., 2021; Schneider et al., 2023). While large datasets of insect images are available for some museum collections (Marques et al., 2018) and crop pests (Ciampi et al., 2023; Wang et al., 2020), the development of databases that manage general entomological images lags behind. Here, we present the first step, to the best of our knowledge, towards this goal. We develop a database with a web interface (https://stardbi.cs.bgu.ac.il/home/welcome), based on the following principles:

1. Focusing on sticky traps as the sole capturing method, and further on a single data acquisition tool, namely standard office scanners. Consequently, the basic entities that the database stores and manipulates are uniform scans of sticky traps and their metadata (time, location, etc.). The major disadvantage of these decisions is neglect of non-flying insects, and of large ones. The latter may escape, and if caught may interfere with the scanning. On the other hand, the entry barrier is rather low: users need traps and acetate sheets (see below), at least a single office scanner per project (around $200), and an internet connection. The low price of sticky traps, and the simplicity of handling and scanning them, make them ideal for large-scale field surveys. A further reduction of the entry barrier, by adding mobile phone or other cameras as an alternative to scanners, is considered in the Discussion section below.
2. A web-based pipeline for image deposition, minimal annotation, visualization, and retrieval. The pipeline is already operational, and is described in the Methods section below.
3. An authorization system that assigns differential view/annotation/ edit permissions to users, regarding different sets of images (e.g., permission to modify images in a specific project vs. in the whole database). The first and major role of this system is preserving data integrity. Yet the policy that it applies has important implications for its usability as a public resource. The details of the authorization system are presented in the Methods section. The controversial topic of policy is suspended to the Discussion.
4. A modular classification scheme. A taxonomic classification system requires training data that includes at least hundreds of manually labeled individuals per class. Thus, a general purpose, yet high resolution, system for taxonomic classification is probably not feasible. Instead we envision a (gradually growing) set of custom-made AI-based classification models. Labeling of individual insects by experts is already implemented in STARdbi, but the other aspects of this vision - training, storage, retrieval and application of classification models - are not yet operational. We are already creating task-specific classifiers on users' requests. The performance of such a classifier is presented in the Results section as case study 1.
5. Development of biomass and biodiversity metrics. The coverage and composition of sticky traps provide estimates of biomass and biodiversity of flying insects (Schneider et al., 2022), which may serve as proxies to the overall features of the local environment. Currently STARdbi provides per-insect coverage of the sticky trap images, as well as the percentage of the trap area covered by insects. The use of such data is presented in the Results section as case study 2. Other metrics that are currently under development are discussed in the Discussion section.

2. Materials and methods

2.1. General

The STARdbi website and database is hosted on a server at Ben Gurion University, Israel. It is accessible through the URL https://stardbi.cs.bgu.ac.il/home/welcome. All code, both back and front ends, is available at https://gitlab.com/stardbi. STARdbi's unit of data management is a field survey performed by a research team, which produces images of sticky traps according to some scientific plan. Currently all our users are also data contributors. The service that we provide to these users is, to the best of our knowledge, unique. We look forward to working with users that perform larger-scale cross-survey projects as well. Below, we describe the pipeline in a top-down fashion, from the unique user's pipeline to the underlying computational infrastructure, which is based on established technology.

2.2. The Pipeline

The STARdbi pipeline (Fig. 1A) includes: (a) sampling flying insects with sticky traps, and covering them with transparent acetate sheets; (b) high resolution scanning of the traps with an office scanner; (c) uploading the scanned images and their metadata (time of placement and removal, location, scanning resolution, etc.) to the STARdbi database, using its web interface; and (d) automatic detection of insects in the images and storage of their position in the database. Once stored and processed, the images as well as bounding boxes around the detected

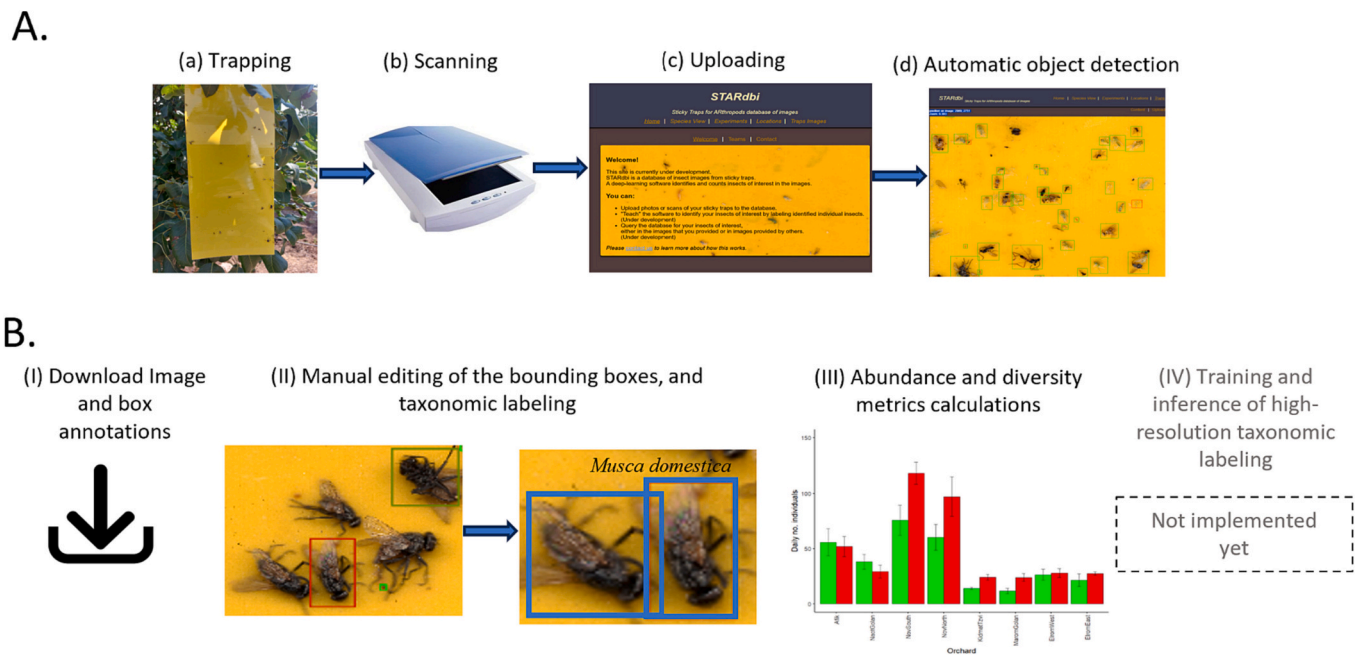


Fig. 1. Scheme of the STARdbi pipeline (A), and major user operations (B).

insects are accessible to users with the relevant authorizations (Fig. 1B). Current user operations include: (I) viewing and downloading of images and bounding box information (position, annotation, and insect-covered area), (II) manual editing of the bounding boxes, and/or taxonomically annotating them, and (III) deriving abundance and diversity metrics. We aim to implement three additional important operations: (IV) training of high resolution (e.g. species-level) classification models, (V) storing the models in the database, and (VI) applying them to the database. A detailed user view of the pipeline as well as best practices for its use can be found in Appendix 1.

2.3. The database

STARdbi uses a mySQL database (https://www.oracle.com/mysql/what-is-mysql/), with the following tables

1. The main hierarchy of tables includes monitoring projects, trap images (each associated with a project) and bounding boxes (each associated with an image).
2. Registered users are assigned differential authorizations (e.g., view, edit, annotate) in one or more projects. A user may have different authorizations in different projects.
3. A hierarchy of taxonomic tables for orders, families, genera, and species. Users may add entries to these tables and use the stored taxa in insect annotation. Users may enter the species, genus, family and order of their focal insects. Broader taxonomic levels may be used for insects that cannot be identified to species level.
4. Image locations.

2.4. Insect detection

The detection of objects within images is a fundamental task in image processing, with established DL methodology (Liu et al., 2018; Wu et al., 2019). The current version of STARdbi uses the Detectron2 implementation of Faster R-CNN (Ren et al., 2015). Given an image, object detection methods predict a set of bounding boxes around the objects and assign each of them to a class, with a confidence score. In STARdbi's object detection stage we focus on a single class, 'arthropod', and leave more detailed class assignments to later stages (see below). Even so, object detection techniques cannot be applied directly to the scanned images due to their size (> 50 megapixel). Following Gallmann et al. (2022), we perform the training and inference on 50% overlapping 2000 × 2000-pixel tiles. Bounding boxes at the tile boundaries, which are prone to errors, are removed. Overlapping bounding boxes in overlapping tiles are merged, and their confidence scores are averaged (Fig. 2).

Three major obstacles reduce the accuracy of insect detection. The first is image resolution, which limits the size of detectable insects. We currently recommend 1200dpi scans, which render some insects (e.g., tiny parasitoids of the family Mymaridae) too small to detect. Higher resolution allows detection of smaller individuals but may considerably extend handling time. Two other, related, problems are over- and under-detection. Over-detection occurs where an insect is split between two, typically overlapping, bounding Boxes (Fig. 3A). This problem may be solved by merging overlapping bounding boxes, which however worsens the related obstacle, under-detection. When two or more insects lie in close proximity, and even overlap, they may be identified as a single individual (Fig. 3B). This problem can be minimized, but not eliminated, by reducing the exposure duration of the traps. The longer the time between trap placement and removal, the more insects are caught and a larger number of them are in close proximity. The current performance of the object detection method is presented in Fig. 4.

2.5. Automatic insect classification

Automatic, AI-based, insect classification is not yet part of the STARdbi pipeline. Users may download images and their box annotations (as Comma Separated Values, csv, files), train models locally, and apply them to unlabeled data. The programs that we applied to this end (case study 1 in the Results section) are available in the STARdbi website as a Google Colab notebook. We also welcome collaborations with other researchers, in which our team will take the role of designing and trouble-shooting classification models to meet specific needs.

The available programs for the generation of classification models include:

1. STARdbi_split_images

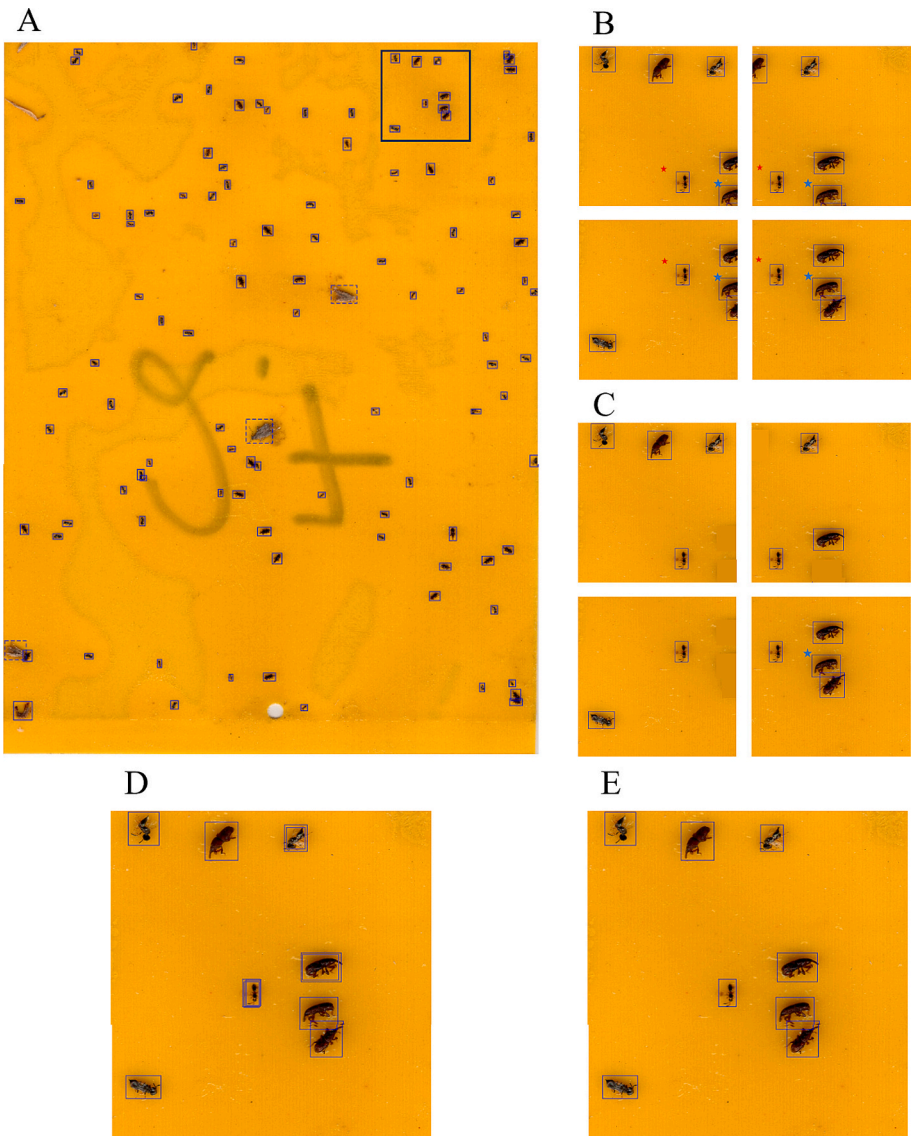


Fig. 2. Illustration of image tiling. A. A sticky trap scan, originating from a grain store with manually-curated regions of interest (bounding boxes). The image is far too large for processing by an object detection algorithm. The black frame encircles the region presented in the other panels. B. Four overlapping tiles. The starred insects appear in all four. The insect marked by a cyan star is cut by tile edges in three out of the four tiles. C. Edge cut insects are removed from the image (in both training and inference). In training we explicitly replace the bounding box by a patch with the background color. D. In the inference stage, after bounding-box prediction and removal of edge-cut boxes, the tile predictions are merged. Note that some predictions, resulting from different tiles, overlap. E. Highly overlapping bounding boxes are merged, to form the final prediction. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Input:

I. A directory with trap image files (jpeg format), and a CSV file with bounding box information, one line per box. Each box is associated with a trap image file, coordinates (in pixels) and a species label, which may be empty. These files can be downloaded from the STARdbi database.
II. An empty output directory.

Output:

The output directory is populated with subdirectories whose names correspond to the insect species that occur in the CSV files. An additional "Unlabeled" subdirectory is also created. Each subdirectory is populated by insect images of the relevant species, extracted from the trap images, based on the bounding box coordinates and annotation. The names of the images include enough information (file name and coordinates) to uniquely associate them with the CSV line from which they were derived.

2. STARdbi_train

Input:

The output directory of STARdbi_split_images, as described above,

Output:

A classification model (ResNet152, He et al. (2015)) in the form of a binary pickle file.

3. STARdbi_inference

Input:

I. A directory with unannotated single-insect image files, with the same naming convention as described in the output of STARdbi_split_images.

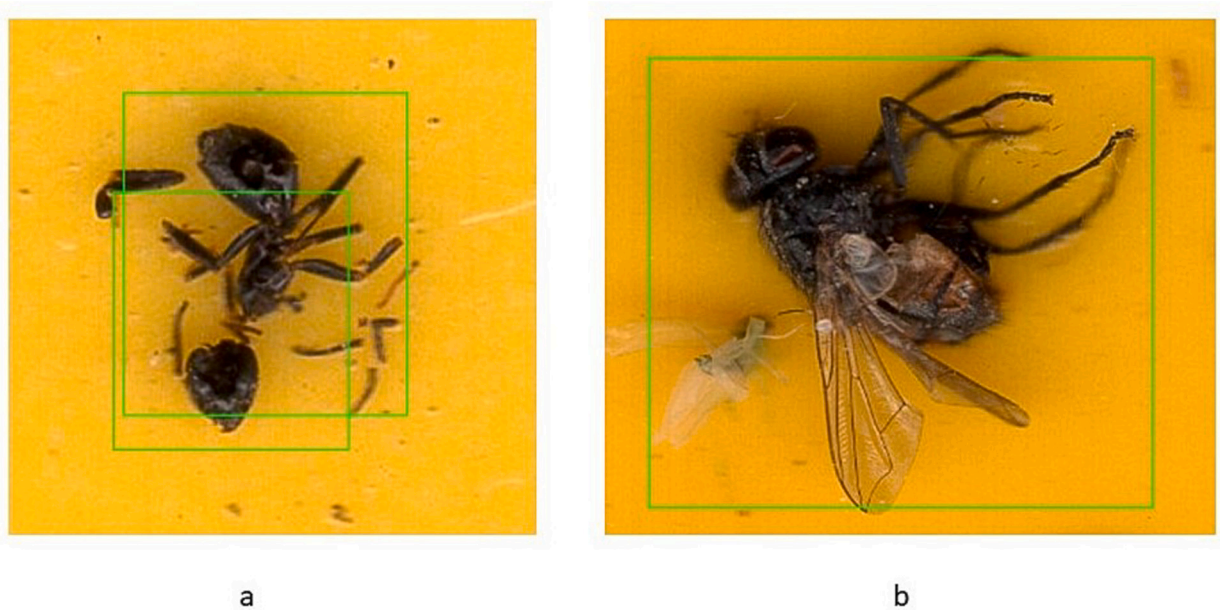


Fig. 3. The challenges of object identification. a. Over-segmentation. A single ant is mistakenly identified as two objects (green bounding boxes) b. Two individuals (a fly and a leafhopper) erroneously identified as a single object and marked by a single bounding box. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

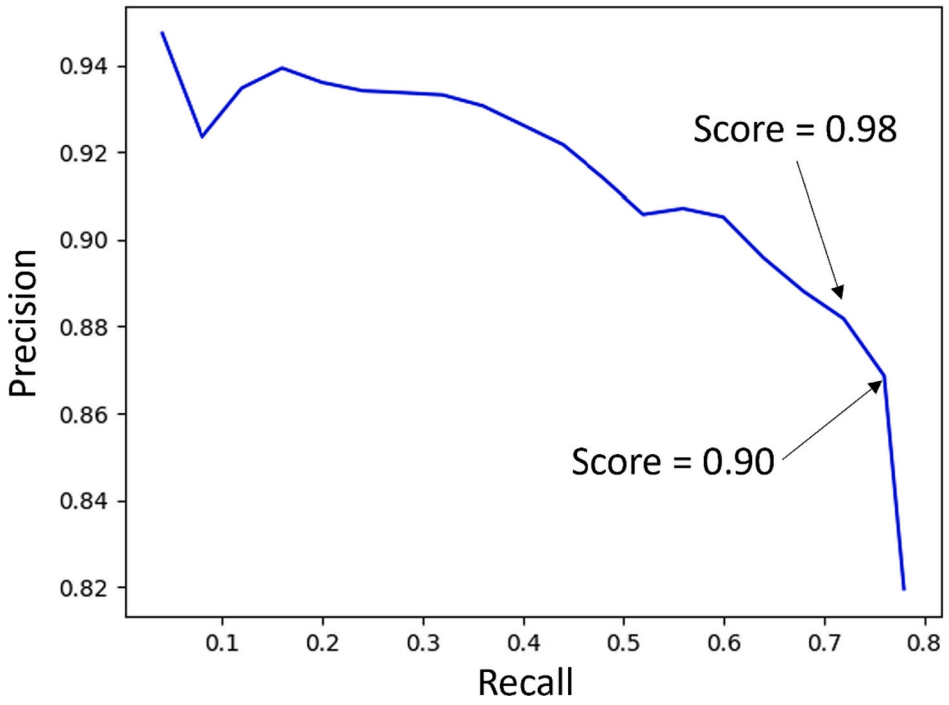


Fig. 4. Performance of the object detection model. The plot depicts bounding box precision (the fraction of insect-containing bounding boxes among all bounding boxes returned by the algorithm), as a function of recall (the proportion of all insects that were identified by the algorithm, Gerovichev et al., 2021). Each point represents an average precision of a 0.04 recall range. The precision for objects detected with a high (>0.98) confidence score is high, and drops for objects detected with lower confidence (<0.90). At the highest recall values, the confidence score is almost 0. The failure to reach higher recall is due to the identification of very close insects as a single entity (see Fig. 3). To produce the object detection model, and estimate its performance, we manually annotated all individual insects in 150 scans of sticky traps from three experiments: a grain storage, orchards (eight locations and three sampling periods along the summers of 2022 and 2023), and cages of lab reared *Ephestia* moths. Overall, the dataset includes 10,923 individual insects of diverse taxa (Lepidoptera, Coleoptera, Hymenoptera, Diptera, etc.) spanning at least an order of magnitude in size (domestic flies vs. chalcid parasitoids).

II. A CSV file in the format of STARdbi_split_images input, with lines that relate to the single-insect image files.
III. The classification model generated by STARdbi_train.

Output:

A CSV file with the same information as the input file with an additional column of predicted class and confidence score.

3. Results

STARdbi is a general-purpose pipeline that can be adapted to multiple uses. Below, we illustrate the types of insights that it can produce through two case studies.

3.1. Case study 1: Circadian activity patterns of stored grain pests and of their natural enemies

Wheat grain stores are infested by pest insects, mostly beetles and moths, which cause considerable food and economic loss. Pest control currently relies on phosphine fumigation, but the emergence of phosphine-resistant pests and environmental considerations call for reductions in insecticide use (Nayak et al., 2019). Several species of parasitoid wasps, which develop on the pests and kill them, inhabit the grain stores as well, providing an opportunity to incorporate biological control in IPM programs for stored grains (Harush et al., 2021). For IPM, it is desirable to time the application of phosphine within the grain piles to coincide with the trough of pest activity and with the peak of parasitoid flight above the pile. This would maximize the insecticide's impact on the pests, while reducing its side effects on the natural enemies. To characterize the circadian activity pattern of grain store insects, sticky traps were hung in three storage facilities in two locations, Netivot and Arugot, in southern Israel. We replaced the sticky traps every 4 h over two days (n = 192 traps). After scanning and object detection by STARdbi, we manually annotated 7070 individuals that were captured on the traps from Netivot, and randomly split them into a training and a test set. Two of the species (the parasitoids *Habrobracon hebetor* and the moth pest *Ephestia kuhniella*) are reared as insectary populations in single-species cages. We placed sticky traps in these cages as well, and thereby obtained >200 additional specimens from each species for training. Overall, the training set includes 6056 individual insects, and the test set includes 1414. We trained a classification model (ResNet152) to distinguish between the main insect classes in our sample: three beetle pests (*Oryzaephilus surinamensis, Sitophilus oryzae, and Tribolium castaneum*), one moth pest (*Ephestia kuehniella*) and three classes of parasitoid natural enemies (*Cephalonomia tarsalis*, *Habrobracon hebetor* and *Pteromalidae* sp.). The class *Pteromalidae* sp. comprises four species that could not be reliably distinguished by human experts on the scanned sticky traps (*Anisopteromalus calandrae*, *Lariophagus distinguendus*, *Pteromalus cerealellae*, and *Thelocolax elegans*). Notably, males of these parasitoids are easy to distinguish (by a red spot on their abdomen) and we could train the model to identify them. The model's classification accuracy was evaluated by comparing the model's predictions on the test dataset (randomly selected 20% of the insects, which were not part of the training data) with identifications of the same individuals by human experts. An identification is considered correct if the ratio between the intersection of their bounding boxes and their union (IoU) exceeds 0.5. Fig. 5 summarizes the model's performance as a confusion matrix, whose rows and columns represent actual and predicted classes respectively. The matrix diagonal represents true predictions, assigning insects to the correct class. Off-diagonal cells indicate false assignments. Notably, the overall accuracy is rather high. All but one of the classes are predicted with >90% accuracy. The exception is the very small class (n = 22) of male *Pteromalidae*. 60% of its members were erroneously assigned to the female *Pteromalidae* class, which was represented by many more individuals (n = 531) in the dataset. As might be expected, all but two errors confused between species within the same insect order (colored squares). We believe that even better results are likely as the database grows and more training examples become available.

To characterize the insects' daily activity schedule, we plotted frequency histograms for the taxa in the test dataset by trapping hours (illustrated in Fig. 6 for the most abundant class of pest and of natural enemies). Both insect classes were most active around mid-day, a pattern that was confirmed by analyzing the whole data set after identification by human experts. These results suggest that both pests and their natural enemies are likely to escape exposure to phosphine if it is applied to the grain mount at noon, as most insects are active and fly above the grain pile during this time. We plan to repeat these surveys at different seasons during the storage period to identify the best times to apply phosphine, when pests are within the grain mount while parasitoids are flying above it.

3.2. Case study 2: Monitoring insect communities in Rosaceae orchards

Agrophotovoltaics (APVs) are solar panels placed above agricultural fields, intended to allow parallel production of crop and power. In an ongoing study to evaluate the biodiversity impacts of this new technology, we surveyed orchard agroecosystems assigned for construction of APVs within the coming years. We used the STARdbi pipeline to characterize communities of flying insects in these plots. To this end, we compared insect trap cover, abundance and body size distributions across seasons, sites and habitats, without identifying specific species. Insects were trapped in eight sites across a steep north-south topographic and climatic gradient in northern Israel. There were three trapping rounds: in June 2022, August 2022 and May 2023. We placed traps within the orchards (n = 6 traps/site/round) and in the adjacent

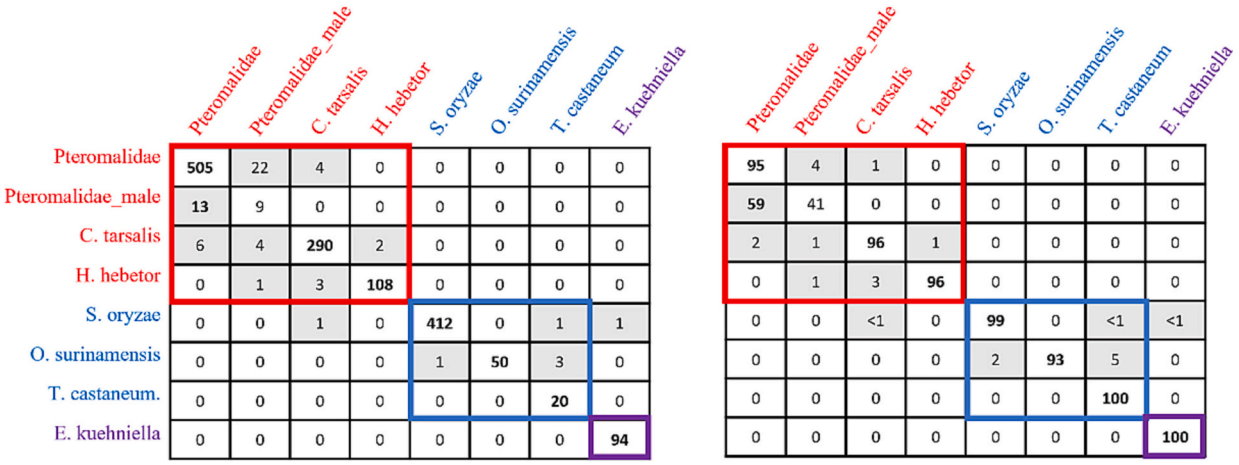


Fig. 5. Confusion matrix. Test set of the Netivot dataset. Rows - actual class, columns - predicted class, left - actual numbers, right – row-wise percentage (rounded to the closest integer), red - parasitoid wasps, blue – wheat-consuming beetles, purple - moths, bold - highest value per row, gray - prediction failure. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

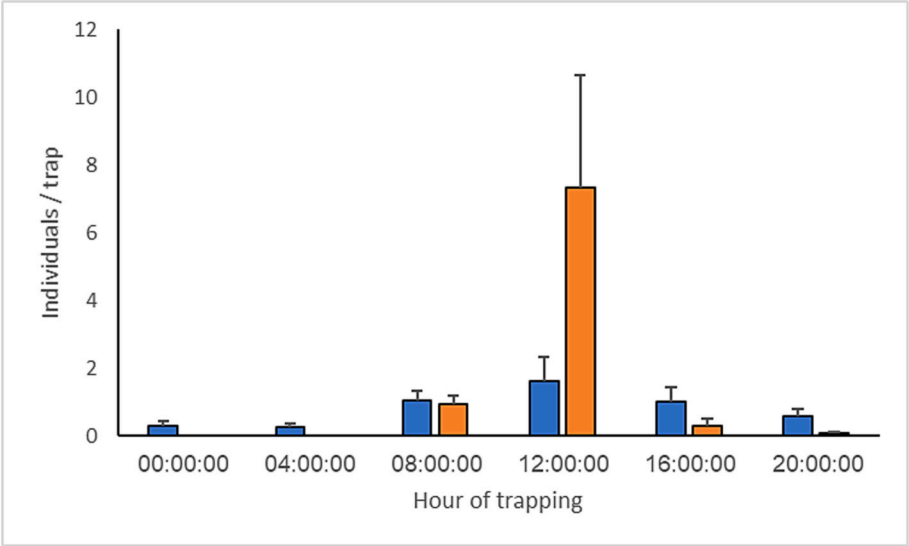


Fig. 6. Frequency distribution of mean ± SE per-trap captures by time of day, for the most abundant classes of pest (the beetle *Sitophilus oryzae*) and natural enemy (the parasitoid family Pteromalidae) in the test dataset.

semi-natural shrubland habitat ($n = 3$ traps/site/round). The traps were deposited in the Margolin House Natural History collections at Oranim College after collection and scanning.

The automated analysis of ca. 46,000 individuals captured on our traps informed us that the percent of trap area covered by insects (a proxy of biomass, Schneider et al., 2022), total insect abundance, and median community-level body size, were higher in the June sample than in May and August. Insect abundance was higher, but median body size was lower, in the orchards than in the semi-natural habitat. Northern (colder) sites tended to have fewer and smaller insects that the southern sites (Figs. 7–9). We plan to repeat these surveys after the construction of the APVs, comparing construction plots to control plots within the same orchards, to assess their effects on community-level measures of insect abundance and diversity. Such a monitoring project, which spans several years, multiple sites and several sampling locations per site, would not be possible if we relied on manual measurements. We plan to run further DL models on this set of images to detect specific insect taxa.

4. Discussion

Insect ecoinformatics is currently constrained by the time and expertise needed to identify captured specimens, as well as by the paucity of global entomological image datasets and tools to analyze them. Our database and associated web interface address these challenges by providing, for the first time, a repository of field-caught insect images that allows users to contribute new images as well as to query existing ones and their metadata. The trapped insects can be annotated, searched, counted, and morphologically characterized. In this manuscript we present two open-source and free AI-based tools, for object detection and for assessing percentage cover, which are already integrated in the web interface. Integrating the training, storage and application of taxonomic classification models is in progress. While we utilize off-the-shelf algorithms for object detection and classification, we view the combination of such models with a user-friendly web-based portal and database as the novel contribution of our work. Further, we aim at continuously updating the AI models used by STARdbi. An obvious way to do that is to retrain the models as new labeled data emerges. Specifically, despite the diversity of the object detection dataset (Fig. 4, legend), it encompasses only a tiny fraction of insect diversity. Thus, STARdbi provides a built-in mechanism for users to manually label some of their images, and trigger retraining. These innovative features are not available in existing datasets of annotated insect images (e.g., Ciampi et al., 2023; Wang et al., 2020). We also acknowledge the rapid development pace of general AI-based methods for image processing, with the speed and accuracy of the algorithms steadily improving. Thus, we aim to periodically replace the networks that are currently implemented in STARdbi by new state-of-the-art ones as better algorithms are released. With this in mind, we don't focus on STARdbi's performance compared to other AI software for insect identification. In fact, we welcome suggestions for improved models from

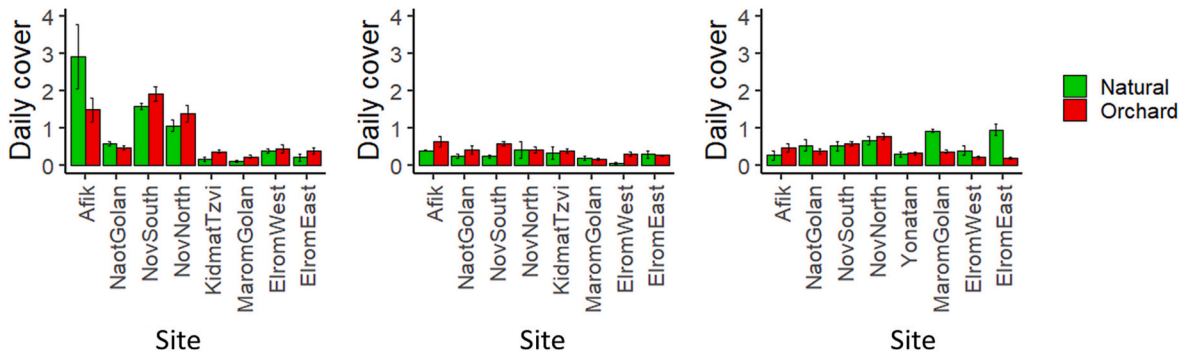


Fig. 7. Mean ± SE percent cover of the sticky traps per day in the June 2022 (left), August 2022 (center) and May 2023 (right) sampling rounds. Sampling sites are presented from south to north. Red and green bars denote the orchard and semi-natural habitats, respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

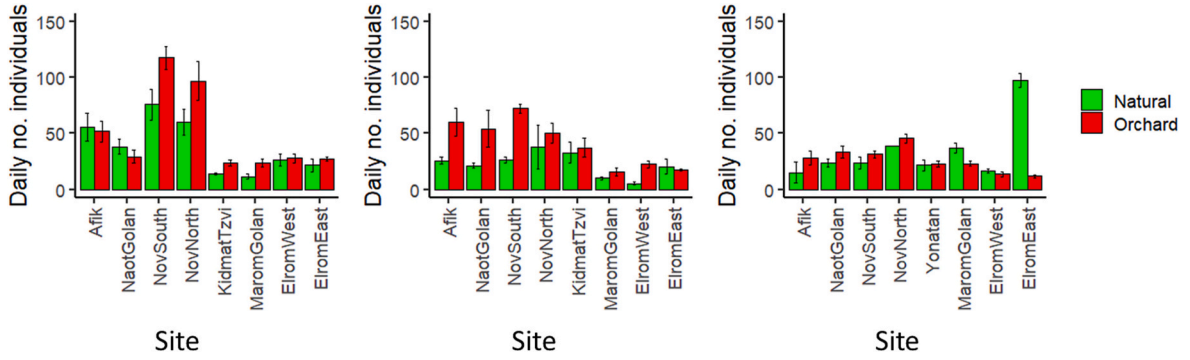


Fig. 8. Mean ± SE number of insects captured/trap/day in the June 2022 (left), August 2022 (center) and May 2023 (right) sampling rounds.

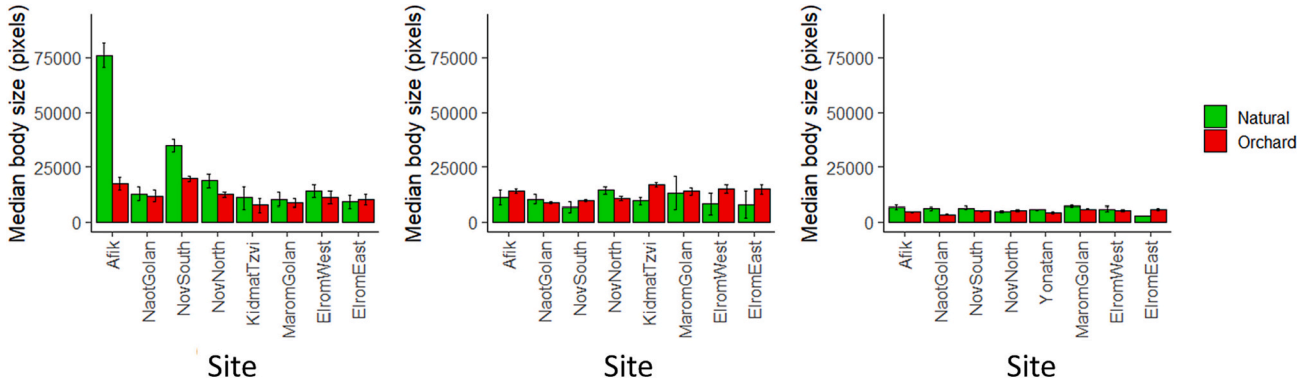


Fig. 9. Median ± SE body size of the captured insect assemblage in the June 2022 (left), August 2022 (center) and May 2023 (right) sampling rounds.

members of the DL community engaged in insect detection and classification. Future improvement may include, for example, models to track the accumulation of trapped insects over time (Geissmann et al., 2022; Rustia et al., 2020), or to identify insects at different taxonomic resolutions (Bjerge et al., 2023). In the next paragraphs, we emphasize STARdbi's prospects and limitations, independent of the specific image processing software that it implements.

The modest price tag of using our pipeline allows large-scale monitoring even when manpower and funds are limited. The use of office scanners for image acquisition, an important component in the process, has already been explored in earlier studies (e.g., Qiao et al., 2008; Xia et al., 2015). Ongoing efforts for high-quality scanning of sticky traps with mobile phones (Faria et al., 2021; Rosado et al., 2022) can make image acquisition even more accessible. Fully automated systems such as camera-equipped traps (e.g., Geissmann et al., 2022; Preti et al., 2021; Rustia et al., 2020; Wang et al., 2020) require less human labor for data acquisition, and provide better temporal resolution. However, they are considerably more expensive and require task-specific equipment, limiting the number of sites that can be monitored. We view the insights gained from fully automated monitoring approaches and from STARdbi as complementary.

The data collected by STARdbi has important inherent limitations. The most severe one, perhaps, is that the workflow only considers arthropods captured on sticky traps. Arthropod images from other sources (e.g. museum specimens, malaise trap catches) are currently left out of the database, and are not training or prediction targets for our machine learning models. In addition, the specimens captured on sticky traps cannot be moved or rotated for inspection, and this complicates their identification in some cases. For example, four species of pteromalid parasitoids in our grain store case study could not be reliably distinguished (neither by entomologists nor by DL models) on the sticky traps but were readily identified by experts when collected from pitfall traps. In other cases, the insects of interest have no congeners in the monitored habitats and can therefore be confidently assigned to species after capture on sticky traps (e.g., Gerovichev et al., 2021 for invasive *Eucalyptus* pests). Finally, captures on the sticky traps are limited to flying or ballooning species and are affected by the traps' color and their placement height. Thus, the insect assemblages on the traps comprise a biased sample of the natural community. However, this limitation is not unique to STARdbi, since all insect monitoring methods have trapping biases. The images stored in the database allow detection of spatial and temporal trends in the caught assemblages, although the trapped species comprise only a subset of the local insect community.

A major component that is still missing from STARdbi is insect classification and identification. It is a standard image processing task, and we (as well as others) provide stand-alone tools to train and use AI-models to complete it (see case study 1 above). However, integrating these utilities into the STARdbi infrastructure is a challenge. An all-in-one model, which identifies all the species that are of interest to any entomologist, is probably not feasible. There are far too many of them. Instead, we will have to make do with multiple models, each identifying either a wide range of low-resolution taxa (e.g., orders), a hierarchy of taxomonic levels (Bjerge et al., 2023), or a limited number of focal species. Allowing users to upload their own models is, unfortunately, a major cyber-security threat, as such models are full-fledged programs that may hide malware. The alternative, towards which we are leaning, is training and storing models using an interface provided by the site and its computational resources.

A major aspect of any data driven project is data availability. Data-science thrives on data, and ecoinformatics is no exception. STARdbi's vision of large-scale surveys of species and biodiversity assessment requires that its raw and processed data be available to the entomological community and the general public. Typically, however, the generators of data wish to exhaust the publication potential of their data before making it publicly available. Further, governments, the major sponsors of most research, may restrict data sharing to protect political interests.

Similar considerations slowed the development of other data-intensive fields. The solution lies in regulation, either by top journals that consider data sharing as a publication requirement, or by funding agencies wishing to maximize the impact of their investment (Michener, 2015). STARdbi will urge its users to make their data as open as possible. We start by raising this issue here.

We have illustrated STARdbi's potential applications for sustainable pest control (Results, case study 1) and conservation of insect diversity (case study 2). We envision additional, more ambitious, uses of the database in the future. One of them involves applying the DL models across several monitoring projects, rather than to one specific project. This would allow, for example, documenting the distribution of a species of interest across habitats or countries, based on traps that had been originally placed for other purposes. Secondly, the current software tool that measures insect sizes can be extended to assess other morphological features of the captured individuals, e.g., size ratios between specific body parts, or color patterns. This would allow placing each analyzed individual within a multi-dimensional trait space. We envision using statistical techniques for dimensionality reduction to group similar-looking insects into clusters, based on their multiple features. The number of clusters and their relative sizes can provide indicators of biodiversity that do not require individual identification. Such biodiversity proxies can provide much-needed tools for rapid and cost-effective diversity assessment for conservation and development projects. Finally, we aim to develop STARdbi also as a teaching resource for citizen science projects. We hope to stimulate the interest of school and college students in monitoring and learning about insect ecology at their doorstep.

Funding

The computational parts of this work were supported by the Data Science Research Center, University of Haifa, by the Israeli Council for Higher Education (CHE) via the Data Science Research Center at Ben-Gurion University of the Negev, and by the Ministry of Science and Technology's India-Israel collaboration program grant number 6294. The computational infrastructure is provided by the Department of Computer Science, Ben Gurion University. Insect sampling was supported by the ICA foundation and by the Israel Ministry of Energy and Infrastructure.

CRediT authorship contribution statement

Tamar Keasar: Writing – review & editing, Writing – original draft, Supervision, Project administration, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Michael Yair:** Writing – review & editing, Software. **Daphna Gottlieb:** Writing – review & editing, Investigation, Data curation. **Liraz Cabra-Leykin:** Writing – review & editing, Investigation, Data curation. **Chen Keasar:** Writing – review & editing, Writing – original draft, Software, Project administration, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization.

Data availability

Image data and code are provided in the STARdbi portal: https://stardbi.cs.bgu.ac.il/home/welcome

Acknowledgements

The authors are grateful to Guy Shani for thoughtful discussions and consultation, to Guy Anchelovich, Guy Braier and Alon Avraham for technical support, and to the devoted honors project students Lior Kotler, Ori M. Shalhon, Ido Yacov, and Tamir Blumberg. Dr. Animesha Rath, Miriam Benita, and Ariel Menachem contributed to the collection and identification of the specimens.

Appendix A. Appendix 1

Practical considerations and how-to:

a. Sticky traps are placed, and removed after a predefined time period. While case-specific, this time period has a considerable effect on the usability of the collected data. While too short periods result in specimen scarcity, long periods increase the frequency of close and even overlapping individuals. Human and machine annotators alike have a hard time coping with close proximity and overlaps. Some good practices:
   i. A rule of thumb: Reduce trap exposure time. Better place two traps than double placement period.
   ii. Stick a note with essential metadata (e.g., position and times of placement and removal) to the trap itself. A mobile application for the generation of machine-readable metadata barcode is on its way. The note's background color should be different then the trap color, to reduce the risk that the marks on the note are identified as insects.
   iii. Use one-sided traps. The traps are somewhat transparent, and double-sided traps are hard to annotate.
   iv. Reduce bycatch of small vertebrates (such as lizards and birds) by using narrow traps and by attaching them to surfaces with raised edges.
   v. Place an acetate sheet on the trap, trying to minimize caught air bubbles that interfere with image annotation. Once covered, the trap is easy to handle and scan.
   vi. Scan the traps using a standard office scanner. We currently recommend scanning at a resolution of 1200 dpi.

References

Bjerge, K., Geissmann, Q., Alison, J., Mann, H.M., Høye, T.T., Dyrmann, M., Karstoft, H., 2023. Hierarchical classification of insects with multitask learning and anomaly detection. Eco. Inform. 77, 102278.

Ciampi, L., Zeni, V., Incrocci, L., Canale, A., Benelli, G., Falchi, F., Amato, G., Chessa, S., 2023. A deep learning-based pipeline for whitefly pest abundance estimation on chromotropic sticky traps. Eco. Inform. 78, 102384 https://doi.org/10.1016/j.ecoinf.2023.102384.

Faria, P., Nogueira, T., Ferreira, A., Carlos, C., Rosado, L., 2021. AI-powered mobile image acquisition of vineyard insect traps with automatic quality and adequacy assessment. Agronomy 11 (4), 731. https://doi.org/10.3390/agronomy11040731.

Gallmann, J., Schüpbach, B., Jacot, K., Albrecht, M., Winizki, J., Kirchgessner, N., Aasen, H., 2022. Flower mapping in grasslands with drones and deep learning. Front. Plant Sci. 12, 774965 https://doi.org/10.3389/fpls.2021.774965.

Geissmann, Q., Abram, P.K., Wu, D., Haney, C.H., Carrillo, J., 2022. Sticky pi is a high-frequency smart trap that enables the study of insect circadian activity under natural conditions. PLoS Biol. 20 (7), e3001689 https://doi.org/10.1371/journal.pbio.3001689.

Gerovichev, A., Sadeh, A., Winter, V., Bar-Massada, A., Keasar, T., Keasar, C., 2021. High throughput data acquisition and deep learning for insect ecoinformatics. Front. Ecol. Evol. 9, 600931 https://doi.org/10.3389/fevo.2021.600931.

Harush, A., Quinn, E., Trostanetsky, A., Rapaport, A., Kostyukovsky, M., Gottlieb, D., 2021. Integrated pest management for stored grain: potential natural biological control by a parasitoid wasp community. Insects 12 (11), 1038. https://doi.org/10.3390/insects12111038.

He, K., Zhang, X., Ren, S., Sun, J., 2015. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778.

Høye, T.T., Ärje, J., Bjerge, K., Hansen, O.L., Iosifidis, A., Leese, F., Mann, H.M.R., Meissner, K., Melyad, C., Raitoharju, J., 2021. Deep learning and computer vision will transform entomology. PNAS 118 (2), 4838. https://doi.org/10.1073/pnas.2002545117.

Júnior, T.D.C., Rieder, R., Di Domênico, J.R., Lau, D., 2022. InsectCV: a system for insect detection in the lab from trap images. Eco. Inform. 67, 101516 https://doi.org/10.1016/j.ecoinf.2021.101516.

Kalfas, I., De Ketelaere, B., Bunkens, K., Saeys, W., 2023. Towards automatic insect monitoring on witloof chicory fields using sticky plate image analysis. Eco. Inform. 75, 102037 https://doi.org/10.1016/j.ecoinf.2023.102037.

Kittichai, V., Pengsakul, T., Chumchuen, K., Samung, Y., Sriwichai, P., Phatthamolrat, N., Tongloy, T., Jaksukam, K., Chuwongin, S., Boonsang, S., 2021. Deep learning approaches for challenging species and gender identification of mosquito vectors. Sci. Rep. 11 (1), 4838. https://doi.org/10.1038/s41598-021-84219-4.

Liu, C., Tao, Y., Liang, J., Li, K., Chen, Y., 2018. Object detection based on YOLO network. In: 2018 IEEE 4th Information Technology and Mechatronics Engineering Conference (ITOEC). IEEE, pp. 799–803.

Marques, A.C.R., Raimundo, M., Cavalheiro, E.M., Salles, F.P., Lyra, C., Von Zuben, F., 2018. Ant genera identification using an ensemble of convolutional neural networks. PLoS One 13 (1), e0192011. https://doi.org/10.1371/journal.pone.0192011.

Michener, W.K., 2015. Ecological data sharing. Eco. Inform. 29, 33–44. https://doi.org/10.1016/j.ecoinf.2015.06.010.

Nayak, M.K., Daglish, G.J., Phillips, T.W., Ebert, P.R., 2019. Resistance to the fumigant phosphine and its management in insect pests of stored products: a global perspective. Annu. Rev. Entomol. 65, 333–350. https://doi.org/10.1146/annurev-ento-011019.

Preti, M., Verheggen, F., Angeli, S., 2021. Insect pest monitoring with camera-equipped traps: strengths and limitations. J. Pest. Sci. 94 (2), 203–217. https://doi.org/10.1007/s10340-020-01309-4.

Qiao, M., Lim, J., Ji, C.W., Chung, B.K., Kim, H.Y., Uhm, K.B., Chon, T.S., 2008. Density estimation of *Bemisia tabaci* (Hemiptera: Aleyrodidae) in a greenhouse using sticky traps in conjunction with an image processing system. J. Asia Pac. Entomol. 11 (1), 25–29. https://doi.org/10.1016/j.aspen.2008.03.002.

Ren, S., He, K., Girshick, R., Sun, J., 2015. Faster r-cnn: towards real-time object detection with region proposal networks. Adv. Neural Inf. Proces. Syst. 28, 91–99.

Rosado, L., Faria, P., Gonçalves, J., Silva, E., Vasconcelos, A., Braga, C., Oliveira, J., Gomes, R., Barbosa, T., Ribeiro, D., Nogueira, T., Ferreira, A., Carlos, C., 2022. EyesOnTraps: AI-powered mobile-based solution for pest monitoring in viticulture. Sustainability 14 (15), 9729. https://doi.org/10.3390/su14159729.

Rosenheim, J.A., Gratton, C., 2017. Ecoinformatics (big data) for agricultural entomology: pitfalls, progress, and promise. Annu. Rev. Entomol. 62, 399–417. https://doi.org/10.1146/annurev-ento-031616-035444.

Rustia, D.J.A., Lin, C.E., Chung, J.Y., Zhuang, Y.J., Hsu, J.C., Lin, T.T., 2020. Application of an image and environmental sensor network for automated greenhouse insect pest monitoring. J. Asia Pac. Entomol. 23 (1), 17–28. https://doi.org/10.1016/j.aspen.2019.11.006.

Rustia, D.J.A., Chao, J.J., Chiu, L.Y., Wu, Y.F., Chung, J.Y., Hsu, J.C., Lin, T.T., 2021. Automatic greenhouse insect pest detection and recognition based on a cascaded deep learning classification method. J. Appl. Entomol. 145 (3), 206–222. https://doi.org/10.1111/jen.12834.

Rustia, D.J.A., Chiu, L.Y., Lu, C.Y., Wu, Y.F., Chen, S.K., Chung, J.Y., Lin, T.T., 2022. Towards intelligent and integrated pest management through an AIoT-based monitoring system. Pest Manag. Sci. 78 (10), 4288–4302. https://doi.org/10.1002/ps.7048.

Salamut, C., Kohnert, I., Landwehr, N., Pflanz, M., Schirrmann, M., Zare, M., 2023. Deep learning object detection for image analysis of cherry fruit fly (*Rhagoletis cerasi L.*) on yellow sticky traps. Gesunde Pflanzen 75, 37–48. https://doi.org/10.1007/s10343-022-00794-0.

Sánchez-Bayo, F., Wyckhuys, K.A.G., 2019. Worldwide decline of the entomofauna: a review of its drivers. Biol. Conserv. 232, 8–27. https://doi.org/10.1016/j.biocon.2019.01.020.

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 Ecol. Evol. 13 (2), 346–357. https://doi.org/10.1111/2041-210X.13769.

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. Ecol. Lett. 26, 1247–1258. https://doi.org/10.1111/ele.14239.

Teixeira, A.C., Ribeiro, J., Morais, R., Sousa, J.J., Cunha, A., 2023. A systematic review on automatic insect detection using deep learning. Agriculture 13, 713. https://doi.org/10.3390/agriculture13030713.

Wang, Q.J., Zhang, S.Y., Dong, S.F., Zhang, G.C., Yang, J., Li, R., Wang, H.Q., 2020. Pest24: a large-scale very small object data set of agricultural pests for multi-target detection. Comput. Electron. Agric. 175, 105585 https://doi.org/10.1016/j.compag.2020.105585.

Wei, M., Zhan, W., 2024. YOLO_MRC: a fast and lightweight model for real-time detection and individual counting of Tephritidae pests. Eco. Inform. 79, 102445 https://doi.org/10.1016/j.ecoinf.2023.102445.

Wu, Y., Kirillov, A., Massa, F., Lo, W.Y., Girshick, R., 2019. Detectron2 [www document]. URL. https://github.com/facebookresearch/detectron2.

Xia, C., Chon, T.S., Ren, Z., Lee, J.M., 2015. Automatic identification and counting of small size pests in greenhouse conditions with low computational cost. Eco. Inform. 29, 139–146. https://doi.org/10.1016/j.ecoinf.2014.09.006.