Zero-Shot Insect Detection via Weak Language Supervision

Benjamin Feuer,[1] Ameya Joshi,[1] Minsu Cho,[1] Kewal Jani,[1] Shivani Chiranjeevi, [2] Zi Kang Deng, [3] Aditya Balu, [2] Asheesh K. Singh, [2] Soumik Sarkar, [2] Nirav Merchant, [3] Arti Singh, [2] Baskar Ganapathysubramanian, [2] Chinmay Hegde [1]

[1] New York University
[2] Iowa State University
[3] University of Arizona

Abstract

Open source image datasets collected via citizen science platforms (such as iNaturalist) can pave the way for the development of powerful AI models for insect detection and classification. However, traditional supervised learning methods require labeled data, and manual annotation of these raw datasets with useful labels (such as bounding boxes) can be extremely laborious, expensive, and error-prone. In this paper, we show that recent advances in vision-language models enable highly accurate zero-shot detection of insects in a variety of challenging environs. Our contributions are twofold: a) We curate the Insecta rank class of iNaturalist to form a new benchmark dataset of approximately 6M images consisting of 2526 agriculturally important species (both pests and beneficial insects). b) Using a vision-language object detection method coupled with weak language supervision, we are able to automatically annotate images in this dataset with bounding box information localizing the insect within each image. Our method succeeds in detection of diverse insect species present in a wide variety of backgrounds, producing high-quality bounding boxes in a zero-shot manner with no additional training cost.

Introduction

Insect pests in the agricultural sector cause infestation and damages to crops resulting in significant economic losses. Improper identification of species (as well as their number density, called the action threshold) could potentially result in unnecessary application of chemicals that could harm beneficial insects, reduce profitability, and have an adverse environmental footprint. While manual scouting remains the gold standard for pest identification and action threshold determination, this is a resource and (expert)labor intensive, yet critical aspect of agriculture. There is significant opportunity for computer vision and AI/ML approaches to contribute to automating this process. However, the task of identification and localization of insects is very challenging due to (a) the large number of species, (b) several distinct species that exhibit remarkably similar visual features, (c) species exhibiting very diverse features along their developmental cycle (nymph vs adult, larva vs pupa vs adult), and (d) images where the insect is difficult to differentiate from the background (for example, green colored insect-pests on green backgrounds).

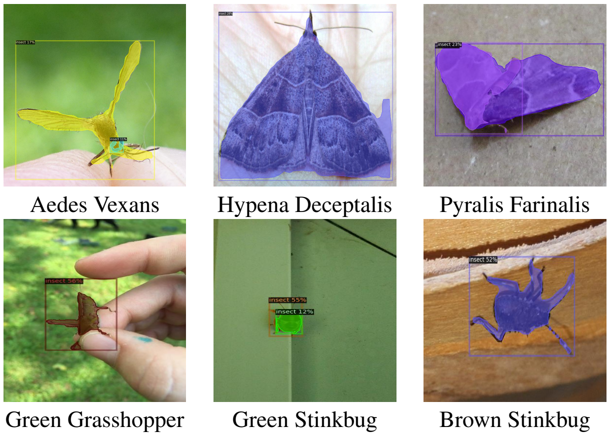

Aedes Vexans
Hypena Deceptalis
Pyralis Farinalis
Green Grasshopper
Green Stinkbug
Brown Stinkbug

Figure 1: Example bounding boxes and segmentation maps with zero-shot DETIC.

The availability of massive open source image datasets (such as iNaturalist (Van Horn et al. 2018)) acquired in a crowd-sourced manner can be leveraged to build powerful deep neural network models that perform accurate insect classification and detection. However traditional deep neural network-based object detectors require high quality annotations (or labels) for all image sample in these datasets. Annotating labels for object detection involves either pixel-by-pixel labelling of class information, or marking of tight bounding boxes for every image in a dataset. Consequently, creating datasets for AI models for insect detection in a supervised manner can be very laborious, time consuming, and prone to errors.

To overcome this, we leverage recent advances in *vision-language modeling*. Starting from CLIP (Radford et al. 2021), the key idea has been to pre-train deep networks that learn to match image data (possibly gathered in an unstructured manner) to associated captions that describe the contents of the images using natural language descriptions. The resulting models are *remarkably robust*: CLIP-style models produce representations of images that transfer very well to a variety of downstream tasks. For many applications these models also enable *zero-shot* performance, i.e., no extra supervision involving additional training data or computation is required. Rather, inference can be performed by simply specifying the category/class using a new natural language description. In our application, we show that merely coupling a recently proposed vision-language object detector (Zhou et al. 2022) with a *single (universal) natural language prompt* provides highly accurate bounding box for a very large dataset of diverse insect-pest images.

In summary, our contributions in this paper are twofold:

1. We curate the Insecta rank class of iNaturalist to form a new benchmark dataset of approximately 6M images consisting of 2526 agriculturally important species. We perform manual quality checking of a subset of these images.
2. Using a vision-language object detection method coupled with weak language supervision, we are able to automatically annotate images in this dataset with bounding box information localizing insect-pests in each image.

Our method succeeds in detection of diverse insect-pests present in a wide variety of backgrounds. In the absence of ground truth, we performed manual quality checks; over a carefully selected subset of images with diverse insect-pest categories, our method exhibited tight bound bounding boxes in large fraction of the samples; therefore, we expect that our new benchmark dataset can be used in the future for building high-quality supervised models as well.

Background: Zero-Shot Detection

Detection models focus on two loosely correlated problems: localizing objects of interest in an image, and assigning labels to them. One popular approach is a two-stage process (Ren et al. 2015; Girshick 2015; He et al. 2020; Lin et al. 2020) wherein the models detect probable object region proposals, and further finetune the bounding boxes and predict classes. In contrast, a single shot detector (Redmon and Farhadi 2018, 2017) not only generates region proposals but also classifies them in a single forward pass. Both of these however rely on high quality, fine grained annotations of localized objects in an image for training. Recent work on weak supervision for detection (Fang et al. 2021; Xu et al. 2021) attempt to resolve the need for such fine-grained labelling by assigning labels to boxes based on model predictions. For example, YOLO9000 (Redmon and Farhadi 2017) assigns labels to boxes based on the magnitude of prediction scores from the classification head. This however requires good quality box proposals *apriori* which may be hard to achieve essentially leading to a circular problem of needing good boxes for good class predictions and vice-versa.

Detic (Zhou et al. 2022) presents an interesting zero-shot solution to this problem by training detection models simultaneously with object detection and image classification datasets. Formally, let $\mathcal{D}_{det} = \{\mathbf{x}_i, \{b_{i,j}, c_{i,j}\}\}$ consist of images with labelled boxes, and $\mathcal{D}_{cls} = \{\mathbf{x}_i, c_i\}$ be a classification dataset with image-level labels. Traditional detection networks consist of a two-stage detector; the first half of the network, $f_D : \mathbb{R}^d \rightarrow \{\mathbb{R}^m \times [0,1]\}$ outputs a set of bounding boxes and corresponding *objectness* scores. The second half, $f_c : \mathbb{R}^m \rightarrow \mathbb{R}^4 \times [c]$ takes in every proposal with an objectness score higher than a threshold and outputs a bounding box with the corresponding prediction. The networks are trained only on D_{det}.

Detic improves upon this by training f_c on both $\mathcal{D}_{det}$ and $\mathcal{D}_{cls}$. The classification head in f_c is also replaced with CLIP (Radford et al. 2021) embeddings as weights to add open-set classification capabilities. Every minibatch consists of mix of samples from $\mathcal{D}_{det}$ and $\mathcal{D}_{cls}$. The training examples from $\mathcal{D}_{det}$ are trained using the standard detection loss (boxwise regression and classification losses). Examples from $\mathcal{D}_{cls}$ are assumed to have a single detected object (the largest detected box) with the image label as the box label. The model is then trained with the following loss:

$$L(I) = \begin{cases} L_{RPN} + L_{Reg} + L_{cls}, & \text{if } I \in \mathcal{D}_{det} \\ \lambda L_{max-size}, & \text{if } I \in \mathcal{D}_{cls} \end{cases}$$

Note that here, L_{RPN}, L_{reg}, and L_{cls} refer to the training losses from (Ren et al. 2015) while $L_{max-size}$ is a cross-entropy loss with the target as the image class. DETIC has two advantages over traditional detectors; (1) it can learn from image classification datasets which are generally larger than detection datasets, and contain a variety of classes, and, (2) the CLIP embeddings used as the classification head allow for a far larger number of classes. Thus, contrary to standard detection models, DETIC does not require fine-tuning, and can be used for zero-shot detection with natural images.

A New Benchmark Dataset

iNaturalist is a citizen science platform where users can upload photographs of specific organisms. The iNaturalist Open Data project is a curated subset of the overall iNaturalist dataset that specifically contains images that apply to the Creative Commons license created by iNaturalist specifically to aid academic research.

We created a workflow tool, iNaturalist Open Download, to easily download species images from the iNaturalist Open Dataset associated with a specific taxonomy rank. We used the tool to download all images of species under the rank class Insecta from the iNaturalist Open Dataset for downstream annotation, curation and use in our model. We choose to only use images identified as "research" quality grade under the iNaturalist framework, which indicates that the labeling inspection for the image is more rigorous than standard and has multiple agreeing identifications at the species level. This results in a total of 13,271,072 images across 95,399 different insect species at the time of writing. The images have a maximum resolution of 1024x1024, in .jpg/.jpeg format and total 5.7 terabytes. Among the 95,399 insect species, we select 2526 species which have been reported to be the most agriculturally important species. This subset of insect classes contribute to 6 million images in total.

Experimental Results

We use the highest performing DETIC model from (Zhou et al. 2022) with the SWINB backbone. The model has been trained on the COCO and Imagenet-21k datasets, with Centernet employed as the region proposal network. We run the

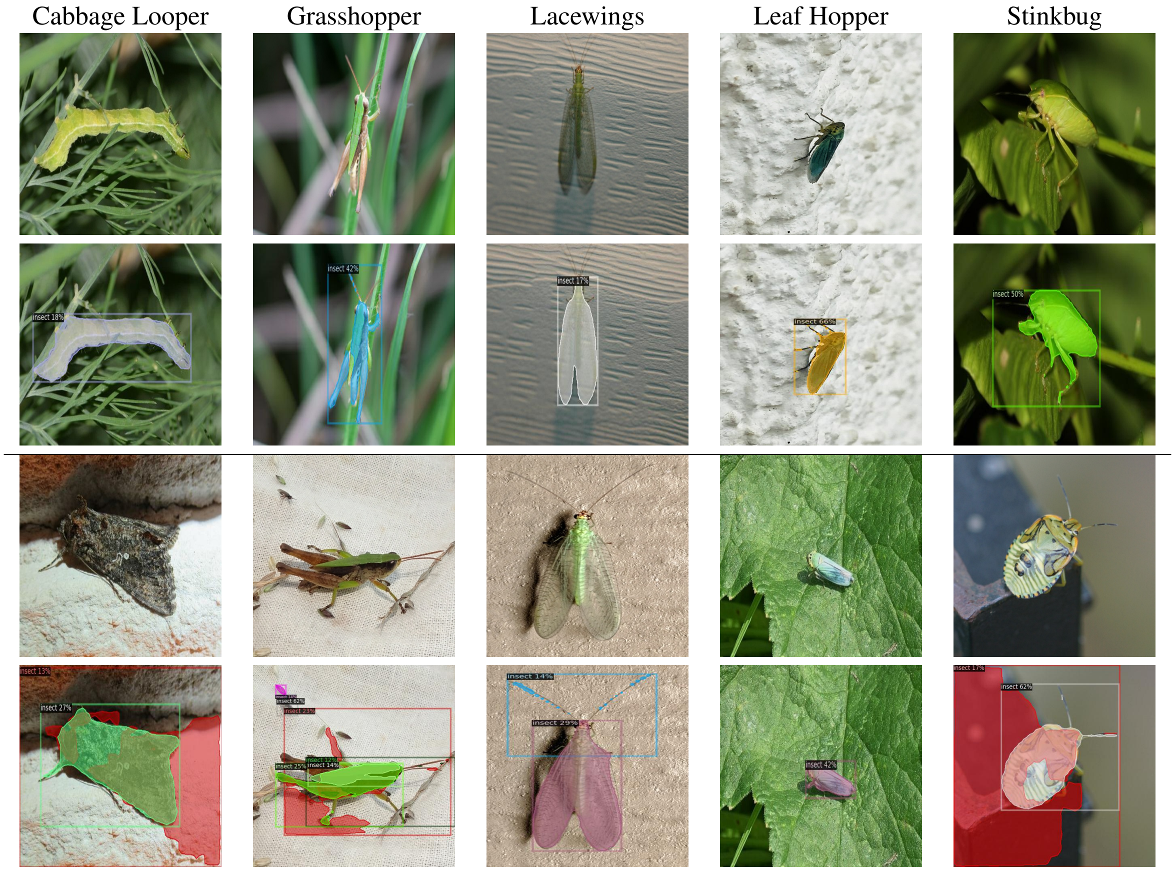


Figure 2: Detic results for green-colored insect-pests.

Species	Accuracy
22-Spotted LadyBird	87.5
Brown Marmorated Stink Bug	88.9
Cabbage White	55.6
Corn Earworm	100.0
Cotton Bollworm	100.0
Fall Armyworm	88.9
Golden Rod Soldier Beetle	83.4
Japanese Beetle	88.9
Red Milkweed Beetle	100.0
Silver-spotted Skipper	100.0
Tobacco Hornworm	70.0
Average	87.6

Table 1: Accuracy of detecting a diverse set of insect species.

Species	Accuracy
Cabbage Looper	88.9
Green Lacewing	66.7
Green Grasshopper	62.5
Green Leaf Hopper	100.0
Green Stink Bug	100.0
Average	83.6

Table 2: Accuracy of detecting insects on challenging backgrounds, specifically green insect-on-green background.

detector on images in our curated benchmark dataset. Our vocabulary only contains the lone word ”insect”, and therefore is an example of (very) weak language supervision. In order to ensure high recall, we use low objectness confidence thresholds of 0.1 and 0.3.

Our results show the considerable promise of zero-shot localization on this dataset. Figure 1 (and Figure 5 in appendix) show example results; in all cases, we see that the insect in the image has been localized with a tight bounding box as well as a high-quality pixel-wise semantic segmentation map.

Table 1 shows the results of our manual quality check over a list of nearly 150 image samples taken from a diverse set of insect species. This diverse dataset spans species across the phylogenic tree of the insecta order (for instance, winged vs wingless, ect). Domain experts were asked to label an output as “correct” if a tight bounding box was achieved, and quality labels were tabulated. We see that performance is high, with a quality metric of nearly 88%. Figure 3 shows example images from this test set.

We also find that our method succeeds in challenging

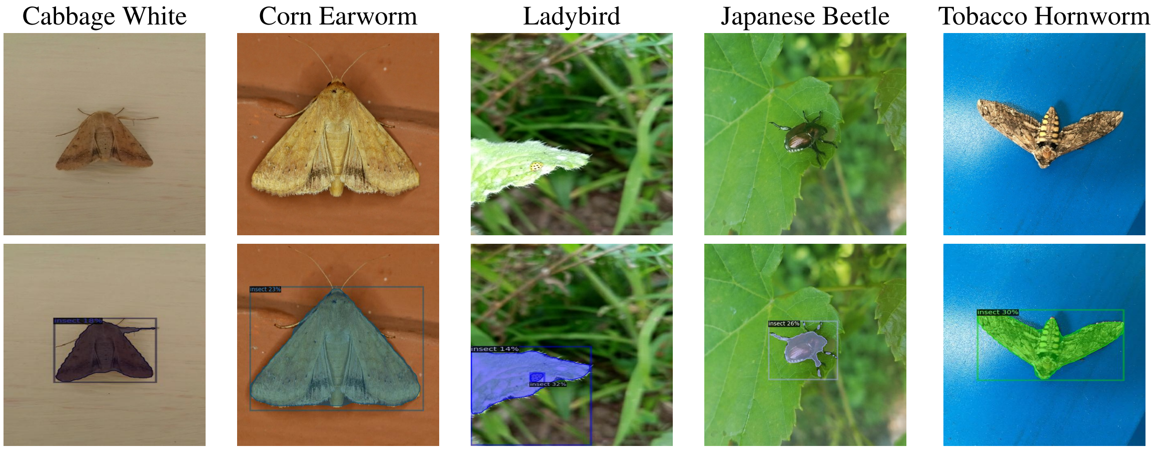


Figure 3: Detic results on a diverse set of insect species.

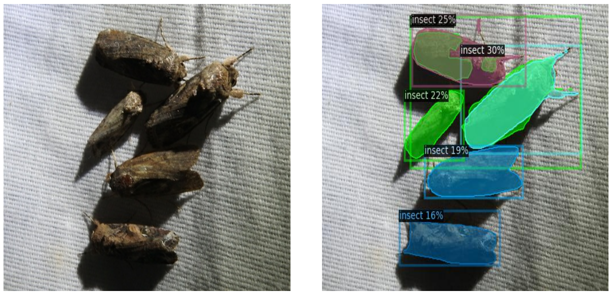


Figure 4: Detection of multiple insects in the same image.

scenarios, such as green-species captured on a green background, which is a fairly common occurrence. Figure 2 shows example results from a subset of such images. Here also, our method succeeds with high accuracy (over 83%); the green grasshopper (with its natural camouflaging property) was the most challenging to detect, as expected. See Table 2 for details.

Finally, we find that our method performs well even when there are multiple insects (of the same species) in the same image. See Figure. 4 that illustrates correct detection and bounding box construction on a major pest (fallarmy worm). Creating correct bounding boxes on multiple insects in the same image is critical to accurately evaluating the action threshold in integrated pest management.

Conclusions

We show that new advances in vision-language models can be used to effectively localize insects in unstructured image data (in a fully zero-shot manner, without requiring any training).

Our preliminary findings pave the way for further improvements in this area. First, our generated bounding box information can be useful in other downstream tasks in insect/pest monitoring, such as visual odometry and decision support. Second, better bounding boxes can be achieved perhaps with improved language supervision and class-specific prompts. Finally, we expect the curated, quality controlled dataset to be of significant interest to the CV community.

References

Fang, S.; Cao, Y.; Wang, X.; Chen, K.; Lin, D.; and Zhang, W. 2021. Wssod: A new pipeline for weakly- and semi-supervised object detection. *arXiv preprint, arXiv:2105.11293*.

Girshick, R. B. 2015. Fast R-CNN. *ICCV*, 1440–1448.

He, K.; Gkioxari, G.; Dollár, P.; and Girshick, R. B. 2020. Mask R-CNN. *T-PAMI*, 42: 386–397.

Lin, T.-Y.; Goyal, P.; Girshick, R. B.; He, K.; and Dollár, P. 2020. Focal Loss for Dense Object Detection. *T-PAMI*, 42: 318–327.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 8748–8763. PMLR.

Redmon, J.; and Farhadi, A. 2017. YOLO9000: Better, Faster, Stronger. In *CVPR*, 6517–6525.

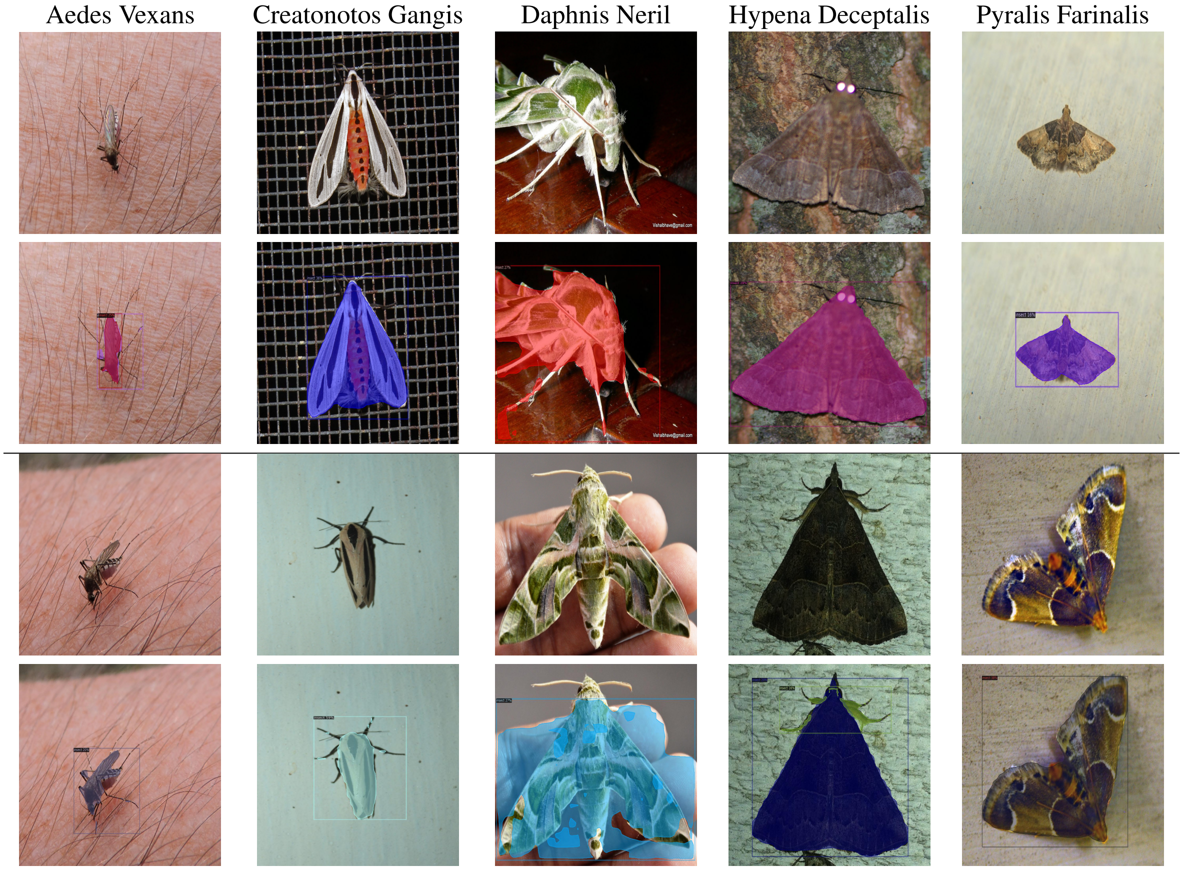
Redmon, J.; and Farhadi, A. 2018. YOLOv3: An Incremental Improvement. *ArXiv*, abs/1804.02767.

Ren, S.; He, K.; Girshick, R. B.; and Sun, J. 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *T-PAMI*, 39: 1137–1149.

Van Horn, G.; Mac Aodha, O.; Song, Y.; Cui, Y.; Sun, C.; Shepard, A.; Adam, H.; Perona, P.; and Belongie, S. 2018. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 8769–8778.

Xu, M.; Zhang, Z.; Hu, H.; Wang, J.; Wang, L.; Wei, F.; Bai, X.; and Liu, Z. 2021. End-to-end semi-supervised object detection with soft teacher. In *ICCV*.

Zhou, X.; Girdhar, R.; Joulin, A.; Krähenbühl, P.; and Misra, I. 2022. Detecting twenty-thousand classes using image-level supervision. *arXiv preprint arXiv:2201.02605*.



Aedes Vexans

Creatonotos Gangis

Daphnis Neril

Hypena Deceptalis

Pyralis Farinalis

Figure 5: Zero-shot bounding box generation and segmentation map obtained using DETIC.