BeetleFlow: An Integrative Deep Learning Pipeline for Beetle Image Processing

Fangxun Liu[1*], **S M Rayeed**[2*],
Samuel Stevens[1], Alyson East[3], Cheng Hsuan Chiang[1], Colin Lee[1], Daniel Yi[1], Junke Yang[1], Tejas Naik[1], Ziyi Wang[1], Connor Kilrain[1], Elijah H. Buckwalter[1], Jiacheng Hou[1], Saul Ibaven Bueno[1], Shuheng Wang[1], Xinyue Ma[1], Yifan Liu[1], Zhiyuan Tao[1], Ziheng Zhang[1], Eric Sokol[4], Michael Belitz[5], Sydne Record[3], Charles V. Stewart[2], Wei-Lun Chao[1]

[1]The Ohio State University [2]Rensselaer Polytechnic Institute [3]The University of Maine
[4]National Ecological Observatory Network (NEON), Battelle [5]Michigan State University

*Corresponding authors: liu.12122@osu.edu, rayees@rpi.edu

Abstract

In entomology and ecology research, biologists often need to collect a large number of insects, among which beetles are the most common species. A common practice for biologists to organize beetles is to place them on trays and take a picture of each tray. Given the images of thousands of such trays, it is important to have an automated pipeline to process the large-scale data for further research. Therefore, we develop a 3-stage pipeline to detect all the beetles on each tray, sort and crop the image of each beetle, and do morphological segmentation on the cropped beetles. For detection, we design an iterative process utilizing a transformer-based open-vocabulary object detector and a vision-language model. For segmentation, we manually labeled 670 beetle images and fine-tuned two variants of a transformer-based segmentation model to achieve fine-grained segmentation of beetles with relatively high accuracy. The pipeline integrates multiple deep learning methods and is specialized for beetle image processing, which can greatly improve the efficiency to process large-scale beetle data and accelerate biological research.

1 Introduction

In entomology and ecology research, biologists often need to collect a large number of insects to study, among which beetles are one of the most common species. Beetles account for around 25% of all known species in the world [1], therefore, they have significant research value in a variety of fields such as evolution and biodiversity, for their species richness and widespread distribution. In practice, biologists often use pins to mount beetles collected at the same site and time on a tray. A tray contains from several to over 60 beetle specimens. After collection, biologists can have up to thousands of trays and take a picture of each tray to digitize. This procedure brings in a large amount of beetle data organized by trays, but how to process them for further study becomes a new question.

Given the importance of beetle study and the challenges, we develop a 3-stage deep learning pipeline to process large-scale beetle images. In the first stage, we utilize an open-vocabulary detector, Grounding DINO [2], for beetle detection and a vision-language model, LLaVA-NeXT [3], for final verification. This approach achieves a high accuracy of 97.81%. In the second stage, we crop each detected beetle from the tray image and save it as a single image, with optional sorting and metadata matching. In the third stage, we leverage a transformer-based model, Mask2Former [4], to segment each beetle into 5 or 9 morphological parts. The model achieves a mean Intersection over Union (mIOU) of 85.11% for 5 class segmentation and 77.38% for 9 class segmentation.

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

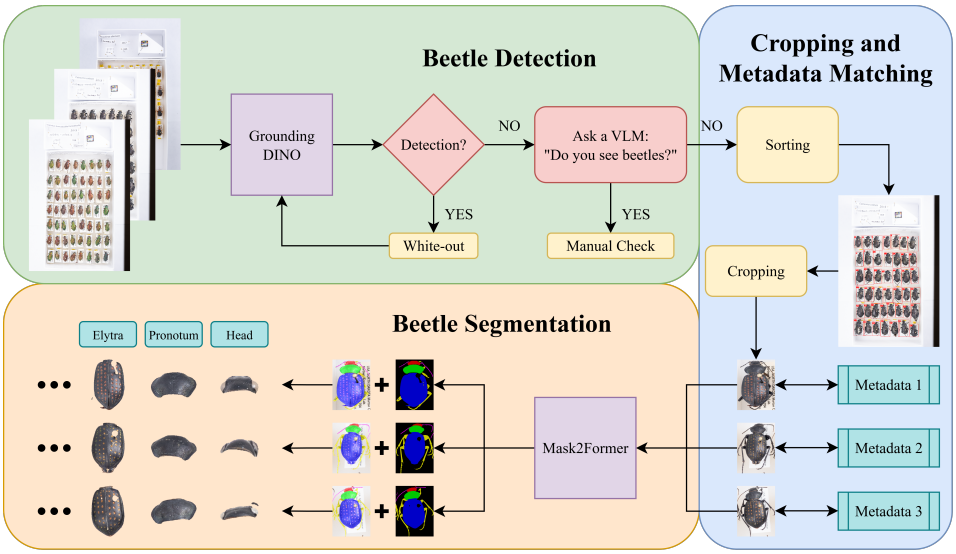

Figure 1: An overview of the 3-stage pipeline: individual detection, cropping/metadata matching, and body-part segmentation stages.

Multiple downstream research on beetles can be conducted based on the high-throughput data processed by our pipeline. Moreover, the pipeline has the potential to generalize to more biological data processing cases, as we have observed a similar detect-and-segment pattern in other biological pipelines, such as QuPath [5] for cells and PlantCV [6] for plants. This pattern is applicable to a variety of biological data processing workflows, and our work on beetles sets an example for insects, which are one of the most numerous groups of organisms in the world.

2 Related Work

While traditional object detectors like the R-CNN family [7] and YOLO [8] are limited by predefined datasets, open-vocabulary detection methods [9, 10] now leverage language prompts to detect arbitrary objects. Similarly, Vision-Language Models (VLMs) [11, 12] extend the reasoning of large language models [13–15] to the multimodal domain for joint text-image reasoning. In parallel, the state-of-the-art in semantic segmentation has shifted from Convolutional Neural Networks (CNNs) like U-Net [16] to transformer-based models [17–19]. These models leverage self-attention for global context, which is critical for distinguishing morphologically similar parts in biological imaging. These advancements are relevant to beetle studies, where machine learning, supported by large-scale beetle sampling programs like NEON [20], is already applied for identification and classification. Existing works range from traditional algorithms on extracted features [21] to deep learning methods, such as CNNs for identification [22] and deep vision models for fine-grained taxonomic classification [23, 24].

3 Beetle Image Processing Pipeline

The input to our pipeline is a series of images of trays containing multiple beetles. Each tray image undergoes a 3-stage processing (Figure 1). The code and data are available at `https://github.com/Imageomics/BeetleFlow.git`.

3.1 Iterative Beetle Detection

The first stage is iterative beetle detection. In one iteration, the tray image and the text prompt “a beetle” are sent as input to Grounding DINO, which then outputs the bounding box coordinates for the detected beetles. Next, white masks are placed over all detected beetles based on the bounding boxes, leaving the undetected ones in the tray image. The resulting modified image then proceeds to the next iteration for another round of detection and masking. When no detection is reported, the

iteration stops. The final modified image is then sent to LLaVA-NeXT with the text prompt "Do you see beetles in this image?". We constrain the model to output "YES" or "NO" as the final word for automated check. If it answers "YES", a message is sent to the user to check manually. If it answers "NO", the detection is successful and the bounding box coordinates are sent to the next stage.

3.2 Beetle Image Cropping

The second stage takes the bounding box coordinates as input and outputs individual cropped beetle images, with optional functions of sorting and metadata matching. Given the bounding box coordinates, the pipeline crops the beetles out of the original tray image and saves them as individual images. A specific order can be applied when saving the beetle images. In practice, the beetles on a tray are arranged in regular rows and columns. If digitized metadata for the beetles are available, they are typically provided in a left-to-right, top-to-bottom order by the biologists. The pipeline sorts the beetle images according to the top-left bounding box coordinate and associates the metadata with each beetle. The metadata matched to the beetles are saved in a CSV file for each tray.

3.3 Fine-grained Beetle Segmentation

The third stage takes individual beetle images as input and outputs the morphological segmentation results for each beetle. For this task, we fine-tune two variants of Mask2Former, based on granularity: a 5-class model (segmenting head, pronotum, elytra, legs, and antennas) for basic morphological analysis, and a 9-class model (additionally eyes, mouthparts, tail, and pin). Both models inherently separate the beetle from the background. For each input image, the model generates a colorized mask image and an image overlaid with the masks. The overlaid image is provided for user verification. The mask image can be utilized for morphological part cropping and defective specimen detection.

4 Experiments

4.1 Experimental Setup

4.1.1 Beetle Detection

Datasets. We apply the detection pipeline on the dataset collected by the National Ecological Observatory Network (NEON) from ecological sites across the U.S., along with associated metadata. The dataset contains 1,506 images of trays containing pinned carabid specimens. The beetles were imaged following the optimized image capture guidelines for biodiversity specimens [25].

Evaluation metrics. Each tray in the NEON dataset is associated with respective metadata, including the ground-truth number of beetles on the tray. After running the detection process, the number of detected beetles is compared to the ground-truth number. The exact-match accuracy over the 1,506 tray images is then calculated to quantify the model's performance.

Implementation Details. We utilized the Grounding DINO model with pre-trained weights from the `IDEA-Research/grounding-dino-base` checkpoint. A fixed text prompt, "a beetle.", was used to guide the model in locating specimens. For post-processing, we set the box confidence threshold to 0.3 and the text relevance threshold to 0.2. Only bounding boxes with scores exceeding both respective thresholds are retained. We utilized the LLaVA-NeXT model with pre-trained weights from the `llava-hf/llava-v1.6-mistral-7b-hf` checkpoint for the final verification.

4.1.2 Beetle Segmentation

Datasets. The unlabeled individual beetle images are derived from our pipeline by processing the tray images. For 5-class labeling, we manually labeled 160 beetles and utilized an additional 180 labeled beetles from a previous work, SST [26]. A total of 340 labeled beetles were then partitioned into a training set of 272 and a test set of 68 images. For 9-class labeling, we manually labeled 330 beetles, dividing them into a training set of 264 and a test set of 66 images.

Evaluation metrics. We use mIOU on the test set as the primary metric, averaged across all classes for each image. We also report some per-class IoU scores to facilitate a more granular analysis.

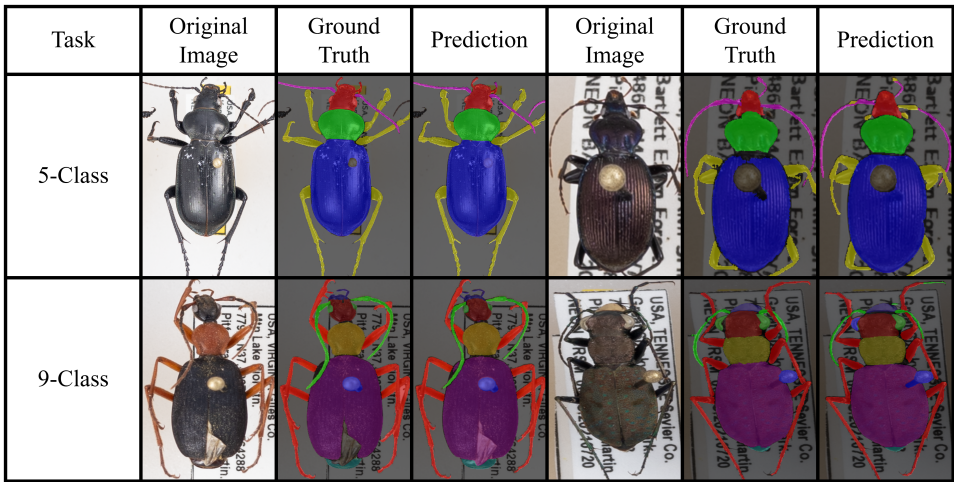

Figure 2: Qualitative segmentation results. Top row: 5-class segmentation (head, pronotum, elytra, legs, antennas). Bottom row: 9-class segmentation (adding eyes, mouthparts, tail, pin). Columns show original, ground truth, and prediction.

Implementation Details. We fine-tuned two Mask2Former models with a Swin-Large backbone, each initialized with weights from the `facebook/mask2former-swin-large-ade-semantic` checkpoint, which was pre-trained on the ADE20K dataset for semantic segmentation tasks. All input images were resized to a resolution of 512×512 pixels. The models were trained for 30 epochs with a batch size of 10, using the AdamW optimizer and an initial learning rate of 1e-4.

4.2 Results

Beetle Detection. We applied the detection process to 1,506 tray images. Of these, 1,473 trays had a perfect match between the number of detected beetles and the ground truth, yielding a total accuracy of 97.81%. Of the 33 failure cases, 32 had a higher detected beetle count than the ground truth, while only one case had a lower count. A majority of these 32 cases were due to fallen beetle heads on the trays, which the model incorrectly detected as separate beetles. The data indicates that our detection process is highly effective at detecting all beetles on a tray with minimal omissions.

Beetle Segmentation. We evaluated the performance of two models on their respective test sets. The mIOU is 85.11% for 5-class segmentation and 77.38% for 9-class segmentation. For per-class IOU results, the model achieves high scores on large morphological parts like "pronotum" (91.85% and 90.99%) and "elytra" (94.69% and 93.97%), while the scores for smaller parts like "legs" (79.57% and 85.39%) and "antennas" (65.93% and 70.08%) are lower. This phenomenon is partly attributable to the sensitivity of IoU to object size. For small objects, deviations of a few pixels can lead to a significant drop. Qualitative results (Figure 2) show that the segmentation is reasonably good.

5 Discussion and Future Work

In this work, we develop an automated pipeline to process large-scale beetle images. The two deep learning-based processes, Grounding DINO detection and Mask2Former segmentation, have proven to be highly accurate. In addition, each tray also contains a scale bar and a color table, and we have developed the functionality to detect and crop them. The scale bar can be used to automatically measure the morphological statistics like length and area, and the color table can be used for the color calibration of images. These applications can be added in future work. In summary, our pipeline greatly improves the efficiency of large-scale beetle image processing, yielding useful outputs for downstream research. This pipeline scheme can be further generalized to other organisms beyond beetles, with the potential to improve the data processing workflow in the biology field.

References

[1] P. M. Hammond, "Species inventory," in *Global Biodiversity: Status of the Earth's Living Resources. A Report Compiled by the World Conservation Monitoring Centre*, B. Groombridge, Ed. London: Chapman and Hall, 1992, pp. 17–39.

[2] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, Q. Jiang, C. Li, J. Yang, H. Su *et al.*, "Grounding dino: Marrying dino with grounded pre-training for open-set object detection," in *European conference on computer vision*. Springer, 2024, pp. 38–55.

[3] H. Liu, C. Li, Y. Li, B. Li, Y. Zhang, S. Shen, and Y. J. Lee, "Llava-next: Improved reasoning, ocr, and world knowledge," January 2024. [Online]. Available: https://llava-vl.github.io/blog/2024-01-30-llava-next/

[4] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, "Masked-attention mask transformer for universal image segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 1290–1299.

[5] P. Bankhead, M. B. Loughrey, J. A. Fernández, Y. Dombrowski, D. G. McArt, P. D. Dunne, S. McQuaid, R. T. Gray, L. J. Murray, H. G. Coleman *et al.*, "Qupath: Open source software for digital pathology image analysis," *Scientific reports*, vol. 7, no. 1, pp. 1–7, 2017.

[6] M. A. Gehan, N. Fahlgren, A. Abbasi, J. C. Berry, S. T. Callen, L. Chavez, A. N. Doust, M. J. Feldman, K. B. Gilbert, J. G. Hodge *et al.*, "Plantcv v2: Image analysis software for high-throughput plant phenotyping," *PeerJ*, vol. 5, p. e4088, 2017.

[7] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580–587.

[8] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.

[9] L. H. Li, P. Zhang, H. Zhang, J. Yang, C. Li, Y. Zhong, L. Wang, L. Yuan, L. Zhang, J.-N. Hwang *et al.*, "Grounded language-image pre-training," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 965–10 975.

[10] X. Gu, T.-Y. Lin, W. Kuo, and Y. Cui, "Open-vocabulary object detection via vision and language knowledge distillation," *arXiv preprint arXiv:2104.13921*, 2021.

[11] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PmLR, 2021, pp. 8748–8763.

[12] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *International conference on machine learning*. PMLR, 2023, pp. 19 730–19 742.

[13] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[14] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan *et al.*, "The llama 3 herd of models," *arXiv preprint arXiv:2407.21783*, 2024.

[15] G. Team, P. Georgiev, V. I. Lei, R. Burnell, L. Bai, A. Gulati, G. Tanzer, D. Vincent, Z. Pan, S. Wang *et al.*, "Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context," *arXiv preprint arXiv:2403.05530*, 2024.

[16] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[18] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[19] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.

[20] D. Hoekman, K. E. LeVan, C. Gibson, G. E. Ball, R. A. Browne, R. L. Davidson, T. L. Erwin, C. B. Knisley, J. R. LaBonte, J. Lundgren *et al.*, "Design for ground beetle abundance and diversity sampling within the national ecological observatory network," *Ecosphere*, vol. 8, no. 4, p. e01744, 2017.

[21] J. Blair, M. D. Weiser, M. Kaspari, M. Miller, C. Siler, and K. E. Marshall, "Robust and simplified machine learning identification of pitfall trap-collected ground beetles at the continental scale," *Ecology and evolution*, vol. 10, no. 23, pp. 13 143–13 153, 2020.

[22] L. Wu, Z. Liu, T. Bera, H. Ding, D. A. Langley, A. Jenkins-Barnes, C. Furlanello, V. Maggio, W. Tong, and J. Xu, "A deep learning model to recognize food contaminating beetle species based on elytra fragments," *Computers and Electronics in Agriculture*, vol. 166, p. 105002, 2019.

[23] S. M. Rayeed, A. East, S. Stevens, S. Record, and C. V. Stewart, "Fine-grained beetle taxonomy with vision models: A benchmark on long-tailed and domain-adaptive classification," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, October 2025, pp. 5093–5099.

[24] S. Rayeed, A. East, S. Stevens, S. Record, and C. V. Stewart, "Beetleverse: A study on taxonomic classification of ground beetles," *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2025. [Online]. Available: https://arxiv.org/abs/2504.13393

[25] A. East, E. G. Campolongo, L. Meyers, S. Rayeed, S. Stevens, I. Zarubiieva, I. E. Fluck, J. C. Girón, M. Jousse, S. Lowe *et al.*, "Optimizing image capture for computer vision-powered taxonomic identification and trait recognition of biodiversity specimens," *Methods in Ecology and Evolution*, 2025.

[26] Z. Feng, Z. Wang, S. I. Bueno, T. Frelek, A. Ramesh, J. Bai, L. Wang, Z. Huang, J. Gu, J. Yoo, T.-Y. Pan, A. Chowdhury, M. Ramirez, E. G. Campolongo, M. J. Thompson, C. G. Lawrence, S. Record, N. Rosser, A. Karpatne, D. Rubenstein, H. Lapp, C. V. Stewart, T. Berger-Wolf, Y. Su, and W.-L. Chao, "Static segmentation by tracking: A frustratingly label-efficient approach to fine-grained segmentation," 2025. [Online]. Available: https://arxiv.org/abs/2501.06749