Imageomics: ML for Biological Knowledge Discovery

Elizabeth G. Campolongo

On behalf of the Imageomics Institute, The Ohio State University

Abstract

A broad goal of the Imageomics Institute is to inspire ML innovation while increasing biological knowledge extraction from images. In furtherance of this goal, we create large and diverse datasets, processing and data exploration tools, and models—big and small—to aid in biological discovery. In this poster we outline some of these datasets, open-source tools, and models to engage with the broader research community.

Processing Tools

LepidopteraLens — Coming Soon!

Pipeline incorporating preprocessing tools (Imageomics' & other open-source) used to analyze images of butterflies:

- Object detection and segmentation (based on YOLOv8),
- Color standardization (with and without colorchecker),
- Automated Landmarking (using ml-morph by A. Porto),
- Quantification of several phenotypic values,

Outputs: PCA of color and pattern variation using the recolorize and patternize workflow by H. Weller and S. Van Belleghem.

LatLonCover

- Generates distribution of landcover types within small and large neighborhoods of given decimal latitude and longitude.
- Uses CropScape land cover designation to determine percentage of each of 7 categories (forest, water, etc.) around given location.
- Available as command line tool, Python API, and HF instance.
- Developed at our Image Datapalooza 2023 workshop.

Classification

Interpretable Transformer

INTR: A Simple Interpretable Transformer for Fine-grained Image Classification and Analysis.

- Built on DETR-R50 backbone.
- Can generate visual representations of INTR's interpretations.
- Demo notebook available for inference time single-image prediction and visualization.

Balanced Butterfly Classification

- 65 classes with approximately 4600 training and 275 test images.
  - Subset of images from the Butterfly Genetics Group (field studies by Chris Jiggins' team at the University of Cambridge). Separated wings.
- Mix of *Heliconius* and other genera of butterflies.

Sample Images

Butterfly from tribe *Ithomiinae*
Pinheiro de Castro, E., Jiggins, C., Lucas da Silva-Brandão, K., Victor Lucci Freitas, A., Zikan Cardoso, M., Van Der Heijden, E., Meier, J., & Warren, I. (2022). Brazilian Butterflies Collected December 2020 to January 2021. Zenodo. https://doi.org/10.5281/zenodo.5561246

Heliconius doris
Jiggins, C., Montejo-Kovacevich, G., Warren, I., & Wiltshire, E. (2019). Cambridge butterfly wing collection batch 3. Zenodo. https://doi.org/10.5281/zenodo.2682458

Caliga eurilochus
Gabriela Montejo-Kovacevich, Letitia Cookson, Eva van der Heijden, Ian Warren, David P. Edwards, & Chris Jiggins. (2020). Cambridge butterfly collection - Loreto, Peru 2018 (1.0.0). Zenodo. https://doi.org/10.5281/zenodo.3569598

Downstream Tasks

- Train and testing data for INTR.
- Fine-grained image classification and difference localization.

Datasets for Animal Behavior Recognition

In-Situ Datasets for Kenyan Animal Behavior Recognition (KABR) from Drone Videos

- Collected using drones that flew over the animals in the Mpala Research Centre in Kenya, providing high-quality video footage of the animal's natural behaviors.
- Behaviors of giraffes, plains zebras, and Grevy's zebras.
- 14,764 unique behavioral sequences, eight different classes.
  - 1,139,893 total frames: 488,638 of Grevy's zebras, 492,507 frames of plains zebras, and 158,748 frames of giraffes.
  - Divided into 850,000 training and 290,000 validation frames.
  - Resolution: 5472 x 3078 pixels; frame rate: 29.97 frames/sec.
- Annotation team led by expert zoologist: behaviors labeled based on distinctive features, using standardized criteria to ensure consistency and accuracy.

Samples

KABR Telemetry Data

The telemetry and drone flight data associated with the KABR dataset. It contains information about the status drone during the missions, such as

- Location (decimal latitude and longitude), altitude, and camera settings (shutter speed, digital zoom, etc.).

Additional animal information is also provided, such as

- Bounding box dimensions of the wildlife in the frame, animal identification (Zebra or Giraffe), and behavior annotation information (Walk, Graze, Head Up etc.).

Downstream Tasks

- Behavior recognition, pose estimation.
- Robotics, development of autonomous navigation algorithms for wildlife data collection, contextualization for KABR dataset.

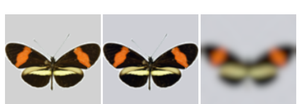
Butterfly Mimicry Recognition Datasets

Subspecies of *Heliconius erato* & *Heliconius melpomene*

- Dorsal full body images of 18 subspecies.
  - 320 specimens, with 3 images per specimen: Original, Bird transformed, and Butterfly transformed (on HF).
  - Low-resolution RGB photos from (Hoyal Cuthill et al., 2019), transformed using AcuityView.
- Dorsal separated-wings of 20 subspecies.
  - Higher-resolution RGB photos from the Butterfly Genetics Group, transformed with newer insect visual acuity estimates.

Sample Images

Heliconius erato ssp. cybria. Views from left to right: RGB, Bird Acuity, Butterfly Acuity



Heliconius melpomene ssp. rosina. Views from left to right: RGB, Bird Acuity, Butterfly Acuity

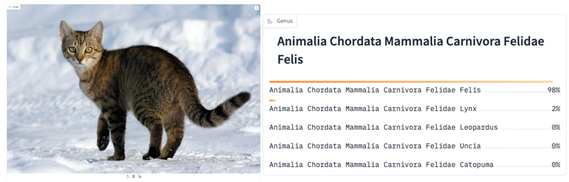
Downstream Tasks

Fine-grained image classification, image generation.

Biological Foundation Model

BioCLIP: A Vision Foundation Model for the Tree of Life

- Based on OpenAI's CLIP, trained on TreeOfLife-10M.
- BioCLIP consistently outperformed existing fine-grained biological classification baselines by 17% to 20% absolute.
- Demo available on Hugging Face (snapshot below).



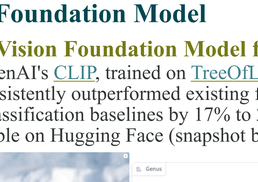
TreeOfLife-10M Dataset

- Largest-to-date ML-ready dataset of images of biological organisms paired with their taxonomic labels.
  - Expands on the foundation established by existing high-quality datasets (eg., iNat21 and BIOSCAN-1M), by further incorporating newly curated images from Encyclopedia of Life (EOL, eol.org, supplies most of the data diversity).
- Over 10 million images covering 454K taxa in the tree of life.
- Every image is labeled to the most specific taxonomic level possible, as well as higher taxonomic ranks in the tree of life (*Kingdom* through *species*).
- Subdivided into training and validation sets, with a 1M image subset of the training data (an ablation study set).

Sample Images

Top Left: *Animalia Chordata Mammalia Artiodactyla Bovidae Ovis aries* (common name urial). Image cc-by-nc-4.0, copyright christianschwarz.

Above: *Fungi Basidiomycota Agaricomycetes Agaricales Entolomataceae Entoloma canoconicum*. Image cc-by-nc-4.0, copyright Alice Shanks.



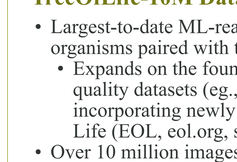
Left: *Digitaria eriantha steud.* Image dedicated to the public domain by ARS Systematic Botany and Mycology Laboratory.

Rare Species Benchmark

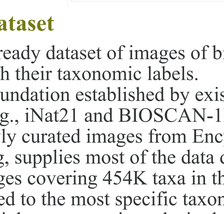
This dataset was generated from EOL alongside TreeOfLife-10M.

- "Rare species": listed on the IUCN Red List (iucnredlist.org) as Near Threatened, Vulnerable, Endangered, Critically Endangered, or Extinct in the Wild.
- Balanced 40 species, with 12K images, all in kingdom *Animalia*.

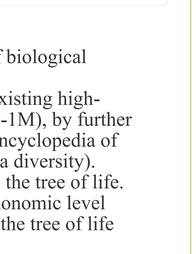
Sample Images



Above: *Animalia Echinodermata Holothuroidea Synallactida Stichopodidae Stichopus herrmanni*. Image cc-by-3.0, copyright Francois Michonneau.



Center: *Animalia Chordata Aves Charadriiformes Glareolidae Glareola nordmanni*. Image cc-by-3.0, copyright Derek Keats.



Right: *Animalia Chordata Aves Galliformes Phasianidae Lophura swinhoii*. Image cc-by-sa-3.0, copyright CharlesLam.

Downstream Tasks

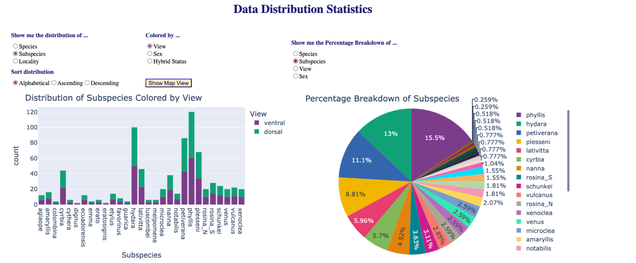
- Image classification, zero-shot classification.
- TreeOfLife-10M and Rare Species were used to train and benchmark BioCLIP, as they can be used for future biology foundation models.

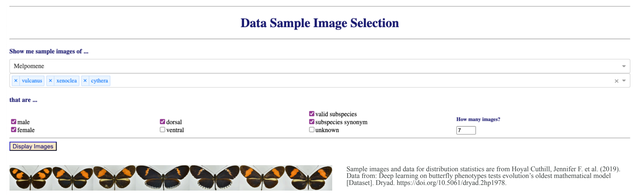
Data Exploration Tools

Data Dashboard

Facilitates data exploration: visualize distribution information and sample images efficiently.

- Telemetry dashboard needs only latitude and longitude, keeps all provided columns.
  - Broader applicability, less organization.
  - Prototype focused on plant and animal images.
- Expanded functionality and scalability in progress.
- Alternate distribution view on ESRI maps, color points by feature.



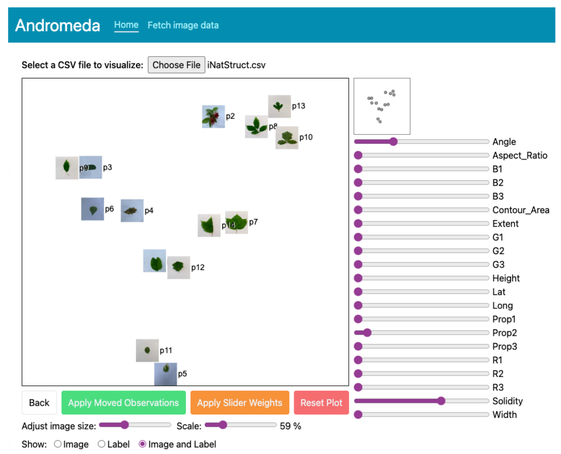


Sample images and data for distribution statistics are from Hoyal Cuthill, Jennifer F. et al. (2019). Data from: Deep learning on butterfly phenotypes tests evolution's oldest mathematical model [Dataset]. Dryad. https://doi.org/10.5061/dryad.2bp1978.

Andromeda: FAIR high-dimensional data exploration

Interact with high-dimensional data in a 2D plot using WMDS.

- Adjust weights and see how positions change; OR
- Adjust position of data points and see which features are more heavily weighted, then apply those weights to the full projection.
- Hugging Face instance also includes fetch data from iNaturalist option with a LatLonCover integration.



This material is based upon work supported by the National Science Foundation under Award No. 2118240.

Hugging Face

Follow us on Hugging Face and GitHub!

GitHub

THE OHIO STATE UNIVERSITY