Accelerating Marine UAV Drone Image Analysis with Sliced Detection and Clustering (MBARI SDCAT)

Duane R. Edgington, Danelle E. Cline, Thomas C. O'Reilly, Stephen H.D. Haddock, John Philip Ryan, Bryan Touryan-Schaefer, William J. Kirkwood, Paul R. McGill, and Rob S. McEwen

Monterey Bay Aquarium Research Institute (MBARI), Moss Landing, California USA *(duane@mbari.org)*

EGU24-2934 ITS1.2/OS4.10

Abstract

Uncrewed Aerial Vehicles (UAVs) can be a cost-effective solution for capturing a comprehensive view of surface ocean phenomena to study marine population dynamics and ecology. UAVs have several advantages, such as quick deployment from shore, low operational costs, and the ability to be equipped with various sensors, including visual imaging systems and thermal imaging sensors. However, analyzing high-resolution images captured from UAVs can be challenging and time-consuming, especially when identifying small objects or anomalies. Therefore, we developed a method to quickly identify a diverse range of targets in UAV images.

We will discuss our workflow for accelerating the analysis of high-resolution visual images captured from a Trinity F90+ Vertical Take-Off and Landing (VTOL) drone in near-shore habitats around the Monterey Bay region in California at approximately 60 meters altitude. Our approach uses a state-of-the-art self-distillation with knowledge (DINO) transformer foundation model and multi-scale, sliced object detection (SAHI) methods to locate a wide range of objects, from small to large, such as schools or individual jellyfish, flocks of birds, kelp forests or kelp fragments, small debris, occasional cetaceans, and pinnipeds. To make the data analysis more efficient, we create clusters of similar objects based on visual similarity, which can be quickly examined through a web-based interface. This approach eliminates the need for previously labeled objects to train a model, optimizing limited human resources. Our work demonstrates the useful application of state-of-the-art techniques to assist in the rapid analysis of images and how this can be used to develop a recognition system based upon machine-learning for the rapid detection and classification of UAV images. All of our work is freely available as open-source code.

A) Motivation

The intersection of population dynamics and ecology demands a synoptic understanding of physical, chemical, and biological processes at the air-sea interface. Bridging the resolution gap between surface vehicles and satellites, Unmanned Aerial Vehicles (UAVs) emerge as a cost-effective and safer alternative to crewed aircraft. UAVs can fly under cloud cover, mitigate atmospheric effects, and gather data at centimeter-scale spatial resolutions.

MBARI aims to coordinate UAVs with ships and Autonomous Underwater Vehicles (AUVs) for science surveys. We have begun this work by deploying a Trinity F90+ Vertical Take-Off and Landing (VTOL) drone. Here, we discuss our workflow for accelerating the analysis of high-resolution visual images captured from near-shore habitats around the Monterey Bay and Santa Cruz region in California at approximately 60 meters altitude.

Figure 1. Trinity flight vertical take off, Davenport Landing, California USA

First Application

- Bio-surveys over Monterey Bay
- Fly repeated transect lines and areas over a specified region
- Line-of-sight initial brief surveys, as the first step towards longer beyond line-of-sight
- Vertical Takeoff and Landing (VTOL) UAV is a good choice for this application: long-range, launch/land in confined space
- Utilize commercial COTS cameras (visible RGB to start)
- Assess datasets for jellies, plankton/kelp, birds, bioluminescence, fish, and more

How to cite

Edgington, D. R., Cline, D. E., O'Reilly, T., Haddock, S. H. D., Ryan, J. P., Touryan-Schaefer, B., Kirkwood, W. J., McGill, P. R., and McEwen, R. S.: Accelerating Marine UAV Drone Image Analysis with Sliced Detection and Clustering (MBARI SDCAT), EGU General Assembly 2024, Vienna, Austria, 14–19 Apr 2024, EGU24-2934, https://doi.org/10.5194/egusphere-egu24-2934, 2024.

B) Challenges

Small Objects

Finding small objects in large images is challenging for both humans and machines. Our camera, model SONY RX1R II, captures 7952 x 5304-pixel images or 42 megapixels. A pixel resolves to 0.77 cm.

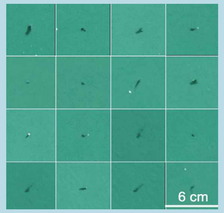


Figure 2. Cluster of floating kelp

Surface Reflections

Images acquired by the UAV camera contain signals from multiple sources.

Water-leaving radiance is light reflected by objects at or under the surface, and our science applications mostly utilize this signal.

There are additional light sources that interfere with water-leaving radiance. Specular reflected sunlight produces sparkly "sun glint" on the water surface, and reflected "skylight" produces a blue or gray sheen on the water surface. Sun glint and skylight interfere with our ability to detect and identify objects of interest in the water, and so we design our aerial surveys and protocols to reduce those effects. The first two guidelines below are recommended to us by Dr Liane S Guild et al at the NASA Airborne Science Program, based on their extensive experience with aircraft-based remote sensing over water:

- Fly at Sun elevations between 30-45 deg, to minimize sun glint while maintaining acceptable water-leaving radiance. Solar elevation of less than 30 degrees results in low water-leaving radiance, which is problematic. There are one or two elevation "windows" each day, depending on latitude and season. Note that cloud cover may result in reduced sun glint even at higher Sun elevations.
- Orient survey legs to fly directly toward and away from the Sun, for more symmetrical lighting across the image width and to optimize the effects of a polarizing filter placed over the camera lens.

We place a circular polarizing filter over the lens of the UAV Sony RX1R II camera, which can significantly reduce sun glint and sky reflection when rotated at an appropriate angle. We rotate the filter before flight to the optimal angle for flights directly into or away from the Sun. Note that the camera isn't pointed straight down while in flight, but is pitched forward several degrees.

Reflections can cause many false detections, quickly overwhelming the downstream processing. Some reflections are mitigated by following the above guidelines and by polarizing filters on the camera, but software removal can also be helpful. Using an in-painting method adapted from specularity removal for endoscopic images, we can remove some, but not all, of the reflections.

Figure 3. Example reflectance removal with in-painting

Imbalanced Data

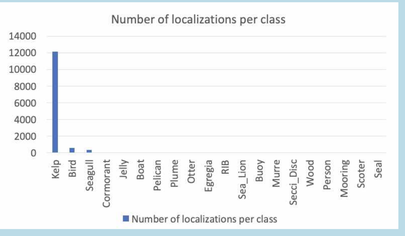


Figure 4. Localizations annotated with assistance from SDCAT from 6 missions (13,348 localizations from 619 images) demonstrate the long-tail problem in real-world data. x axis: Number of Instances per Class. y axis: Classes Sorted in Decreasing Order of Number of Instances.

The frequency of observation of different classes varies dramatically, with some common in our survey areas (e.g., Kelp) and others rare (e.g., Otters), which is a challenge for automatic object classification. Perfectly balanced data improves performance in classification models but is unrealistic in real-world data. A small portion of the classes form the majority of data, while other classes of interest lack sufficient data to be representative – a long-tail distribution.

C) Approach: Detection

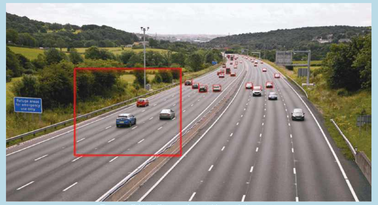


Figure. SAHI: Slicing Aided Hyper Inference

SAHI (Slicing Aided Hyper Inference) is a method that improves the detection accuracy of small objects by running object detection on slices across the image. This method is slower than the saliency map detector but has the advantage that it can detect and classify in one step. It is optimized to run across slices and combine the results.

We found that models not necessarily trained on UAV images still work for general object detection by using them in a class-agnostic way with SAHI.

Figure 4. Results SAHI detection from a YOLOS (Vision Transformer) model trained with COCO 2017 (118k annotated images of everyday objects). Twelve slices, image rescaled to 30% of the original size, maximum area 30,000, minimum confidence threshold 0.1

With enough training data, this same approach is a complete workflow to detect and classify targets. Early results from a trained YOLOV5x model on the data we labeled with the help of this workflow are promising.

Figure 5. Results SAHI detection from a model trained with our UAV data. Five slices, image rescaled to 30% of the original size, maximum area 30,000, minimum confidence threshold 0.1. Seagull is detected as before but now is also classified.

Fine-grained saliency map detection

Figure 6. (left) original image, (right) fine-grained saliency map

A faster method than SAHI leverages a fine-grained saliency map based on the luminance and saturation of the image. This produces a map on which a conventional blob detector is then run to detect blobs. Luminance is slightly weighted to strengthen brighter objects.

$$\text{saliency\_lum} = 1.2 * (\text{center\_lum} - \text{surround\_lum})$$

$$\text{saliency\_sat} = \frac{\text{center\_sat} - \text{surround\_sat}}{4}$$

$$\text{saliency\_level} = \frac{\text{saliency\_sat} + \text{saliency\_lum}}{2}$$

Both methods can be combined using NMS to leverage the strengths of each. This allows for continued discovery by combining SAHI, a supervised method, with unsupervised saliency detection.

D) Approach: Clustering DINO + HDBSCSN

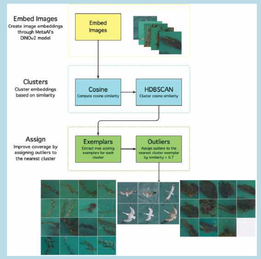


Figure 7. Cluster Algorithm Workflow

We create clusters of similar objects based on visual similarity to make the data analysis more efficient. These can be quickly examined through grids exported by the toolkit.

Our approach uses a state-of-the-art DINO transformer foundational model trained on 142 million images. Embeddings from images are collected into clusters based on cosine similarity. We found that the VITS14 worked the best to create clusters that were easy to interpret. To improve cluster coverage, outliers are assigned to the nearest cluster using exemplars with the highest score within a cluster.

Filtering

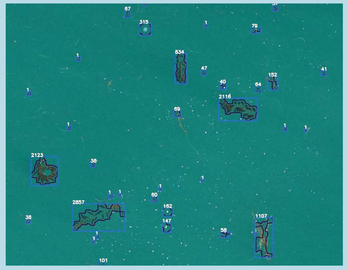


Figure 8. Example saliency scores. We found a lower value of 300 suitable for most objects of interest.

In addition to the maximum and minimum area, detections with a low saliency score or very low standard deviation in the saliency map are removed before clustering. The saliency is determined by a function that penalizes small objects with low variance based on the luminance and saliency map output.

$$factor = \frac{imagewidth}{1000}$$

$$\text{saliency} = \frac{\text{area} \times (\text{mean\_luminance} + \text{mean\_saliency} + 0.1 \times \text{area}) - \text{variance}}{\text{factor}}$$

E) Conclusions and Next Steps

Conclusions

Our approach uses a state-of-the-art DINO transformer foundational model and multi-scale, sliced object detection and saliency maps to locate a wide range of objects, from small to large, such as individual jellyfish, flocks of birds, kelp forests or kelp fragments, small debris, occasional cetaceans, and pinnipeds.

To make the data analysis more efficient, we create clusters of similar objects based on visual similarity, which can be quickly examined. This approach eliminates the need for previously labeled objects to train a model, optimizing limited human resources.

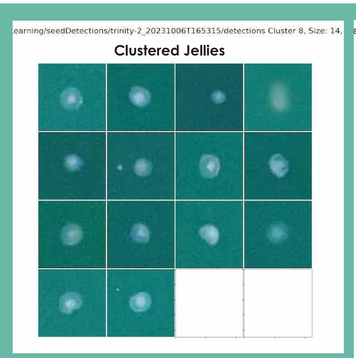
This method has been successfully applied to six missions, and localizations have been used to create an object detection model to improve object detection.

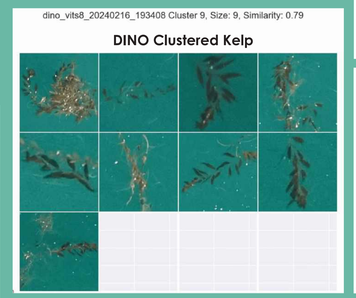
Our work demonstrates the helpful application of state-of-the-art techniques to assist in rapidly analyzing images.

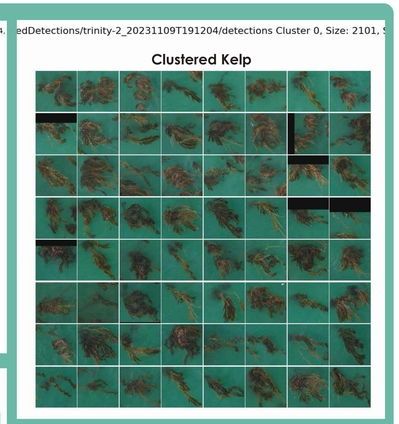
All of our work is freely available as open-source code at
https://github.com/mbari-org/sdcat

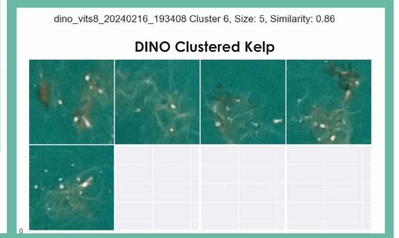
Next steps

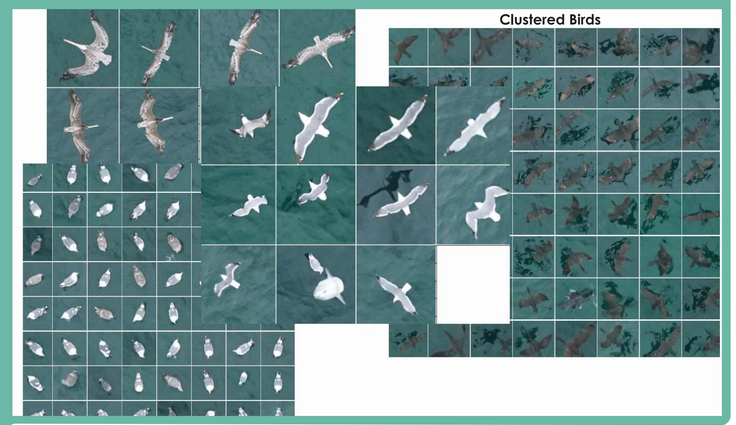
- Use the clustering algorithm for predicting classes. This could address the long-tail problem.
- Label more clustered detections, with a focus on rarely seen objects.
- Training on image slices or augmented data to improve the object detection model.











References

SAHI
Akyon, F. C., Altinuc, S. O., & Temizel, A. (2022). Slicing Aided Hyper Inference and Fine-Tuning for Small Object Detection. In 2022 IEEE International Conference on Image Processing (ICIP) (pp. 966-970). Bordeaux, France. doi:10.1109/ICIP46576.2022.9897990. Link

DINO
Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., & Joulin, A. (2021). Emerging Properties in Self-Supervised Vision Transformers. In Proceedings of the International Conference on Computer Vision (ICCV).

Fine-Grained Saliency
Sebastian Montabone and Alvaro Soto. "Human detection using a mobile platform and novel features derived from a visual saliency mechanism." In Image and Vision Computing, Vol. 28, Issue 3, pages 391–402. Elsevier, 2010. DOI:10.1016/j.imavis.2009.06.006.

NMS
R. Girshick, J. Donahue, T. Darrell, and J. Malik. "Rich feature hierarchies for accurate object detection and semantic segmentation." In CVPR, 2014.

Inpainting
Telea, Alexandru. "An image inpainting technique based on the fast marching method." Journal of Graphics Tools 9.1(2004): 23-34.

VTOL
Quantum Systems. (n.d.). Trinity F90 Pro. Retrieved from Link
Quantum Systems. (n.d.). F90 User Manual V2.3.0.49. Retrieved from Link