Species-level Detection and Tracking of Caribbean Coral Reef Fish

Levi Cai[1,2*] Austin Greene[2*] Daniel Yang[1,2*] Nadège Aoki[1,2] Sierra D. Jarriel[2]
Jasper Ha[2,3] Nathan Formel[2] T. Aran Mooney[2] Yogesh Girdhar[2]
[1]Massachusetts Institute of Technology [2]Woods Hole Oceanographic Institution [3]UC San Diego
https://warplab.github.io/corefish/

Abstract

Species-level detection and tracking of fish from videos is a crucial capability for ecologists to study and understand trends in bio-diversity as they relate to ecosystem changes. However, these capabilities still trail their terrestrial equivalents. This paper presents a dataset and initial experiments to understand the current state and limitations for this technology, highlighting the constraints encountered in an ecological fieldwork scenario. Our dataset focuses on Caribbean coral reef fish, where ecologists have gathered consistent data over 8 years across 5 distinct reef sites. This dataset leverages diver-collected visual survey data, unlike alternative datasets, which typically either rely on stationary, baited camera traps or alternative methods of sourcing data, such as caught animals or tourist captured videos. Our dataset consists of 14,580 expert-annotated frames, including 15,802 unique fish tracks consisting of 98,267 bounding boxes. Of those tracks, 2,410 of them include species-level annotations from 54 species found in these waters. Our preliminary experiments highlight the challenges in applying typical fish datasets and detection methods to a novel ecological context.

1. Introduction

Marine biologists use diver surveys to gather information about biodiversity and species-level trends about certain ecosystems, such as coral reefs. These surveys are conducted regularly in the same regions and are typically analyzed at the species level to determine broader ecosystem trends [8, 19]. This data is difficult and time consuming to analyze, requiring specialist domain expertise, often resulting in vast quantities of unanalyzed data.

Visual detectors have shown incredible promise for automating video analysis of biodiversity data, especially in terrestrial settings [4, 10]. These are trained on massive datasets of terrestrial organisms to build broad "animal" detectors, which can then be used in conjunction with species-level classifiers for finer grained analysis. In the past few years, many large datasets consisting of detection and classification tasks for fish and other marine species have been released [7, 9, 12, 13, 16–18]. However, those datasets still are insufficient for training similar detectors and classifiers, and currently no known method of fish detection and classification has achieved similar adoption, both from broader detection levels (at the fish-level) and especially at the species level.

We hypothesize this difficulty is due to several key issues, including (1) a lack of scale and diversity of data, (2) a lack of high resolution and high quality data due to attenuation and turbidity of water, and (3) the relatively small scale of many of the organisms of interest. Underwater data is often collected and annotated as video clips, which may contribute to the lack of data diversity as subsequent frames are often similar. However, we also note that during data annotation human labelers often rely on motion and multiple viewpoints to successfully identify marine organisms in images, highlighting an avenue for further research,

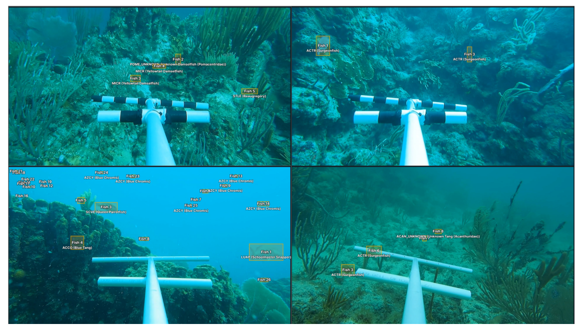

Figure 1. CoReFish dataset example images. Our dataset consists of expert-annotated bounding box tracks across frames from ego-centric videos collected by divers performing visual surveys as they swim across reef sites in the US Virgin Islands. Some of the challenges in detecting and classifying fish from such visual data include the range of sizes, visual similarities between the fish and the reef background environment, and visual degradation from imaging through a medium.

[*]Equal contribution

that tracking or motion will help in detection. Furthermore, tracking helps in modeling and extrapolation, since more area is covered over samples (when compared with a single image frame in a region), and properly tracked organisms can avoid being double counted.

1.1. Related Works

Large-scale visual detectors for ecology have seen widespread adoption in terrestrial settings. MegaDetector [4] and SpeciesNet [10], which have been trained on extremely large camera trap datasets such as iWildCam [5], iNaturalist [20], and CalTech CameraTraps [4], resulting in over 50M images. These have been used to significantly reduce time during post-processing and analysis to answer varied ecological questions about bio-diversity and trends.

Similar efforts for the marine domain are increasing. Many datasets have started to proliferate, such as NOAA Labeled Fishes in the Wild [6], [17]. Some of the largest and most recent efforts include VIAME FishTrack23 [7], FishNet (KAUST) [13], and FathomNet (MBARI) [12]. Though the number of publicly available, well-labeled, fish images is nearly 1M, this is still an order of magnitude fewer than data in the terrestrial domain, and much less diverse. The marine domain is particularly challenging, due to difficulty of human access to oceans, few experts, and optical issues like light attenuation and turbidity. Notably, we believe many of these datasets do not capture realistic settings that marine ecologists rely on for fish surveys.

2. Dataset

We present the WHOI Reef Solutions Initiative-Caribbean Coral Reef Fish (CoReFish) dataset, which contributes a detection and tracking-oriented image dataset targeted at training and evaluating models at the species-level with ego-motion (from diver perspectives). It consists of roughly 14K annotated frames from 162 diver transects, resulting in over 15K individually tracked fish, of which over 2K are labelled at the species-level. In particular, while other datasets such as VIAME FishTrack23 support ecologist use-cases that leverage baited camera traps, this dataset is useful for swimming fish surveys.

2.1. Collection

This dataset is assembled from fish surveys collected by divers in the United States Virgin Islands National Park, St. John, annually from the years 2016 to 2023 across 5 distinct coral reef sites shown in Figure 2. Following established procedures [2], an Olympus PEN Lite E-PL5 camera in an underwater housing is mounted to a PVC apparatus, as shown in Figure 3, and divers swim 30-meter transects, slowly. The PVC apparatus is 1m in length outward from the camera, with a 25cm cross bar at 50cm from the camera, and a 50cm cross bar at 1m from the camera. The cross bars

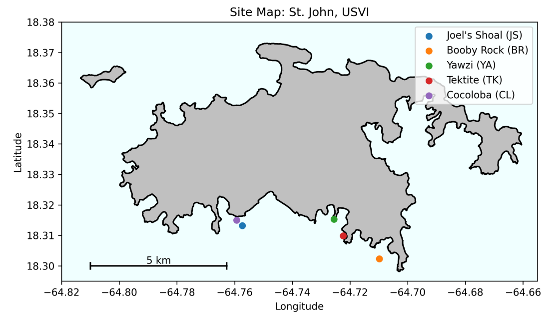

Figure 2. Map of data collection sites from 2016 until 2023.

are used by biologists to estimate fish size, a crucial biological metric for studying energetics or other ecological traits. Videos were captured in 1080p (1920x1080) resolution at 30 fps.

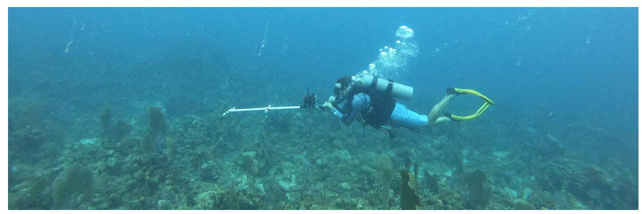

Figure 3. Image showing the data collection process. Every year since 2016, our scientists have run consistent diver video surveys for fish population.

2.2. Processing and labeling

Initially, we have data from 162 transects consisting of variable-length, 30 fps ego-centric video. We extract 30 second samples from each of these transect videos and then downsample the framerate to 3 fps. Thus, each transect yields 90 frames to be annotated.

We utilize data labeling services through iMerit on the AWS Sagemaker platform [1] to perform bounding box annotations of fish across our selected frames. Critically, the AWS Sagemaker interface enables moving back and forth between consecutive frames, which empirically aids in the detection of fish which may be quite small as well as visually degraded giving the harsh underwater imaging environment. Intuitively, one can reason that perceived motion of an object between frames that is not explained by ego-motion is likely to be originating from an independent source such as a fish. From this process, each of our subsampled transect ego-videos contains annotated bounding boxes of unique fish moving across frames.

Unfortunately, species-level fish annotation is a more specialized process, requiring expert domain knowledge to properly annotate. We import our annotated image frame sequences, consisting of many fish tracklets, into LabelBox [14] and partner with in-house biologists to perform this final level of expert species-level annotation. Because of the

Property	Value
# transects/clips	162
# frames	14580
# bboxes	100074
# unique fish/tracks	15802
Mean fish per frame	8.9
# species	54
# tracks labeled at species-level	2410

Table 1. Overview properties of the CoReFish dataset.

intensive nature of this step, only a subset of clips were annotated at the species level.

2.2.1. Temporal and spatial splits

While this dataset can be used for cross-validation tasks or various other purposes, we also provide standardized train/val/test splits by *year*. It is important for fair evaluation of detectors and trackers to test for *extrapolation* performance, and so images from the same video should be included in the same split. If need be, for purposes such as testing continual learning algorithms, these can be split based on time, but it is recommended that if a frame t_0 and t_f are included from the same video in a particular split, then all frames between $[t_0, t_f]$, inclusive, are also included in the same split, to avoid testing for *interpolation* performance (which will typically be much higher than extrapolation performance). Splits based on site (spatial splits) are also valid, but we opted for by year in order to validate general trends as studied in [8].

2.3. Statistics

Our dataset statistics are described in Table 1. Overall, the dataset captures a wide range of species, tracks, coral reef sites and water conditions. Like other natural datasets, the species distributions are long-tailed as shown in Figure 5.

2.4. Comparison to other datasets

Our dataset enables benchmarking on three challenging tasks — detection of fish at the object level, detection and classification of fish at a species level, and then tracking of fish across sequential frames. Each of these tasks are commonly encountered from a field scientist's perspective and provide varying granularity of data with which to assess biodiversity of a region. As shown in Table 2, VIAME FishTrack23 [7] comes closest to our dataset in terms of characteristic but is visually significantly different, likely a result from the use of stationary camera traps, often in greyscale. Many other datasets contain unnatural data, where fish are not imaged *in-situ* (e.g. a fisherman holding the fish in the air). No other dataset contains diver-centric egomotion which is among the most common methods for ecologists to visually survey an environment.

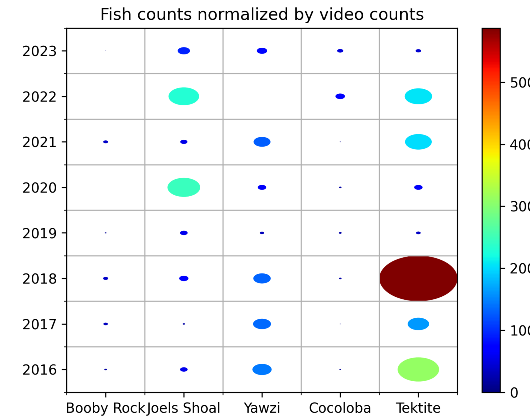

Figure 4. Number of fish by site and year, normalized by the number of transects/videos used. Size and color both represent number of fish, the blanks in Booby Rock and Yawzi sites in 2020 and 2022 had no transects labeled from those years. The years 2016 and 2017 are used for test data because they can be directly mapped to previous ecological studies.

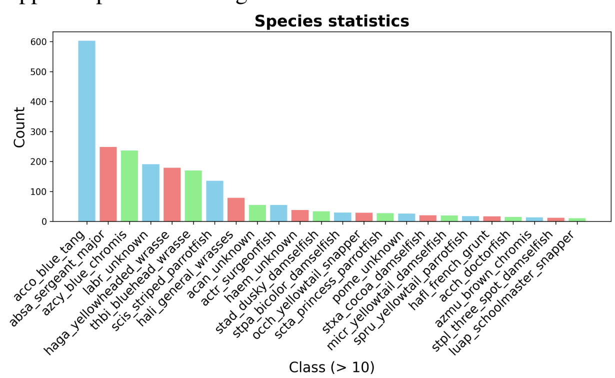

Figure 5. CoReFish species distributions.

3. Benchmark Experiments

3.1. Fish-level Detection

We established benchmark performance for fish-level detection tasks, which are shown in Table 3. In general we attempted to understand answers to the following questions which are common for ecologists and practitioners training their own models:

1. Are pre-existing datasets with coral reef fish sufficient to train detectors of coral reef fish in our coral fish survey context?
2. Given a pre-trained fish model, is it helpful to fine-tune on our new dataset, or should one train from scratch (though, still pre-trained on COCO)?
3. Given other datasets, is it beneficial to train all together rather than fine-tuning?
4. How well do models trained on our data extrapolate to slightly out of sample environments?

Dataset	Images	Detection	Classification	Species	Tracking	Natural	Coral Reef Fish	Ego-motion
NOAA Puget Sound Nearshore Fish 2017-2018 [9]	77739	✓						
Brackish Fish [17]	14674	✓			✓			
FathomNet (Gnathostoma, Nov. 2024) [12]	18289	✓	✓	✓		✓		
DeepFish [18]	620	✓	✓	✓		✓		
Kakadu Fish AI [11]	44412	✓	✓	✓		✓		
FishNet (KAUST) [13]	94778	✓	✓	✓			✓	
AIMS OzFish (Bounding Box) [16]	1758	✓	✓	~		✓	✓	
VIAME FishTrack23 [7]	1M	✓	✓	✓	✓	✓	✓	
CoReFish (Ours)	14670	✓	✓	✓	✓	✓	✓	✓

Table 2. Comparison of major fish datasets used to train detection, classification, and tracking models. We note that AIMS OzFish provides an alternative dataset consisting of only fish crops for classification tasks. Our dataset is most similar to the VIAME FishTrack23 dataset, also consisting of natural video clips annotated for tracking at the species-level. However, most of their imagery is taken from stationary camera traps, often in grey-scale, and are not suitable for evaluating methods that exhibit ego-motion from diver-held surveys.

We see that training on pre-existing datasets only, is insufficient, with training on AIMS being the best at 0.19 mAP50. What is surprising though is that training on AIMS, with about 1200 images, outperforms training on VIAME Seamap dataset, which is nearly 300K images, and both are of coral reef fish. We hypothesize that the AIMS dataset and ours are both colored (rather than greyscale of VIAME Seamap), and that the AIMS dataset is more explicitly diverse. Kakadu and FathomNet are freshwater fish and deep ocean fish respectively, and as expected, perform poorly.

Also, we find in our preliminary analysis that it is *not* beneficial to fine-tune a pre-existing fish detector model. We see that a model that has been pre-trained on the AIMS dataset and then fine-tuned on our dataset, performs worse than having only fine-tuned on our dataset starting from a pre-trained COCO model [15]. We suspect there are minor problems associated with catastrophic-forgetting in these cases. However, simultaneously training with more data *does* provide benefits, seeing gains from 0.364 mAP50 on CoReFish[fish] only data to 0.369 mAP50 when trained with AIMS and Kakadu datasets (even though Kakadu is a substantially different domain and showed poor performance originally).

Pre-training	Fine-tuning	Test	mAP50	mAP95
COCO	AIMS	CoReFish[fish]	0.19	0.082
COCO	Kakadu	CoReFish[fish]	0.071	0.030
COCO	FathomNet-Gnathostoma	CoReFish[fish]	0.008	0.003
COCO	VIAME-SEAMAP-fish	CoReFish[fish]	0.02	0.009
COCO	CoReFish[fish]	CoReFish[fish]	0.364	0.173
COCO	AIMS + CoReFish[fish]	CoReFish[fish]	0.352	0.169
COCO	AIMS + Kakadu + CoReFish[fish]	CoReFish[fish]	**0.369**	**0.182**
AIMS	CoReFish[fish]	CoReFish[fish]	0.35	0.169

Table 3. We compare fine-tuning vs. fully pre-trained networks (all are YOLO11m trained at 1280p resolution), alongside similar fish detection datasets. The VIAME dataset was subsampled to only include the Seamap dataset, which consists of 320,142 images of coral reef fish and thus should be similar to our dataset.

Model	Test dataset	mAP50	mAP95
CoReFish[fish]	GoPro videos	0.615	0.369
CoReFish[fish]	Robot downward-facing videos	0.116	0.0392
CoReFish[fish]	Robot forward-facing videos	0.043	0.0119

Table 4. Demonstrating slightly out-of-sample test performance of a YOLO11m model trained on only our dataset. All these datasets are taken from the same sites, but using different cameras and platforms. GoPro videos are from divers. Robot videos are collected from the CUREE platform [3], which use OAK-D cameras in a forward-facing and downward-facing manner.

Pre-training	Fine-tuning	Testing	mAP50	mAP95
AIMS	CoReFish[species]	CoReFish[species]	0.036	0.019
Kakadu	CoReFish[species]	CoReFish[species]	0.033	0.016
COCO	CoReFish[species]	CoReFish[species]	**0.081**	**0.045**

Table 5. Species-level detection performance benchmarks, comparing fine-tuning from different pre-trained models.

3.2. Species-level Detection

We also ran preliminary benchmarks for species-level detection, which are shown in Section 3.2. We see that species-level detection with YOLO is substantially more difficult and that again, naive YOLO11 pre-training on pre-existing datasets is not sufficient. This highlights the difficulty of this fine-grain classification and detection task.

4. Conclusions

We provided a unique dataset of coral reef fish that can be used by ecologists to train new visual detection models, and also to evaluate how well these models work in realistic and directly applicable settings. The dataset is unique in that it covers Caribbean coral reef fish, labeled to the species-level, and provides classification, detection, and tracking tasks all at once. Furthermore, to our knowledge, it is the only known densely-labelled fish dataset that demonstrates ego-motion, which is useful for many ecologists leveraging fish surveys or robotic solutions for monitoring. Furthermore, we provide initial insights that fine-tuning on consecutive fish detection datasets tends to hurt performance, whereas concurrent training helps.

References

[1] Amazon Web Services. Amazon sagemaker ground truth, 2024. 2

[2] Anonymous Anonymized. Anonymized. *Marine Pollution Bulletin*, 136:282–290, 2018. 2

[3] Anonymous Anonymized. Anonymized. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 11411–11417, 2023. 4

[4] Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in Terra Incognita. In *Computer Vision – ECCV 2018*, pages 472–489, Cham, 2018. Springer International Publishing. 1, 2

[5] Sara Beery, Arushi Agarwal, Elijah Cole, and Vighnesh Birodkar. The iWildCam 2021 Competition Dataset, 2021. arXiv:2105.03494 [cs]. 2

[6] George Cutter, Kevin Stierhoff, and Jiaming Zeng. Automated Detection of Rockfish in Unconstrained Underwater Videos Using Haar Cascades and a New Image Dataset: Labeled Fishes in the Wild. In *2015 IEEE Winter Applications and Computer Vision Workshops*, pages 57–62, 2015. 2

[7] Matthew Dawkins, Jack Prior, Bryon Lewis, Robin Faillettaz, Thompson Banez, Mary Salvi, Audrey Rollo, Julien Simon, Matthew Campbell, Matthew Lucero, Aashish Chaudhary, Benjamin Richards, and Anthony Hoogs. FishTrack23: An Ensemble Underwater Dataset for Multi-Object Tracking. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 7152–7161, Waikoloa, HI, USA, 2024. IEEE. 1, 2, 3, 4

[8] Jason P. Dinh, Justin J. Suca, Ashlee Lillis, Amy Apprill, Joel K. Llopiz, and T. Aran Mooney. Multiscale spatiotemporal patterns of boat noise on U.S. Virgin Island coral reefs. *Marine Pollution Bulletin*, 136:282–290, 2018. 1, 3

[9] Dara M. Farrell, Bridget Ferriss, Beth Sanderson, Karl Veggerby, Lauren Robinson, Anusua Trivedi, Shreyaan Pathak, Sreya Muppalla, Jane Wang, Dan Morris, and Rahul Dodhia. A labeled data set of underwater images of fish and crab species from five mesohabitats in Puget Sound WA USA. *Scientific Data*, 10(1):799, 2023. Publisher: Nature Publishing Group. 1, 4

[10] Tomer Gadot, Ștefan Istrate, Hyungwon Kim, Dan Morris, Sara Beery, Tanya Birch, and Jorge Ahumada. To crop or not to crop: Comparing whole-image and cropped classification on a large dataset of camera trap images. *IET Computer Vision*, 18(8):1193–1208, 2024. 1, 2

[11] Andrew Jansen, David Walden, Samantha Walker, and Constanza Buccella. A deep learning dataset for underwater object detection of tropical freshwater fish species in northern Australia, 2022. 4

[12] Kakani Katija, Eric Orenstein, Brian Schlining, Lonny Lundsten, Kevin Barnard, Giovanna Sainz, Oceane Boulais, Megan Cromwell, Erin Butler, Benjamin Woodward, and Katherine L. C. Bell. FathomNet: A global image database for enabling artificial intelligence in the ocean. *Scientific Reports*, 12(1):15914, 2022. Publisher: Nature Publishing Group. 1, 2, 4

[13] Faizan Farooq Khan, Xiang Li, Andrew J. Temple, and Mohamed Elhoseiny. FishNet: A Large-scale Dataset and Benchmark for Fish Recognition, Detection, and Functional Trait Prediction. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 20439–20449, 2023. ISSN: 2380-7504. 1, 2, 4

[14] Labelbox, Inc. Labelbox: The training data platform for ai teams, 2024. 2

[15] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common Objects in Context. In *Computer Vision – ECCV 2014*, pages 740–755, Cham, 2014. Springer International Publishing. 4

[16] Daniel Marrable, Kathryn Barker, Sawitchaya Tippaya, Mathew Wyatt, Scott Bainbridge, Marcus Stowar, and Jason Larke. Accelerating Species Recognition and Labelling of Fish From Underwater Video With Machine-Assisted Deep Learning. *Frontiers in Marine Science*, 9, 2022. Publisher: Frontiers. 1, 4

[17] Malte Pedersen, Joakim Bruslund Haurum, Rikke Gade, and Thomas B Moeslund. Detection of Marine Animals in a New Underwater Dataset with Varying Visibility. 2, 4

[18] Alzayat Saleh, Issam H. Laradji, Dmitry A. Konovalov, Michael Bradley, David Vazquez, and Marcus Sheaves. A Realistic Fish-Habitat Dataset to Evaluate Algorithms for Underwater Visual Analysis. *Scientific Reports*, 10(1): 14671, 2020. arXiv:2008.12603 [cs, eess]. 1, 4

[19] Melita A. Samoilys and Gary Carlos. Determining Methods of Underwater Visual Census for Estimating the Abundance of Coral Reef Fishes. *Environmental Biology of Fishes*, 57 (3):289–304, 2000. 1

[20] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The iNaturalist Species Classification and Detection Dataset. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8769–8778, Salt Lake City, UT, 2018. IEEE. 2