FishTrack23: An Ensemble Underwater Dataset for Multi-Object Tracking

Matthew Dawkins[1], Jack Prior[2], Bryon Lewis[1], Robin Faillettaz[3], Thompson Banez[4], Mary Salvi[1], Audrey Rollo[2], Julien Simon[3], Matthew Campbell[2], Matthew Lucero[4], Aashish Chaudhary[1], Benjamin Richards[2], Anthony Hoogs[1]

[1]Kitware Inc., USA {matt.dawkins,bryon.lewis,mary.salvi,aashish.chaudhary,anthony.hoogs}@kitware.com
[2]NOAA NMFS, USA {jack.prior,matthew.d.campbell,audrey.rollo,benjamin.richards}@noaa.gov
[3]DECOD, Ifremer, INRAE, L'Institut Agro, Lorient, France {robin.faillettaz,julien.simon}@ifremer.fr
[4]California Department of Fish and Wildlife, USA {matthew.lucero,thompson.banez}@wildlife.ca.gov

Abstract

Tracking and classifying fish in optical underwater imagery presents several challenges which are encountered less frequently in terrestrial domains. Video may contain large schools comprised of many individuals, dynamic natural backgrounds, highly variable target scales, volatile collection conditions, and non-fish moving confusers including debris, marine snow, and other organisms. Additionally, there is a lack of large public datasets for algorithm evaluation available in this domain. The contributions of this paper is three fold. First, we present the FishTrack23 dataset which provides a large quantity of expert-annotated fish groundtruth tracks, in imagery and video collected across a range of different backgrounds, locations, collection conditions, and organizations. Approximately 850k bounding boxes across 26k tracks are included in the release of the ensemble, with potential for future growth in later releases. Second, we evaluate improvements upon baseline object detectors, trackers and classifiers on the dataset. Lastly, we integrate these methods into web and desktop interfaces to expedite annotation generation on new datasets.

1. Introduction

An increasing number of optical sensors are being deployed underwater for applications such as robotics, exploration, sustainable fisheries management [1], wind farm inspection, and other applications. In most of these domains, the amount of imagery collected by autonomous underwater vehicles (AUVs), towed camera rigs, and drop cameras can be prohibitively expensive to analyze manually, necessitating automated methods to derive insights from data. Simultaneously, machine learning techniques are highly dependent on having curated, accurate, and representative datasets to create operationally-ready solutions. With the advancement of deep learning for detection [30], instance segmentation [18] [4], and tracking [12], a high quantity of annotations is required to train neural networks containing millions of weights.

While not required for every application, object tracking can be used to identify the same individual across a sequence and, if accurate enough, further improve population census counts. Alternatives to tracking for population estimation include metrics, such as MaxN [8,9], which utilize individual frames with a maximum number of target species within a chosen time window, then further extrapolated over multiple collects. To date, while there have been a few fish datasets posted online [17,20,33], there has not been a large release of annotations at the track level. In this work, we introduce FishTrack, an ensemble dataset approaching 1 million boxes and frames, with ∼25k individual tracks divided into a test and train split. It is our hope that this dataset can be used to benchmark competing tracking algorithms, as applied to the underwater domain. It currently contains data collected by the National Oceanic and Atmospheric Administration (NOAA), Institut Français de Recherche pour l'Exploitation de la Mer (Ifremer) and California Department of Fish and Wildlife (CDFW), but can also be expanded to include new datasets in future releases. The data is publicly available for download[1], alongside availability via academic torrent[2].

1.1. Prior Work

There are a number of existing fish detection datasets including DeepFish [33], Fish4Knowledge [17], and FishClef [20]; however, these datasets contain only single frame labels not linked into tracks, and in the case of [20], low resolution cameras. This ensemble contains both species labels, box labels, and object tracks on a variety of higher definition cameras. It is also a magnitude larger than other datasets currently available. One exception is the work of [5], also containing tracks, of which an updated ver-

[1] https://viame.kitware.com/#/collections

[2] https://academictorrents.com

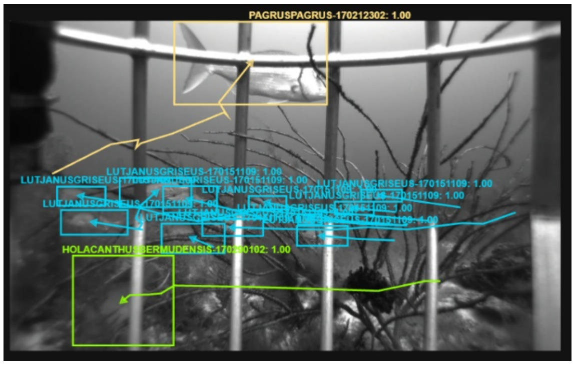

PAGRUSPAGRUS-170212302: 1.00
LUTJANUSGRISEUS-170151109: 1.00
HOLACANTHUSBERMUDENSIS-170200102: 1.00

Figure 1. A high number of overlapping targets

sion is included as part of this ensemble. DeepFish [33] contains polygonal annotations which the initial release of this dataset lacks, though adding segmentation masks is a goal for later releases, for the purpose of both improving performance and allowing the comparison of instance segmentation techniques. The work of [6] contains single object tracking annotations from a moving underwater platform, while the annotations in this archive are all multi-object. Lastly, WATB benchmark [36] focuses on single-target tracking as well, across a wider range of animal.

2. Dataset Overview

2.1. Challenges

There are several challenges associated with accurately tracking fish across a video, some of which are shared by tracking problems in other domains, while others unique to underwater processing. The most difficult sequences contain many individuals with varying scales swimming near each other, leading to a very high level of short- and long-term occlusions (see Figure 1). Target confusors with similar motion signatures include debris (Figure 2), marine snow, moving sediment when platforms hit the benthic substrate, heterogeneous backgrounds when the cameras move, as well as other organisms (invertebrates, jellyfish, etc.).

Variable natural backgrounds, bottom substrates, and water turbidity can lead to certain individuals having very low contrast or being camouflaged (Figure 3). Due to turbidity, fish may come in and out of visibility, even without any other objects obstructing them (Figure 4). In some cases, there are virtually no single frame appearance features visible for an individual, just a subtle motion signature (Figure 5). Lastly, there is a wide range of optical sensors deployed in different cases, of varying resolutions and color properties (e.g. color vs greyscale, baited vs unbaited, different white balancing capabilities). Many of the sensors in this release rely on ambient lighting, though others also utilize artificial lighting attached to the collection platform.

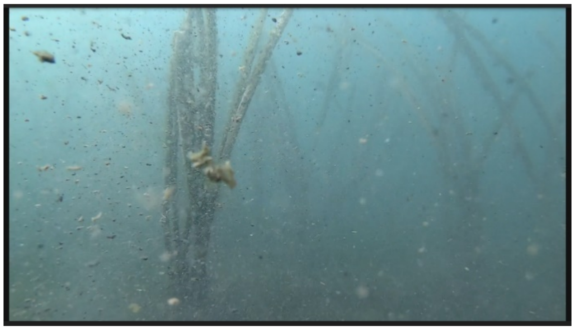


Figure 2. Moving debris and sediment in the water

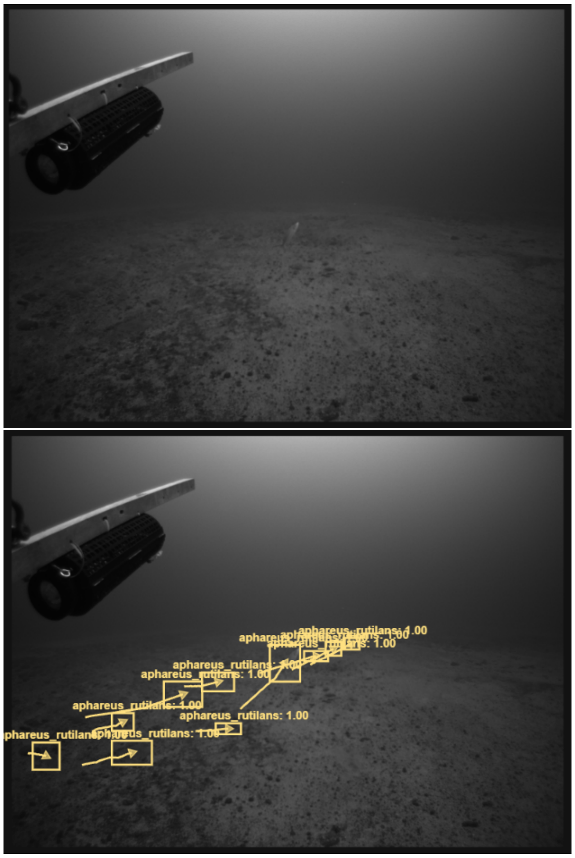


Figure 3. Low-contrast movers, with and without track display, across the background substrate which camouflages the targets.

2.2. Statistics and Properties

For consistent evaluation of methods, we divide our release into a recommended train and test set, containing 23k and 3.5k tracks respectively. Videos were manually selected to generate this split, ensuring that both a range of difficulties and multiple locations were included in the test set. Approximately two hundred species labels are included, primarily from ocean-dwelling fish, along with a smaller quantity of freshwater fish from lakes and rivers. In most cases

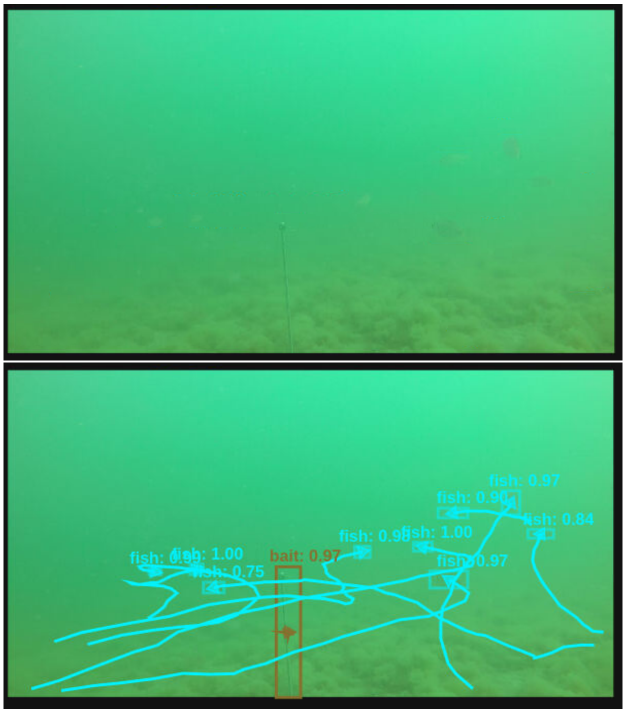


Figure 4. Low contrast detections disappearing into the background due to lighting and water conditions. The same image is shown with and without fish and bait tracks drawn on the video.

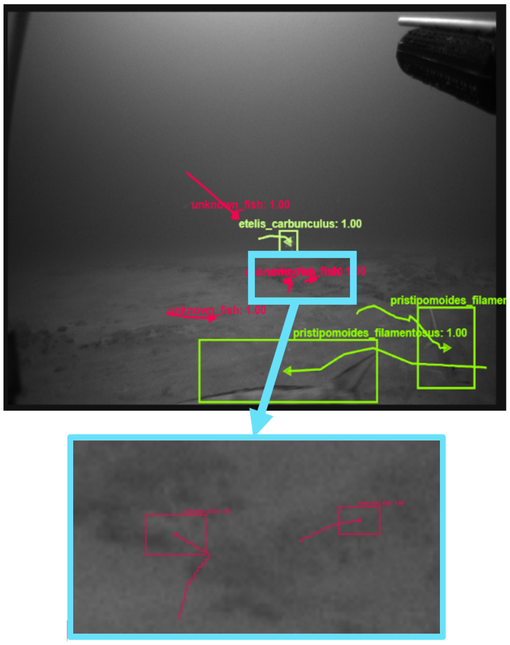


Figure 5. Very small, low-SNR targets which are usually only observable from a slight motion signature, identified via moving the video slider back and forth in the annotation software.

the species of the fish is provided, though in cases where the contrast is too low to determine the species, alongside a portion of videos (∼10% of the archive), an ‘unknown_fish’ label is used. Data was annotated at either 5hz or 10hz, even when the native video frame rate was higher. This served as a balance between annotator time and full video coverage. Resolution of the video frames span from 1280x720 (720p) to 1920x1080 (1080p) from an assortment of sensors. Collected data is all of the optical spectrum, though a portion is black and white instead of full color imagery.

The dataset is available for standard download[1], in addition to an academic torrent[2]. Tracks are provided both in a CSV format and a variant of the COCO JSON schema [26]. For those interested in adding data to future releases, an open-source web-based annotator, alongside an additional desktop application, is provided. Within these interfaces, new data can either be annotated from scratch, or alternatively, baseline trackers can be run on new datasets then corrected for rapid annotation. New annotations can also be added from external tools in one of the above formats and submitted for review through the interface for inclusion in future releases of the ensemble. The full dataset is comprised of subsets of data further described in the following sections.

2.3. Dataset Subcomponents

2.3.1 Ifremer DropCam

The Ifremer DropCam dataset, collected in the Bay of Biscay for the BAITFISH project, is a collaboration amongst scientists, fishermen, and stakeholders to develop baited traps for commercially targeted species, particularly gilthead seabream (*Spondyliosoma catharus*). It focuses on recording and analyzing the feeding behaviors of fish around different types of traps and tested baits (i.e., cockles, mussels, or other baits mixed with biodegradable/water-soluble plastic materials to lengthen the diffusion). Several hundreds of 9-hour videos were recorded in the summers of 2019, 2020, and 2021 using a drop cam system, offering an extended dataset for this temperate species. GoPro Hero 4, 5 and 7 Black, recording at 24 fps and 1080p resolution, were used to collect the 9-hour videos, during daytime hours only (see Figure 6). Three species of fish are annotated in the dataset, with gilthead seabream largely dominating in abundance (see Acknowledgement).

2.3.2 Game of Trawls

The Game of Trawls dataset was also collected in the Bay of Biscay, but inside a pelagic trawl prototype. Recording was performed with either a GoPro 4 Black and a Zed Mini camera using artificial lighting (4 x 2000 lumens), the data from which can be observed in Figure 7. Five genus/species of fish are annotated in the dataset. The project strives to develop intelligent fishing gears using underwater imaging,

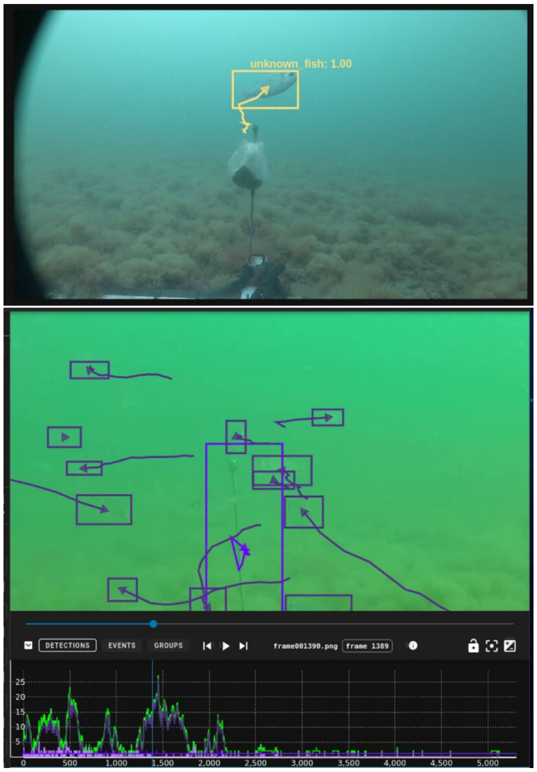


Figure 6. Ifremer DropCam Example Imagery

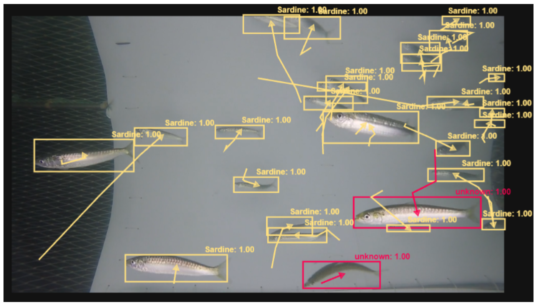


Figure 7. Example of the Game of Trawls Dataset

acoustic communication, artificial intelligence, and computer vision techniques. It integrates these technologies to remotely control and automate a selective device in fishing gears, such as beam or pelagic trawls, to actively detect and guide the escapement and selection of target species. The ecological impact of the fishing gears is therefore drastically reduced as the species are sorted mid-water, maximizing their probability of survival. While this is the one dataset which includes a non-natural background, it is mostly solid or consists of netting. With the one exception of slightly less range in terms of lateral movement towards and away from the camera, most movement properties of the fish are shared with those in natural conditions and the other sets, with frequent fish occlusions.

2.3.3 SEAMAP Reef Fish Video Survey

The SEAMAP video survey has been conducted annually since 1992. Traditional fisheries gears (e.g., trawls) often damaged the habitat and were less effective for reef species that the U.S. National Marine Fisheries Service is required to manage. In the late 80's underwater video became more accessible and researchers began investigating how to use the technology to evaluate abundance of marine resources. The survey is primarily conducted on the high-relief shelf edge break habitat of the northern Gulf of Mexico from the U.S.-Mexico international border to the Dry Tortugas, FL. Data is typically collected at depths ranging from 15-200 m. The primary deployment gear in the survey has been a stationary, ground-tended array utilizing various camera technologies and providing a 360 field-of-view by stitching imagery from five cameras and stereo-vision in a single orientation via an attached satellite camera (see Figure 8). In this dataset, only the view from a single camera was annotated (not the stitched composite array). All imagery in the subset is black and white, without color information, as shown in Figures 1 and 9. While species classification is not the primary goal of the FishTrack compilation, one of the central difficulties of this particular subset is the long-tail distribution of species labels. For example, it contains tens of thousands of red snapper images (*Lutjanus Campechanus*) vs. just a couple of instances of other species of fish, with a large variance across all 150 species.

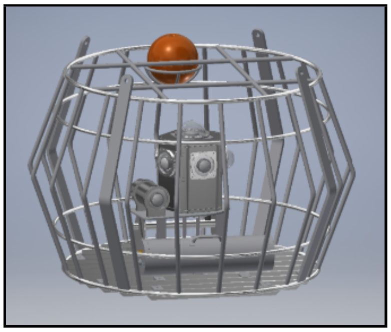


Figure 8. SEAMAP Collection Platform

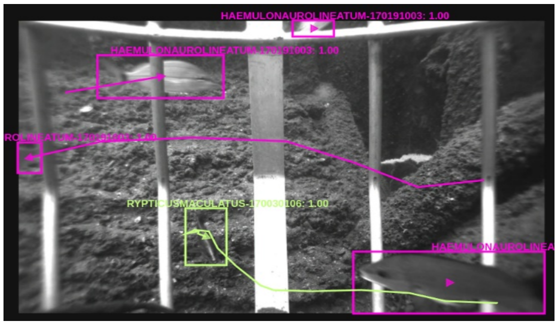

HAEMULONAUROLINEATUM-170191003: 1.00
HAEMULONAUROLINEATUM-170191003: 1.00
RYPTICUSMACULATUS-170031106: 1.00
HAEMULONAUROLINEA

Figure 9. SEAMAP Example Imagery

2.3.4 CDFW Habitat Monitoring Project

In freshwater environments, highly mobile species are difficult to sample. Boat electroshocking, gill netting and angling are common methods of surveying lakes and reservoirs. These methods allow for detection of abundance, diversity, size distribution, and condition factors. However, they do not provide site nor species specific habitat occupation over time. To address these issues, underwater camera traps were deployed to capture various habitat types in a southern California reservoir (Lake Jennings). Camera mounts were constructed from ¼ inch PVC pipe to create a 1x1 foot square base with a 1x1 foot square vertical top. GoPro Hero 7 Black action cameras were used to collect 1080p, 30hz video as shown in Figure 10, below.

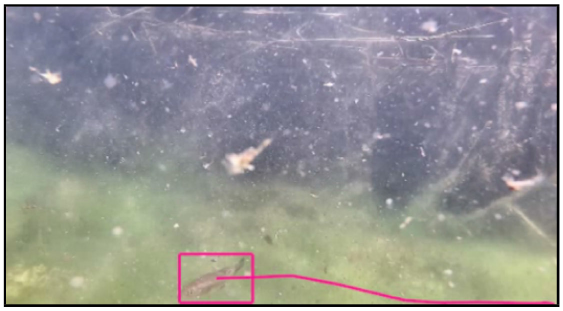


Figure 10. CDFW Habitat Monitoring Project Example Imagery

Structure is an important component of fish habitat [2]. Juvenile fish often avoid predation by occupying structurally complex areas where predators cannot forage effectively [34] and in doing so increases advantageous foraging areas for juveniles [39]. To promote a balanced fishery and the longevity of juvenile fish, supplemental fish habitats were installed within selected reservoirs. From 2017 to 2019 670 units were placed of habitat structure at 35 locations. Two types of supplemental habitat were placed into the lake; tree limbs and spider blocks (constructed from irrigation tubing anchored via concrete block). Most habitat was placed between 4-9 m deep to benefit warm water fish species [27]. Habitat units were placed to create "communities" that increase localized productivity that contribute to maintaining the warmwater fisheries while the lake is in its reduced capacity. Although habitat improvements are common in fisheries management, little has been documented outside of electroshocking to determine species habitat preference. The fish assemblage within the lake consists of Florida-strain largemouth bass (*Micropterus salmoides floridanus*), bluegill (*Lepomis machrochirus*), redear sunfish (*Lepomis microlophus*), green sunfish (*Lepomis cyanellus*), black crappie (*Pomoxis nigromaculatus*), channel catfish (*Ictalurus punctatus*), blue catfish (*Ictalurus furcatus*), brown bullhead (*Ameiurus nebulosus*), common carp (*Cyprinus carpio*), golden shiner (*Notemigonus crysoleucas*), inland silverside (*Menidia beryllina*) threadfin shad (*Dorosoma petenense*), and seasonally planted rainbow trout (*Oncorhynchus mykiss*).

Monitoring was conducted utilizing four cameras simultaneously. One camera was placed at each location for an average of 70 minutes of filming per location. This was done once per month for a total of 4 months (February, March, April and June). At each location, two staff members would enter the water and swim down to the specified depth and location to place the camera mount with a clear view of the habitat. Once filming was completed, staff would retrieve the camera and mounts.

2.3.5 Modular Optical Underwater Survey System

The NOAA MOUSS dataset contains videos collected around the main Hawaiian islands for stock assessment of the Deep 7 Bottomfish species, the most culturally important and highly valued of the deep-water bottomfish species in Hawaii. The Deep 7 species include Ehu (squirrelfish snapper, *Etelis carbunculus*), Gindai (Brigham's snapper, *Pristipomoides zonatus*), Hapu'upu'u (Seale's grouper, *Hyporthodus quernus*), Kalekale (Von Siebold's snapper, *Pristipomoides sieboldii*), Lehi (silverjaw snapper, *Aphareus rutilans*), Onaga (longtail snapper, *Etelis coruscans*), and 'Ōpakapaka (pink snapper, *Pristipomoides filamentosus*).

The system (Figure 12) is rated to 500 m and can effectively identify fish at depths of up to 250 m in Hawaiian waters using only ambient light. Baited drop cameras were typically deployed for 15-minute periods. Collected imagery is mostly static, though a small amount of camera motion can be observed, due to the elevated camera's position combined with currents. Data has been collected across multiple years, though only a small subset from 2017 was annotated and used as a part of this release (Figure 13).

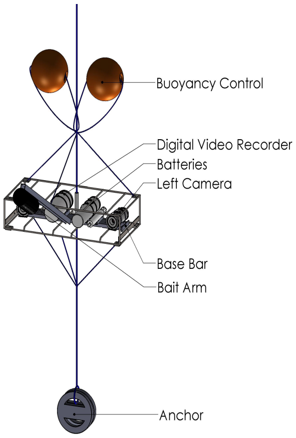


Figure 11. MOUSS Collection Platform

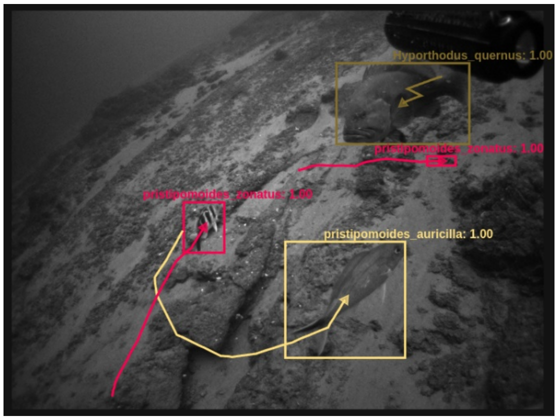

Hyporthodus_quernus: 1.00
pristipomoides_zonatus: 1.00
pristipomoides_zonatus: 1.00
pristipomoides_auricilla: 1.00

Figure 12. MOUSS Example Imagery

3. Baseline Algorithms

In training baseline models, we used a combination of open-source off-the-shelf deep networks available on GitHub [41], custom implementations of certain algorithms [32], and modified versions of publicly available methods with several additions [7, 10, 35, 38, 40]. These methods were then integrated into a single chained processing pipeline for end users [14], where individual components could be swapped out for evaluation while keeping other components the same, e.g. switching a detector or classifier but maintaining the same tracking algorithm.

For object detection, thus far we have trained only one detection architecture on the aggregate set, Cascade RCNN [7]. This model was trained in three separate variations, however, with each adaption present at both train and inference. The first (CRCNN), using the network architecture as defined in [7, 10], only with a custom train harness containing automatic learning rate adjustment and early stopping for when validation loss was not improving. This model was trained via resizing larger imagery to be at most 800x800 pixels when input to the first layer of network. A standard RGB input is used to train the detector, where greyscale imagery is simply duplicated to have the same values for each color channel in the case of black and white imagery. Secondly, we consider an adaptation (CRCNN-MS) that runs the network on multiple scales of the input, particularly the original image resized to a 800x800 maximum resolution, alongside the base image enlarged by a small margin (1.25x, via linear interpolation). The larger image is tiled into smaller chips (again 800x800 maximum, to conserve GPU memory), only without any downsampling, with the goal of improving small object detection. At inference time, detections near tile boundaries are assigned lower probabilities than mid-tile entries (reduced probability by a factor of 10), a tile step length less than the maximum chip size is applied (600), and non-maximum suppression applied to reduce cross-scale and cross-tile duplicates. Lastly, we consider an initial (basic) method to integrate motion features into the network (CRCNN-MT) wherein two independent motion detectors are combined with a greyscale version of the input frame, and fed into the detector (visualized in Figure 13). In this case, the baseline motion image is derived from the intensity difference from a 2- and 4-second windowed average of frames. No hue shifting data augmentations are performed in this variant, as they would distort the disparate information available in each channel. Simultaneously, no tiling or windowing of the image was performed, though a larger input perspective field size was used in the base layer of the network (1200x800 max size). All object detectors were trained via merging all fish categories into a single category, and treating all non-fish categories as background.

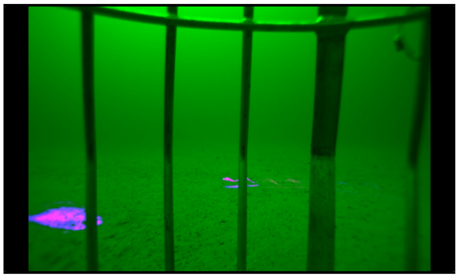


Figure 13. Composite image comprised of greyscale version of the input (green channel) alongside two different types of motion (red/blue) fed into training.

For tracking methods, we integrated three competing approaches into the system. The first, a variant of the method described in [32] (TUT or "Tracking the Untrackable"). The second, a multi-target tracking version of the SiamMask approach [25] (SM-MOT). Lastly, the more modern ByteTrack tracker [41] (BYTE). The TUT method combines three different sub-classifiers into a single score of whether or not a new detection belongs to an existing track. One for kinematic (motion) prediction, one for appearance, and one for local interactions. Each classifier and the resultant fusion model utilize structured recurrent neural networks on top of base features in order to incorporate temporal information into the model. In the case of appearance, features from a siamese network are used. For motion, raw image-plane positions. For interaction, a bin for the number of detections in an expanded 4x4 grid surrounding each detection. Our implementation adds a fourth feature type (bounding box consistency) and can also be run in either an offline or online fashion. A Hungarian matrix [24] combined with thresholding is used to make final linking decisions. Detections are utilized from one of the aforementioned integrated detectors. The MOT version of SiamMask similarly uses the same detections, but only for track initialization on new targets and the detectors are run at a lower frame rate than target output frame rate. New boxes which don't overlap significantly with existing tracks (IOU=0.5) and which exceed a fixed detection score (default=0.5) are used to initialize new instances of the single-object trackers for a particular object. A secondary pruning step compares overlaps on top of entire tracks to reduce duplicates. ByteTrack aims to associate all detections of tracks into tracks, regardless of threshold, and either builds on top of other tracking techniques or run as an independent tracker itself.

Final track-level re-classifiers offer the opportunity to filter either at the fish or species-level, allowing for noise suppression of entire tracks in order to boost system performance. For this purpose we utilized ResNext101 [40] (RESNEXT), EfficientNetV2 [35] (ENET2), track averaging of classification values along tracks for each class (RESNEXT-TA), and averaging plus the addition of separate high and low resolution classifier networks (RESNEXT-TA-LS). In the latter, if the extracted image chip around a fish (with a 1.2x scaling factor) was less than 7000 square pixels in the native image, it was considered low resolution. In all models performing species classification, inputs were converted from RGB to greyscale to reduce biases across sequences containing color and those without. Additionally hard negative mining of background samples were performed via running detectors on the datasets and utilizing high scoring detections which didn't overlap with annotations as a 'background' class. Other more simple methods of filtering include excluding tracks under a fixed length (default=3 states), or those with minimal image plane movement, i.e. objects that are more likely stationary background distractors.

4. Evaluation

No single metric is perfect for direct comparison between methods, though we initially focused on utilizing MCC (Matthew's Correleation Coefficient [11]) for classifiers, mAP (mean average precision) for detectors, and IDF1 [31] for trackers. MCC has been favored against alternative metrics such as F1 in cases where there are larger class imbalances, as it provides a more class-conscious comparison. With respect to IDF1, other metrics such as multi-object tracking accuracy (MOTA) skew slightly towards measuring detection performance instead of track continuity [3], the latter of which is often helpful to minimize to reduce annotator time correcting tracks. For trackers, the same detection approach was used in each method (CRCNN-MS) and evaluated on the test set. For classifiers, MCC represents performance of classifying detection states resulting from the default tracker model (CRCNN-MS + TUT), in the sense that temporal aspects of each classifier were used when available (e.g. averaging along the tracks) but classifiers were still evaluated at the per-frame level. A background category is introduced for detections which do not significantly overlap with any truth. This would imply a slight skew towards longer tracks having a greater affect on final values. In the case of classifiers, species-level metrics were computed, instead of just 1 general 'fish' category as in the case of measuring detection and tracking performance. Preliminary results are shown in Table 1, though additional metrics will be posted on our project page[1].

Detection Method	**mAP**
CRCNN	0.775
CRCNN-MS	0.812
CRCNN-MT	0.781

Tracking Method	**IDF1**
CRCNN-MS+TUT	56.9
CRCNN-MS+SM-MOT	49.1
CRCNN-MS+BYTE	66.3

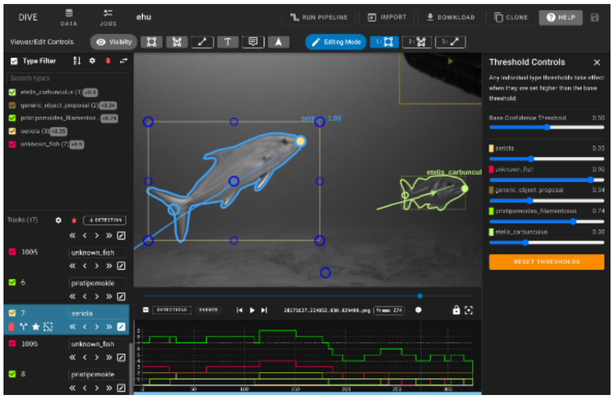
Classification Method	**MCC**
RESNEXT	0.86
RESNEXT-TA	0.90
RESNEXT-TA-LS	0.91
ENET2	0.87

Table 1. Preliminary Algorithm Comparisons

5. System and Annotation Tools

The FishTrack compilation is designed to be independent of the annotation tooling used to generate annotations. However, the baseline algorithms have been integrated into three interfaces within the open-source VIAME toolkit [14], including its latest user interface DIVE[3], shown in Figure 14. Within VIAME, new annotations can either be generated from scratch (with standard optimizations such as track interpolation, annotating at different frame rates, etc...), or the outputs of baseline models corrected. A majority of the annotations in this release were generated using this toolkit ($\sim$90%), most manually, though a portion via correcting and refining baseline algorithm outputs for expediency ($\sim$25%, but still with final manual verification/correction). Additional features of the desktop and web user interface include: the ability to train novel detection or classifier models, user-initialized object tracking using single target trackers [25] to speed up annotation (only retrained on fish data), automatic box to polygon converters trained for different problems [22, 23], stereo camera support, adaptive filtering based on species types, custom attribute assignment for secondary annotations such as occlusions, alongside other specializations. In addition to the data itself, an optional baseline scoring package is provided alongside FishTrack to encourage consistent comparisons across the test set, supporting generic metrics such as mAP, MOTA, IDF, and other track continuity scores.

[3] https://kitware.github.io/dive/



Threshold Controls

Any individual type thresholds take effect when they are set higher than the base threshold

Base Confidence Threshold 0.50

seriola 0.35

unknown_fish 0.90

generic_object_proposal 0.34

pristipomoides_filamentosus 0.74

etelis_carbunculus 0.30

RESET THRESHOLDS

Figure 14. DIVE Annotation Interface. The interface can be run either in a web server or from desktop installers. Outside of the core annotator, file managers and model training utilities allow for the training of new models by end-users.

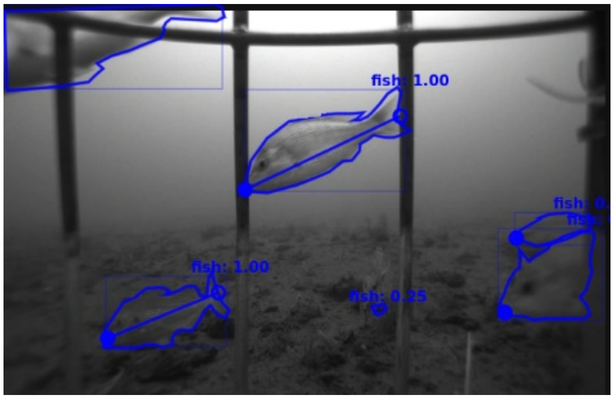


Figure 15. A near-term goal of future releases of FishTrack is to populate all tracks with polygons and head and tail positions, through a combination of automatic processes and manual correction.

6. Discussion and Conclusions

A few trends can be identified in the results presented in Section 4. For object detectors, our current motion model offered limited improvement in contrast to the multi-scale approach, indicating that either handling smaller objects was more important to improving performance than motion, or that we need to further improve how motion is integrated into the approach. There are a number of potential reasons for this, for instance the networks may be predispositioned to use standard image intensity the most, as they were seeded with models trained on the COCO [26] challenge. Additionally the loss of color information may have adversely affected the network. From the CNN architecture point of view, there are a number of alternative ways motion could be integrated, including a separate motion fork that's fused in later network layers instead of directly at the input [13], unsupervised pre-training over motion channels, 3D convolution in base layers of temporally stacked images, recurrent neural networks, or vision transformer-based methods [29].

Upon manual observation, many tracking errors are in the form of track switches in clusters of fish, and breaks across partial occlusions indicating that initial work needs to performed for appearance tracking in the domain, and occlusion aware re-identification descriptors. While the BYTE method outperformed TUT, it still had a number of disparate track breaks, indicating that the kalman-based filter used within the approach might not always be ideal for sporadic low speed fish maneuvers. Compared to more widespread tracking challenges, such as MOT17 and MOT20 [15], our tracking scores were lower, as the difficulty of the problem might suggest. There are a number of additional avenues for algorithmic improvements now that the dataset and baseline models have been generated. Some examples include enlarging the number of tracking algorithms we are comparing [16, 19, 21, 28, 37], improved temporal handling in classifiers besides naive averaging along a track, and refining the kinematic components of the tracking algorithms. Lastly, there are a couple of paths forward for improving the utility of the existing datasets. Adding pixel-level masks alongside head/tail markers will increase the number of algorithms that can be evaluated and trained against the dataset, and support related applications such as automated measurement across tracks (Figure 15). Due to the size of the dataset, this process would likely occur via automatic conversion [22] followed by manual review and correction, in order to minimize annotator effort.

It is our hope that this compilation can spur the development and evaluation of detection, classification, and tracking algorithms. Additionally, we hope it encourages and bootstraps the annotation of related data in novel locations. With the incorporation of additional data, new releases can be made annually or bi-annually, and used to support both future coding challenges and model improvements.

Acknowledgement

We would like to especially thank Alexa Abangan for the DropCam component (who is not listed as co-author due to procedural reasons, but may be considered as such), alongside Chris Demarke and Jeremy Taylor (for MOUSS annotations). Part of this paper was funded by the NOAA Fisheries Strategic Initiative on Automated Image Analysis (DG133W10CQ0040) and by the European Maritime and Fisheries Fund and France Filière Pêche (PFEA390021FA1000002, Ifremer Game of Trawls S2 project). The findings and conclusions in the paper are those of the authors and do not necessarily represent the views of NOAA, Ifremer, or the governments of the USA or France.

References

[1] Magnuson-stevens fishery conservation and management act, 1996. Public Law, 94:265. 1

[2] Robert D Barwick, Thomas J Kwak, Richard L Noble, and D Hugh Barwick. Fish populations associated with habitat-modified piers and natural woody debris in piedmont carolina reservoirs. *North American Journal of Fisheries Management*, 24(4):1120–1133, 2004. 5

[3] Keni Bernardin and Rainer Stiefelhagen. Evaluating multiple object tracking performance: the clear mot metrics. *EURASIP Journal on Image and Video Processing*, 2008:1–10, 2008. 7

[4] Daniel Bolya, Chong Zhou, Fanyi Xiao, and Yong Jae Lee. Yolact: Real-time instance segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9157–9166, 2019. 1

[5] Océane Boulais, Simegnew Yihunie Alaba, John E Ball, Matthew Campbell, Ahmed Tashfin Iftekhar, Robert Moorehead, James Primrose, Jack Prior, Farron Wallace, Henry Yu, et al. Seamapd21: A large-scale reef fish dataset for fine-grained categorization. In *Proceedings of the FGVC8: The Eight Workshop on Fine-Grained Visual Categorization, Online*, volume 25, 2021. 1

[6] Levi Cai, Nathan E McGuire, Roger Hanlon, T Aran Mooney, and Yogesh Girdhar. Semi-supervised visual tracking of marine animals using autonomous underwater vehicles. *International Journal of Computer Vision*, 131(6):1406–1427, 2023. 2

[7] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6154–6162, 2018. 6

[8] Matthew D Campbell, Adam G Pollack, Christopher T Gledhill, Theodore S Switzer, and Douglas A DeVries. Comparison of relative abundance indices calculated from two methods of generating video count data. *Fisheries Research*, 170:125–133, 2015. 1

[9] Mike Cappo, E Harvey, and M Shortis. Counting and measuring fish with baited video techniques-an overview. In *Australian Society for Fish Biology Workshop Proceedings*, volume 1, pages 101–114. Australian Society for fish biology Tasmania, 2006. 1

[10] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, et al. Mmdetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019. 6

[11] Davide Chicco and Giuseppe Jurman. The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC genomics*, 21(1):1–13, 2020. 7

[12] Gioele Ciaparrone, Francisco Luque Sánchez, Siham Tabik, Luigi Troiano, Roberto Tagliaferri, and Francisco Herrera. Deep learning in video multi-object tracking: A survey. *Neurocomputing*, 381:61–88, 2020. 1

[13] Matthew Dawkins, Jon Crall, Matt Leotta, Tasha O'Hara, and Liese Siemann. Towards depth fusion into object detectors for improved benthic species classification. In *International Conference on Pattern Recognition*, pages 415–429. Springer, 2022. 8

[14] Matthew Dawkins, Linus Sherrill, Keith Fieldhouse, Anthony Hoogs, Benjamin Richards, David Zhang, Lakshman Prasad, Kresimir Williams, Nathan Lauffenburger, and Gaoang Wang. An open-source platform for underwater image and video analytics. In *2017 IEEE winter conference on applications of computer vision (WACV)*, pages 898–906. IEEE, 2017. 6, 7

[15] Patrick Dendorfer, Hamid Rezatofighi, Anton Milan, Javen Shi, Daniel Cremers, Ian Reid, Stefan Roth, Konrad Schindler, and Laura Leal-Taixé. Mot20: A benchmark for multi object tracking in crowded scenes. *arXiv preprint arXiv:2003.09003*, 2020. 8

[16] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6824–6835, 2021. 8

[17] Robert B Fisher, Yun-Heh Chen-Burger, Daniela Giordano, Lynda Hardman, Fang-Pang Lin, et al. *Fish4Knowledge: collecting and analyzing massive coral reef fish video data*, volume 104. Springer, 2016. 1

[18] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 1

[19] Ahsan Jalal, Ahmad Salman, Ajmal Mian, Mark Shortis, and Faisal Shafait. Fish detection and species classification in underwater environments using deep learning with temporal information. *Ecological Informatics*, 57:101088, 2020. 8

[20] Alexis Joly, Hervé Goëau, Hervé Glotin, Concetto Spampinato, Pierre Bonnet, Willem-Pier Vellinga, Robert Planqué, Andreas Rauber, Simone Palazzo, Bob Fisher, et al. Lifeclef 2015: multimedia life species identification challenges. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 6th International Conference of the CLEF Association, CLEF'15, Toulouse, France, September 8-11, 2015, Proceedings 6*, pages 462–483. Springer, 2015. 1

[21] Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. Transformers in vision: A survey. *ACM computing surveys (CSUR)*, 54(10s):1–41, 2022. 8

[22] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 7, 8

[23] Anton S Kornilov and Ilia V Safonov. An overview of watershed algorithm implementations in open source libraries. *Journal of Imaging*, 4(10):123, 2018. 7

[24] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955. 6

[25] Bo Li, Wei Wu, Qiang Wang, Fangyi Zhang, Junliang Xing, and Junjie Yan. Siamrpn++: Evolution of siamese visual tracking with very deep networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4282–4291, 2019. 6, 7

[26] 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: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 3, 8

[27] WE Lynch Jr and DL Johnson. Angler success and bluegill and white crappie use of a large structured area. ohio cooperative fish and wildlife research unit, school of natural resources, ohio state university. appendix v of evaluation of fish management techniques. final report. Technical report, Project F-57-R Study 10. Columbus, OH, 1988. 5

[28] Gerard Maggiolino, Adnan Ahmad, Jinkun Cao, and Kris Kitani. Deep oc-sort: Multi-pedestrian tracking by adaptive re-identification. *arXiv preprint arXiv:2302.11813*, 2023. 8

[29] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12179–12188, 2021. 8

[30] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015. 1

[31] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *European conference on computer vision*, pages 17–35. Springer, 2016. 7

[32] Amir Sadeghian, Alexandre Alahi, and Silvio Savarese. Tracking the untrackable: Learning to track multiple cues with long-term dependencies. In *Proceedings of the IEEE international conference on computer vision*, pages 300–311, 2017. 6

[33] 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. 1, 2

[34] Jacqueline F Savino and Roy A Stein. Predator-prey interaction between largemouth bass and bluegills as influenced by simulated, submersed vegetation. *Transactions of the American Fisheries Society*, 111(3):255–266, 1982. 5

[35] Mingxing Tan and Quoc Le. Efficientnetv2: Smaller models and faster training. In *International conference on machine learning*, pages 10096–10106. PMLR, 2021. 6, 7

[36] Fasheng Wang, Ping Cao, Fu Li, Xing Wang, Bing He, and Fuming Sun. Watb: Wild animal tracking benchmark. *International Journal of Computer Vision*, 131(4):899–917, 2023. 2

[37] He Wang, Song Zhang, Shili Zhao, Qi Wang, Daoliang Li, and Ran Zhao. Real-time detection and tracking of fish abnormal behavior based on improved yolov5 and siamrpn++. *Computers and Electronics in Agriculture*, 192:106512, 2022. 8

[38] Qiang Wang, Li Zhang, Luca Bertinetto, Weiming Hu, and Philip HS Torr. Fast online object tracking and segmentation: A unifying approach. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pages 1328–1338, 2019. 6

[39] Earl E Werner, James F Gilliam, Donald J Hall, and Gary G Mittelbach. An experimental test of the effects of predation risk on habitat use in fish. *Ecology*, 64(6):1540–1548, 1983. 5

[40] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017. 6, 7

[41] Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Fucheng Weng, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. Bytetrack: Multi-object tracking by associating every detection box. In *European Conference on Computer Vision*, pages 1–21. Springer, 2022. 6