International Journal of Computer Vision (2023) 131:1406–1427
https://doi.org/10.1007/s11263-023-01762-5

Semi-supervised Visual Tracking of Marine Animals Using Autonomous Underwater Vehicles

Levi Cai[1] · Nathan E. McGuire[2] · Roger Hanlon[3] · T. Aran Mooney[4] · Yogesh Girdhar[2]

Received: 29 April 2022 / Accepted: 2 January 2023 / Published online: 1 March 2023

Abstract

In-situ visual observations of marine organisms is crucial to developing behavioural understandings and their relations to their surrounding ecosystem. Typically, these observations are collected via divers, tags, and remotely-operated or human-piloted vehicles. Recently, however, autonomous underwater vehicles equipped with cameras and embedded computers with GPU capabilities are being developed for a variety of applications, and in particular, can be used to supplement these existing data collection mechanisms where human operation or tags are more difficult. Existing approaches have focused on using fully-supervised tracking methods, but labelled data for many underwater species are severely lacking. Semi-supervised trackers may offer alternative tracking solutions because they require less data than fully-supervised counterparts. However, because there are not existing realistic underwater tracking datasets, the performance of semi-supervised tracking algorithms in the marine domain is not well understood. To better evaluate their performance and utility, in this paper we provide (1) a novel dataset specific to marine animals located at http://warp.whoi.edu/vmat/, (2) an evaluation of state-of-the-art semi-supervised algorithms in the context of underwater animal tracking, and (3) an evaluation of real-world performance through demonstrations using a semi-supervised algorithm on-board an autonomous underwater vehicle to track marine animals in the wild.

Keywords Semi-supervised learning · Visual tracking · Marine animal tracking · Autonomous underwater vehicles

Communicated by Silvia Zuffi.

✉ Levi Cai
cail@mit.edu

Nathan E. McGuire
nmcguire@whoi.edu

Roger Hanlon
rhanlon@mbl.edu

T. Aran Mooney
amooney@whoi.edu

Yogesh Girdhar
ygirdhar@whoi.edu

[1] Massachusetts Institute of Technology and Woods Hole Oceanographic Institution Joint Program, Woods Hole 02543, MA, USA

[2] Applied Ocean Physics and Engineering Department, Woods Hole Oceanographic Institution, Woods Hole 02543, MA, USA

[3] Marine Biological Laboratory, Woods Hole 02543, MA, USA

[4] Biology Department, Woods Hole Oceanographic Institution, Woods Hole 02543, MA, USA

1 Introduction

This work proposes the use of *semi-supervised visual tracking* (SST) algorithms on *autonomous underwater vehicles (AUV)* to track marine animals with no prior training. Semi-supervised algorithms have shown remarkable success in generic tracking benchmarks, but these benchmarks do not provide sufficient evidence of their performance in the underwater domain. This work aims to first establish the effectiveness of semi-supervised trackers in marine tracking tasks through domain-specific benchmarking, and in addition, demonstrate the use of a semi-supervised tracker in the real-world. Our contributions are thus to provide (1) a unique and underwater-specific dataset consisting of videos of mobile marine animals in their natural environment taken by following them with a moving camera system, (2) an evaluation of current state of the art semi-supervised tracking algorithms on this dataset using metrics relevant to the problem of marine animal tracking, and (3) a novel robotic demonstration of using a semi-supervised tracker in the real-

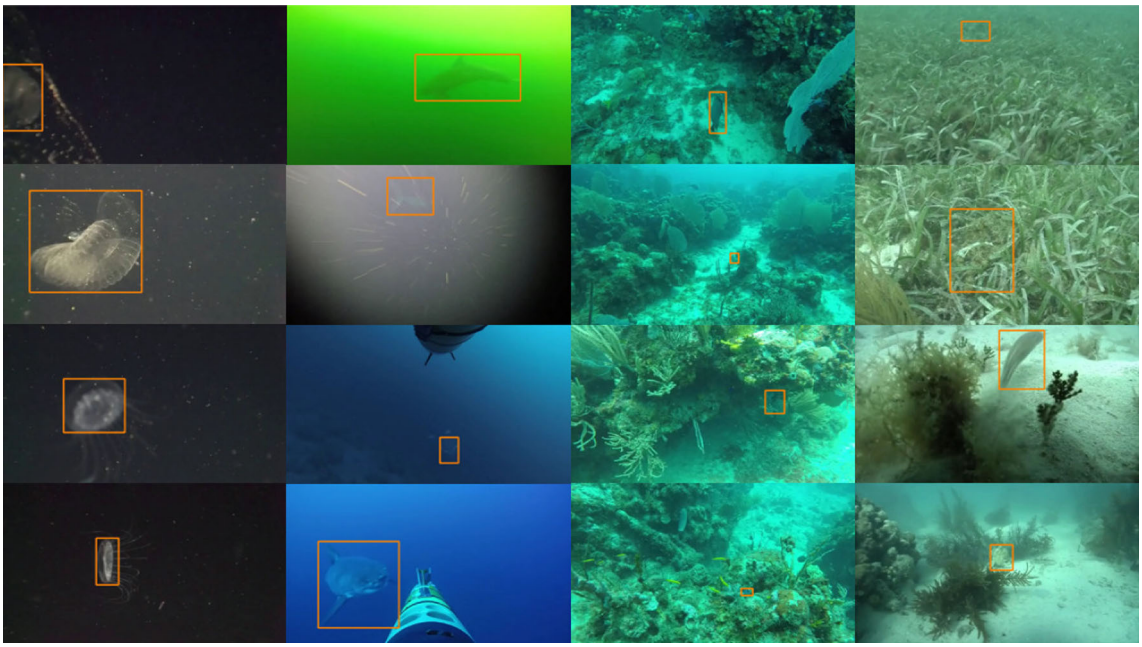

Fig. 1 We present an initial dataset for evaluating performance of visual semi-supervised trackers for tracking marine animals in the wild. This dataset attempts to capture difficulties (and sometimes benefits) in tracking across species, environments, and behaviors. It consists of 33 densely-labelled sequences, collected by both divers and AUVs in real marine animal tracking scenarios, averaging over 1-min in length (Color figure online)

world by deploying it on an AUV to track a marine animal in the wild.

1.1 Visual Tracking of Marine Animals

In-situ visual observations of marine organisms can provide valuable insights into their biology that can be difficult to discern using other modes of observations. These observations can be used to characterize details of animal behavior and their interactions within that ecosystem. However, gathering in-situ observations using current approaches, especially in the case of marine animals, can be expensive, time-restrictive due to depth effects, and often dangerous. To achieve these observations, marine biologists have long relied on diver-based operations (Hanlon et al., 1999; Hanlon & McManus, 2020), tags (Kukulya et al., 2015; Mooney, 2020), and occasionally using human-occupied vehicles (HOVs) (Priede et al., 2020) or remotely operated vehicles (ROVs) (Katija et al., 2021) to gather visual observations. Each of these approaches is uniquely suited depending on the animal, environment, costs, hazards, and equipment that are present.

Quantifying behavior requires long sequences of behavioral interactions and this becomes increasingly difficult with mobile animals. One of the most productive methods of measuring behavior is known as "focal animal sampling" in which video is acquired in a very disciplined manner by continually filming either a single animal, or pairs of animals for long periods to enable quantification and statistical analyses (Bateson & Martin, 2021). The reason for this is that key behaviors are not predictable and thus large video data sets over long continuous periods are required to capture both ongoing and episodic events. Divers become exhausted and cold with this demanding method, and the animal tracking can lead them to deeper water or beyond safe retreat to the surface vessel. Moreover, the bubbles from SCUBA and the changing shape of the diver can bias the target species' behavior, thus negatively influencing the natural interactions of the target species. Manually controlled ROVs can assist to a degree but they are prone to being pulled by currents that push the tether and visually lose track of the target species. Moreover, ROVs often cannot be deployed close to the seafloor in complex coral reef like environments which can pose significant danger to both the robot (entanglement) and the reef. ROVs have been adapted to record static benthic animals with a regimented sampling routine (Williams et al., 2009) but the ability to follow mobile animals is far more challenging. AUVs have the potential to follow and record mobile animals without the tether problem.

Autonomous underwater vehicles (AUVs) equipped with cameras and higher powered computers have potential to greatly supplement visual observations of animal behaviors in more difficult to access regions of the ocean, or to provide longer term observations in cases with limited human support. However, fully autonomous AUVs can currently only

track a limited set of organisms in a small set of environments. Currently, trackable animals (i) are either tagged (Kukulya et al., 2015), meaning a device must be physically attached to them that emits an acoustic signal to localize against, (ii) are found in relatively simple visual situations (such as the deep midwater column) (Yoerger et al., 2021), or (iii) have significant amounts of labelled visual data available (Katija et al., 2021). These can be limiting in many circumstances because tags are difficult to install, most marine animals live in visually complex environments ranging from coral reefs to hydrothermal vents, and the underwater community generally has limited labelled data available for many species of interest.

1.2 Semi-supervised Trackers

Semi-supervised trackers have been proposed as alternatives to fully-supervised tracking methods because they require *no target-specific pretraining*, are able to *run in real-time*, and have shown high accuracy on a number of difficult visual object tracking benchmarks. Semi-supervised tracking algorithms, at run-time, only require an initial bounding box of the target to be provided by a human, and then the algorithm autonomously localizes the target in all subsequent frames. Their performance on existing benchmarks suggests they may be useful to the AUV community, by providing an alternative strategy to enable mostly autonomous visual data gathering without the need of significant visual training data or additional tracking equipment. This means AUVs could track a more diverse set of animals and in a wider range of environments. However, these methods have only been evaluated on generic datasets, which have few realistic examples of tracking animals fully underwater, so their effectiveness in these environments is not well known.

1.3 Challenges Unique to the Underwater Domain

Underwater environments have many unique visual characteristics that are under-represented in generic datasets that are used to evaluate semi-supervised trackers. Some examples include: depth and distance dependent color absorption, presence of marine snow, extreme deformations and body shapes of marine animals, camouflaging, and large variations in visual complexity of underwater habitats forming the background. Hence a dedicated evaluation of their effectiveness specifically in realistic underwater settings is necessary. For instance, while the largest existing generic dataset, LaSOT (Fan et al., 2020), contains many examples of underwater animals, these sequences are primarily taken above-water or in aquariums with idealized lighting. Although there are a few underwater tracking sequences, they contain mostly turtles in relatively clear and simple cases.

The most immediate visual distinctions are caused by water as a medium itself, which causes light distortion and absorption. In the former, following Snell's law, light traveling between media of varying densities causes light to bend, this is prevalent in cases where a camera is placed inside a waterproof housing, and so light must pass from water, to glass or plastic, and then air before reaching the camera lens. In addition, water absorbs different frequencies of light at different rates (Akkaynak & Treibitz, 2019), red light being the most easily absorbed and blue light the last to be absorbed. This causes significant loss of color-based information in environmentally-lit situations, especially deeper underwater. Underwater caustics near the surface, where light refracts and reflects underwater causing bright and constantly moving patterns, can also serve as significant distractors.

In deeper waters or at night, where environmental light is less of a concern, active lighting from a diver or vehicle can simplify the color absorption problem. Unfortunately, such relatively bright lights override the dark adaptation of the target animals night vision, and thus affect its behavior. In addition, again due to light absorption, visible distances are typically short. Furthermore, marine snow, small biological particulate matter, is prevalent in the ocean (Wang et al., 2021), and active light reflects off of these particles. This can cause either general haziness or become sharp and ubiquitous visual features.

The types of habitats and the animals themselves are also unique and present significant challenges. Habitats such as coral reefs, mid-water, hydrothermal vents, kelp forests, sea grass, etc. provide various visual challenges, several of which are shown in Sect. 3. Marine organisms exhibit a variety of different swimming, foraging and defensive behaviors that can disrupt the capabilities of imaging to follow them. Camouflage by many marine animals is highly advanced and diverse. An extreme example among invertebrates are octopuses, cuttlefish and squid which can instantly change their camouflage pattern, and even their body shape as well as their skin 3D texture, all of which is exceptionally difficult to discern with digital imagery. These animals as well as many fishes have specific behaviors to actively avoid detection and tracking, through camouflage, inking, darting, hiding, and so forth.

Finally, collection of raw data can be especially difficult underwater, and so too is the development of large databases of imagery. Although extensive datasets exist from the underwater domain, they are either aiming a stationary camera to a stationary subject, or aiming at characterizing fish biodiversity with wide angle lenses and thus not focusing on single target tracking.

Depending on depth and distance from shore or the hazards of the environment or animals themselves, significant equipment, extensive training, and labor could be required. Even in relatively shallow water, below 10 m, longer-term

tracking of marine animals requires a SCUBA certification and specific camera housings. If the animal varies its depth quickly, this can be dangerous to the diver. In other instances, only rare vehicles are capable of collecting data, for instance in the deep sea.

In this paper, we review related works in Sect. 2, present the dataset and evaluation results in Sect. 3, present the real-world AUV tracking results in Sect. 5, and discuss the results and their implications in Sect. 6 and finally give concluding remarks in Sect. 7.

2 Related Work

This work is most closely related to works in autonomous vision-based tracking for marine animals and the development of datasets for evaluating real-time, semi-supervised tracking methods. In the following section we discuss how this work is situated in these contexts.

2.1 Autonomous Vision-Based Marine Animal Tracking

We focus on the current use of AUVs for marine animal tracking. We first discuss *passive* video tracking of animals. In passive contexts, video cameras collect data of marine animals, but the *images themselves* are not used to inform AUVs where to look. These strategies tend to fall in two categories: surveys and acoustic tag-based tracking. In the former, vehicles such as the MBARI i2MAP Dorado or the Seabed AUV (Williams et al., 2009) perform pre-programmed surveys in ocean, collecting video along the pre-determined track, which is analysed afterwards. Other passive visual observation gathering by AUVs is accomplished through acoustic tags. Animals such as sharks, cetaceans, penguins, turtles, some larger fish, etc. can be outfitted with an acoustic tag that an AUV can then use to localize its position relative to the vehicle. These AUVs, such as the REMUS series, are then equipped with cameras, often many oriented in several directions, that record video to a memory card and again analysed after the mission (Kukulya et al., 2016). Tags are only able to be attached to specific species of animals, and typically require significant deployment effort and training to affix. In addition, videos collected in this manner have less quality guarantee, and animals may not stay in the frame of a single camera for very long.

AUVs equipped with *active* visual tracking capabilities have recently been developed. The Mesobot platform (Yoerger et al., 2021) is developed to track slow-moving animals in the mesopalegic zone (300–1000 m depths), also known as the Ocean Twilight Zone, which has very little light. This system consists of a grey-scale, stereo-camera system for tracking and provides its own light. Animals are tracked through established color segmentation and blob-tracking methods, and the Mesobot was used to successfully track jellyfish and larvaceans in-situ for several hours. However, these methods only work in very simple tracking scenarios, which is unique to the mesopalegic zone and the types of animals that live there, where jellyfish illuminated by on-board lights are easily distinguished from the dark ocean backdrop. Katija et al. (2021) introduced the use of tracking-by-detection and deep learning strategies to increase robustness and track in more complicated scenarios and demonstrated its effectiveness on-board the MBARI MiniROV. In this case, a deep convolutional neural network, RetinaNet (Lin et al., 2017) is pre-trained on the target(s) of interest. During run-time, the network is used to detect targets, and a data association strategy is used to determine which detections correspond to the appropriate target. In both of these scenarios, once the target is localized in each frame, the vehicle is commanded to update its position to center the target in the camera frame.

These systems have provided invaluable insights for researchers, however, Mesobot is only able to track simple organisms in the deep sea and the MBARI MiniROV can only track animals for which they have significant training data already collected. Our work aims to show if semi-supervised trackers can enable AUVs to track a much larger range of animals and in more varied habitats where significant target-specific training data for fully supervised neural networks is not available.

2.2 Underwater Visual Datasets for Marine Animal Tracking

Due to the popularity of machine learning methods for land-based animal classification, several datasets have recently been developed for marine animal classification as well. Most prominent among these are FathomNet, VIAME, AIMs Ozfish, MBARI VARS, and DeepFish (Katija et al., 2022; Dawkins et al., 2017; Schlining & Stout, 2006; OzFish Dataset, 2022; Saleh et al., 2020). All of these datasets are useful for training deep learning-based networks for classification and can be incorporated into fully-supervised trackers. However, each is highly localized to a specific set of animals. For instance, AIMs Ozfish and DeepFish both are collected from nearby Australia, and primarily contain images of vertebrate fish. VIAME is a more general underwater animal dataset, but mostly contains imagery from smaller organisms that visit baited camera traps. In addition, these types of sequences are insufficient for evaluating tracking scenarios with moving cameras, or for organisms in the open ocean. FathomNet and the MBARI VARS datasets are perhaps the only datasets that explicitly attempt to capture data useful for evaluating active tracking tasks. However, the MBARI VARS dataset is difficult to access publicly, and FathomNet's dataset, at the time of this writing, only includes deep

ocean species such as jellyfish, larvaceans, and slow-moving seafloor organisms found off the coast of Monterey Bay in California.

None of these datasets is large enough to provide robust training for fully-supervised tracking methods to work on the full range of marine animals of interest, such as those illustrated in Table 2, though they are invaluable starts. Furthermore, none provide longer video sequences that allows evaluation of semi-supervised trackers on realistic tracking tasks where there is significant camera motion, high-frame rates of at least 10 fps, which are necessary to enable reasonable feedback control in AUVs, and a realistic set of behaviors where baited traps are not used.

2.3 Deep Learning Approaches for Semi-supervised Tracking

While semi-supervised visual tracking has a fairly long literature, only recently has it started to gain more widespread attention with the introduction of deep learning-based feature extractors and classifiers. One of the first deep learning-based SSTs, MDNet (Nam & Han, 2016), achieved the highest performance on an early generic object tracking benchmark VOT (Kristan et al., 2013). However, MDNet could not run in real-time. Since then, several innovations, especially in deep learning-based trackers, have enabled SSTs to achieve real-time tracking speeds while continuing to consistently be top performers in accuracy on multiple generic object tracking benchmarks.

Of these innovations, two architectures, which we will refer to as either *Siamese-based* and *online discriminator-based*, in particular have dominated the most recent benchmarks. Tao et al. (2016) first introduced Siamese-based networks as a way to achieve the accuracy of deep learning-based trackers with real-time speed. In these architectures, a pre-trained deep neural network backbone is used to extract features from an initial template image of the target. Then using the backbone network, features are extracted from subsequent images, and the region (in the extracted feature space) that is most similar to that of the features of the template image, is labelled as the target of interest in that frame. Subsequent Siamese-based trackers innovated on the types of backbone networks used for feature extraction or downstream optimization tasks such as providing full masks or selecting between distractors (Li et al., 2019; Wang et al., 2018; Li et al., 2019).

Siamese-based networks suffer in performance however during object appearance changes, since they rely only on the appearance of the object in the first frame for reference. Later, Danelljan et al. (2019) introduced an online learning component in ATOM, where recent frames, in addition to the first frame, are used to train a discriminator that is then used to classify subsections of the current tracking frame as the target or not. This further improved performance in the case of appearance changes, and many later trackers added post-processing steps to gain additional performance boosts (Bhat et al., 2019; Danelljan et al., 2015, 2017; Chen et al., 2021; Wang et al., 2021).

It is interesting to note, as described in Wang et al. (2021), that Siamese-based networks are effectively special cases of online discriminator-based architectures, where the online learning step is removed, which may be important in how they are analysed. In particular, Siamese-based networks are generally robust in the long-term, meaning if the target appearance returns to something similar to the first frame, the tracker can recover. In contrast, in the online discriminator case, because later frames are not labelled, small errors in appearance-learning accumulate, and can cause the tracker to drift from focusing on the correct features resulting in lower long-term robustness (Mueller et al., 2016). However, when this tradeoff is carefully addressed, the online visual tracking problem seems to tend towards online learning approaches, because the best predictor of an object's appearance at time t is its appearance at $t-1$.

In some of these cases, trackers can catastrophically fail either from appearance change or from self-drift, even when the object remains in the sequence at all times. We thus believe it is important for evaluation datasets to contain longer (in duration) sequences, where object appearances change over time, in order to verify these issues.

2.4 Generic Benchmarks for Semi-supervised Tracking

Many general purpose semi-supervised tracking datasets and benchmarks exist, such as LaSOT, GOT10K, VOT2020, DAVIS, OxUvA, OTB100, TrackingNet, Youtube-BB, NfS, and UAV123 (Fan et al., 2020; Huang et al., 2021; Kristan et al., 2020; Wu et al., 2015; Caelles et al., 2019; Xu et al., 2018; Mueller et al., 2016; Galoogahi et al., 2017; Valmadre et al., 2018). However, as noted in LaSOT, many of these datasets such as OxUvA, TrackingNet, and Youtube-BB, are only labelled intermittently, so many frames are not guaranteed to have a human-checked ground truth. This is not well-suited for evaluating trackers that need to run on vehicles where high-frame rates are required. As in LaSOT, we thus focus on comparing against the following densely labelled datasets, also noted in Table 1.

OTB100 (Wu et al., 2015), *VOT2020* (Kristan et al., 2020), *Need for Speed (NfS)* (Galoogahi et al., 2017) are smaller standard datasets, OTB and VOT are some of the earliest benchmarks to have been introduced, with VOT2020 being a slightly updated version of VOT, using a different bounding box methodology and swapping out a few sequences. NfS was developed for evaluating extremely fast frame-rate systems, with frames at 240 fps. However, in all 3 of these

Table 1 Comparison of densely-labelled and high-fps benchmark datasets for evaluating semi-supervised trackers

	OTB100	VOT2020ST	NfS	GOT-10K	LaSOT	UAV123	UAV20L	VMAT (Ours)
Num sequences	100	60	100	9695	1550	123	20	33
Frame rate	30	30	240	10	30	30	30	30
Min duration (s)	2.4	1.4	0.7	2.9	33.3	3.6	57.2	14.6
Mean duration (s)	19.7	11.1	16.0	14.9	83.4	30.5	97.8	74.9
Max duration (s)	129.1	50.0	86.1	141.8	379.9	102.8	184.2	248.2
Fully underwater animal sequences	No	Only 1	Only 1	Yes	Only 1 species	No	No	Yes
Realistic tracking sequences	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes

Here we consider fully underwater animal sequences where both the camera and animal are underwater in nature, and not in an aquarium or similarly artificial setting. Here we consider realistic tracking sequences those that are aimed at longer duration, densely-labelled, and >10 fps (for vehicle control purposes) sequences

datasets combined, only 2 sequences are of underwater animals at all, both in fairly ideal lighting conditions.

GOT-10K (Huang et al., 2021) was developed to have by far the most sequences of other current datasets, to be used for both training and evaluation purposes. Some of these sequences also include underwater animal tracking. However, these sequences are very short, on average only 14.9 s in duration across the whole dataset, or only 9.5 s in duration when considering animal classes. These are too short to effectively evaluate performance for longer-term deployments on real vehicles.

UAV123 and UAV20L (Mueller et al., 2016) are part of the same dataset, developed to evaluate tracking of humans and vehicles from on-board unmanned aerial vehicles (UAVs). UAV20L consists of 20 sequences that are longer in length than the rest of the UAV123 sequences, and is most like our dataset in terms of its overarching goals. However, this dataset is only focused on humans and ground vehicles.

LaSOT (Fan et al., 2020) contains 1550 sequences that are densely labelled, at 30 fps, and capture many realistic tracking scenarios, and is likely most useful for evaluations similar to ours. However, among those sequences, while many are taken of underwater situations, most occur in artificial environments such as aquariums, or are above the surface of the water. We found only 12 sequences that were consistent with our goals, but do not have species diversity, as 11 are of turtles and 1 of an alligator.

For instance, LaSOT focused on longer sequences, GOT10K increased number of classes, OxUvA provided significant numbers of fully occluded or fully lost targets, etc. Our approach most resembles UAV20L, which aims to provide a more domain-specific dataset with longer tracking sequences that are realistic to the overall tracking problem on a mobile platform.

While LaSOT and UAV20L closely resemble our tracking evaluation goals, they do not have sufficient coverage of the underwater tracking domain to provide convincing evaluations of the range of underwater-specific issues.

3 The Visual Marine Animal Tracking Dataset

3.1 Design Principles

In order to evaluate expected performance of semi-supervised trackers in a variety of challenging marine tracking tasks, we needed to develop a new evaluation dataset. Our dataset, which we refer to as the Visual Marine Animal Tracking (VMAT) dataset, is used for evaluation only, and is domain-specific, similar to the UAV20L dataset (Mueller et al., 2016) in terms of size and scope, but targeted at AUV deployments for tracking underwater animal targets. We thus focus on collecting data that is diverse across animals, habitats, behaviors, and tracking scenarios as possible, which we show in Table 2. We also aim for each sequence to be as *realistic* as possible, where a camera is actively tracking an animal, sequences are densely-labelled, where every frame is labelled, and are in as long duration as possible (without loss of sight). We ensure frame rates are over 10 fps which we believe is the minimum frame rate necessary to perform reliable feedback control of most underwater vehicles, though higher is better.

3.2 Data Collection, Processing, and Annotation

To build a diverse dataset, we gathered sequences from several different existing scientific sources, in addition to collecting our own sequences in the field. All sequences in the dataset were collected in active tracking scenario, where a single target is the focus for the entirety of the track. From each source, we selected the longest contiguous sequences that did not have full occlusions or loss-of-sight.

We began by collecting existing videos from nearby scientists. These consisted of difficult sequences of octopuses exhibiting a variety of behaviors (from highly conspicuous to effectively camouflaged) in seagrass, coral, and sandy habitats; these were filmed via SCUBA in Puerto Rico by co-author Roger Hanlon using a Panasonic HD HVX200.

Table 2 List of sequences in the dataset and the corresponding sequence ID, animal, habitat, and behavior contained within the sequence

ID	Animal	Habitat	Motion/Behavior	# Frames
0	Octopus	Seabed, sand and grass	Stop and go, fast	1340
1	Octopus	Seabed, sand and grass	Slow crawl	1129
2	Octopus	Seabed, dense seagrass	Slow crawl	2971
3	Octopus	Seabed, dense seagrass	Stop and go, slow	2042
4	Shark	Midwater, shallow, clear	Constant swim	438
5	Shark	Seabed, near-bottom	Constant swim	1092
6	Shark	Midwater, deep, marine snow	Fast swim	750
7	Dolphin	Midwater, turbid	Fast swim	690
8	Larvacean	Midwater, deep, marine snow	Slow swim	601
9	Larvacean	Midwater, deep, marine snow	Slow swim	901
10	Jellyfish	Midwater, deep, marine snow	Slow swim	512
11	Jellyfish	Midwater, deep, marine snow	Slow swim	511
12	Striped Fish	Coral	Fast darting	750
13	Parrotfish	Coral	Medium swim	2610
14	Parrotfish	Coral	Medium swim	1080
15	Parrotfish	Coral	Medium swim	1680
16	Lionfish	Coral	Stationary	4770
17	Angelfish	Coral	Medium swim	3870
18	Boxfish	Coral	Medium swim	2580
19	Blue Tang	Coral	Fast swim	1350
20	Blue Tang	Coral	Fast swim	600
21	Squid	Rocky Seabed	Medium swim	2550
22	Squid	Rocky Seabed	Medium swim	5550
23	Octopus	Rocky Seabed	Crawling	2190
24	Snapper	Coral	Fast swim	3180
25	Shark	Coral	Medium swim	2490
26	Shark	Coral	Medium swim	5514
27	Shark	Coral	Medium swim	1566
28	Stingray	Seagrass	Medium swim	7448
29	Jack	Coral	Medium swim	3960
30	Barracuda	Coral	Fast swim	2363
31	Sea turtle	Midwater	Slow swim	3180
32	Jellyfish	Seagrass	Slow swim	1920

These sequences were obtained via a focal animal sampling routine for a study of adaptive camouflage.

To add midwater tracking scenarios with significantly larger and faster animals, Amy Kukulya and Roger Stokely, from the Woods Hole Oceanographic Institution, provided sequences of sharks and dolphins. These sequences were all collected by REMUS vehicles (Kukulya et al., 2015) tracking tagged animals. In these cases, even though the vehicles were actively tracking an acoustic beacon physically attached to the animals, they were only *passively* collecting visual data. This meant that the animals rarely stayed in frame, so the sequences are relatively short. These sequences also show variation in lighting at different depths, as shallower tend to be blue, deeper are black, and water that is high in chlorophyll tends to be very green. It is also important to note the simpler backgrounds in these midwater examples. The deeper of these sequences also show significant marine snow. These can be seen in the second column of Fig. 1.

We also then sought to add deeper ocean tracks. The Mesobot team, from their expeditions in the Pacific Ocean as published in Yoerger et al. (2021), were able to provide long tracks of midwater organisms in the ocean twilight zone (100–3000 m depths). These samples were collected fully autonomously by Mesobot via a standard blob tracking algorithm. While Mesobot sequences were longer in length, there do not include as much visual diversity, and so the included sequences are the ones that capture the most diverse motion instances.

Finally, to represent a larger range of species, the authors Levi Cai, Yogesh Girdhar, and Aran Mooney from the Woods Hole Oceanographic Institution, performed several SCUBA dive operations in the U.S. Virgin Islands, St. John in October 2021 and October 2022. We used GoPros in off-the-shelf underwater housings and followed a variety fish in coral reef, seagrass, and sandy environments. For each sequence, we track a single organism for as long as possible while minimizing occlusions. We found a new organism when it was no longer possible, and repeated this until our dive times were complete. Through these tracks, we collected 20 additional sequences across several species, including squid, an octopus, reef sharks, a stingray, a squid, and 9 types of reef fish.

All videos were then reviewed manually by the author and the longest continuous sequences, without loss of sight and with continuing novel viewpoints or appearance changes, of each organism were selected. Because many tracker evaluation metrics are sensitive to differences in both spatial and temporal resolution we standardize all videos to a minimum common standard that is still usable for active AUV tracking. Specifically, we subsample frame rates down to 30 fps and scale resolutions of all videos to 854×480 pixel (480p) using ffmpeg.

Finally, to produce dense ground-truth labels of each target, the authors manually labelled axis-aligned bounding boxes using LabelBox (Labelbox, 2022) with manual labelling (adding a keyframe) or explicit review no less than every 15 frames. LabelBox then linearly interpolates between labels, and the authors do a last verification by watching the whole sequences. All sequences were labelled by Levi Cai and Yogesh Girdhar, and reviewed by Levi Cai.

3.3 Dataset Attributes and Statistics for Evaluation

The overall dataset contains 33 sequences, with a total of 74K frames. On average, the sequences are 75 s long, with a longest sequence duration of 248 s. We emphasize the importance of longer *duration* sequences to capture variability in animal behaviors. This is different from the *NfS* philosophy which focuses on extremely fast behaviors. More general statistics and comparisons are shown in Table 1.

After the selection process, the final dataset consists of 17 different types of marine animal: octopuses, sharks, dolphins, larvaceans, jellyfish, squid, turtle, stingray, and 9 species of fish (striped fish, parrotfish, lionfish, angelfish, boxfish, snapper, barracuda, jack, and blue tangs). They are distributed over several visually distinct marine habitats including coral reefs, sea grass, the shallow mid-water column, the deep mid-water column, and near sandy and rocky seabeds. And finally exhibit several types of swimming behaviors such as fast, medium, and slow constant swimming, darting, crawling, and stop and go maneuvers.

It is common to label sequences with attributes that may cause tracker difficulties and group evaluations by those attributes in order to determine which attributes cause issues during tracking. For this, we select 6 standard attributes including: scale variation (SV), low resolution (LR), partial occlusions (PO), difficult backgrounds (DB), and similar objects (SO), with a description of each and how they are determined in Table 3. Because of the inherent physical active tracking scenario and longer durations, all or most of the sequences in this dataset have camera motion, in-plane rotation, and show object appearance changes from rotation or deformation.

In order to evaluate metrics that are more specific to the underwater domain and habitats, we include 7 additional attributes that were manually determined. These are mid-water (MW), seabed (SB), coral reef (CR), seagrass (SG), intermittent sand or rocks (IS), and active lighting (AL) or passive lighting (PL). The descriptions of these attributes are also listed in Table 3. We note that CR, SG, and IS are all subsets of SB and are more notable for having DB. Because these environments are more typical in shallow regions, and because data was collected during the day, these also have PL. Likewise, MW environments tend to have simple backgrounds (they are generally blue or black), but may have more foreground clutter due to marine snow. In deeper MW environments, it is common to use AL, which amplifies the visual clutter due to marine snow, and so there is coupling inherent in these attributes that may be difficult to separate. A distribution of the sequences and associated attribute labellings is shown in Fig. 2.

4 Evaluation of State-of-the-Art Semi-supervised Trackers

4.1 Tracker Selection

Here we discuss which trackers we selected for comparison. We focus on semi-supervised single-object trackers because we intend to use these on vision-based AUVs for tracking of individual marine organisms where large training datasets do not exist. In real-time, semi-supervised object tracking, given a video or live-stream of images, an initial target is determined from a user-specified bounding box in the initial frame. Only based on this information, the tracker must predict the target bounding box on all subsequent frames, in real-time, without further user input.

While semi-supervised trackers have a long history with many types of architectures, performance in all recent benchmarks such as LaSOT, GOT10K, UAV20L, TrackingNet, and more, are all dominated by those based on deep-learned neural networks. We thus focus on those trackers. We selected 13 recent trackers who self-report highest scores

Table 3 Attributes selected for benchmarking

Attr.	Name	Description
SV	Scale variation	(Auto) Ratio of bbox px exceeds [0.5, 2] from initial bbox
ARC	Aspect ratio change	(Auto) Ratio of bbox aspect ratio exceeds [0.5, 2] from initial bbox
LR	Low resolution	(Auto) Bbox is less than 1000px in area
PO	Partial occlusion	(Manual) Object is partially occluded for more than 1 frame
DB	Difficult background	(Manual) Bbox overlaps with complex background
SO	Similar objects	(Manual) Similar looking objects are nearby target object
Env. Attr.	Name	Description
MW	Midwater	Target does not overlap with seafloor
SB	Seabed	Target overlaps with seafloor at least half the sequence
CR	Coral Reef	Target overlaps with coral reef at least half the sequence
SG	Seagrass	Target overlaps with seagrass at least half the sequence
IS	Intermittent sand/rocks	Target overlaps with sand/rock at least half the sequence
AL	Active lighting	Scene is predominantly illuminated by AUV/diver
PL	Passive lighting	Scene is predominantly illuminated by environment (sun)

The first six attributes are common to all object tracking datasets and the latter seven are peculiar to the underwater environment. SV, ARC, and LR were automatically computed from characteristics of the labels of the corresponding video sequences. The remaining attributes were derived manually because they require some qualitative interpretation

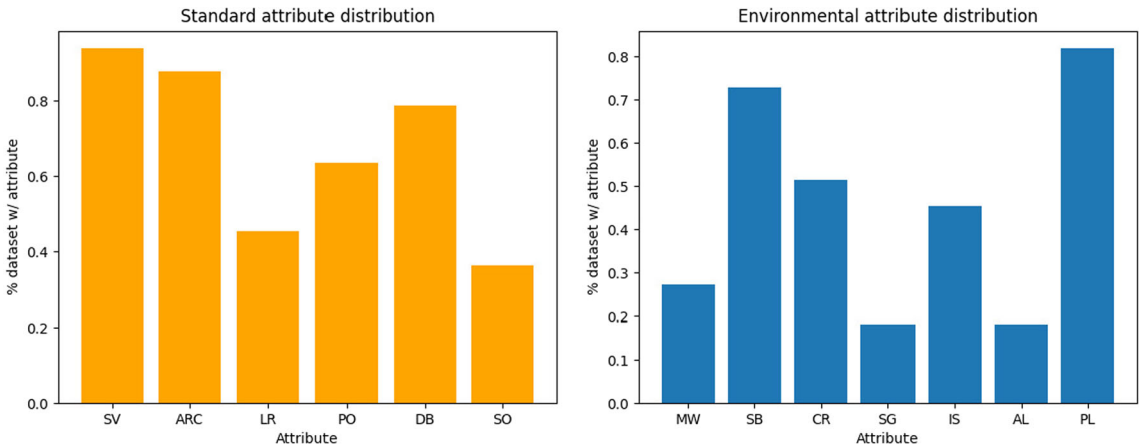

Fig. 2 Distribution of sequences with each attribute in our dataset. On the left are standard generic attributes that we evaluate on and on the right are environmental-based attributes unique to our dataset (Color figure online)

on previous benchmarks, discussed in Sect. 2, with varying underlying architectures. As an additional baseline, we include older top-performing trackers as reported by the official LaSOT benchmark. We further select trackers that are capable of running in real-time (above 10 fps) on currently available hardware and are used specifically for single-object tracking tasks. As described in Sect. 2, we also focus on two primary architectures, Siamese-based and online discriminative-based trackers. For evaluations we have thus selected: SiamRPN++ (Li et al., 2019), DaSiamRPN (Zhu et al., 2018), SiamMask (Wang et al., 2018), ECO (Danelljan et al., 2017), ATOM (Danelljan et al., 2019), DiMP (Bhat et al., 2019), PrDiMP (Danelljan et al., 2020), SuperDiMP, which is a combination of DiMP and the regressor from PrDiMP, KeepTrack and KeepTrackFast (Mayer et al., 2021), TransT (Chen et al., 2021), TrSiam (Wang et al., 2021), and TrDiMP (Wang et al., 2021). For abbreviations and properties we consider refer to Table 4. As is common practice, we run each algorithm as provided (Fan et al., 2020), because the following reasons: they may require different training strategies, they are sensitive to training settings, and as-is trackers are typically already optimized. There are many subtle differences between each tracker that are difficult to list and enumerate, but broadly we can characterize them as Siamese-based: SiamRPN++, DaSiamRPN, SiamMask, ECO, TransT, and TrSiam, or online-discriminator based: DiMP, PrDiMP, SuperDiMP, KeepTrack, and TrDiMP. All these algorithms utilize pre-trained (using generic datasets) deep convolu-

Table 4 Overview of selected trackers, the abbreviations we use to denote them, their source year, general architecture (siamese-based vs. online discriminator based), and an inclusion of a transformer network for additional attention processing

Tracker	Abbrev.	Source	Arch. (S/OD)	Trans. layer?
ECO (Danelljan et al., 2017)	EC	CVPR17	S	N
DaSiamRPN (Zhu et al., 2018)	DS	ECCV18	S	N
ATOM (Danelljan et al., 2019)	AT	CVPR19	S	N
SiamRPN++ (Li et al., 2019)	SR	CVPR19	S	N
SiamMask (Wang et al., 2018)	SM	CVPR19	S	N
DiMP (Bhat et al., 2019)	DP	ICCV19	OD	N
PrDiMP (Danelljan et al., 2020)	PR	CVPR20	OD	N
SuperDiMP	SD	2020	OD	N
KeepTrack (Mayer et al., 2021)	KT	ICCV21	OD	N
KeepTrackFast (Mayer et al., 2021)	KF	ICCV21	OD	N
TransT (Chen et al., 2021)	TT	CVPR21	S	Y
TrSiam (Wang et al., 2021)	TS	CVPR21	S	Y
TrDiMP (Wang et al., 2021)	TD	CVPR21	OD	Y

Note that SuperDiMP is a combination of PrDiMP and DiMP provided by Danelljan et al. (2020) without further publication, and KeepTrackFast uses slight changes in parameters, but is the same architecture as KeepTrack, and is provided in Mayer et al. (2021)

tional neural networks to perform feature extraction. In most cases, we use the ResNet-50 (He et al., 2016) feature extractor when available, though older networks such as ECO relies on VGG-M (Chatfield et al., 2014), and ATOM relies on ResNet-18 (He et al., 2016).

Vaswani et al. (2017) introduced Transformers as a new type of network architecture that produced state-of-the-art results on a variety of neural network related classification and decision making tasks. Only recently however have these innovations been applied into the tracking domain, and are presented in TransT, TrDiMP, and TrSiam. Their performance has not been characterized by a standardized dataset yet, and so we include them here as well. For further details on all trackers, please refer to each respective publication (Table 4).

4.2 Results of Evaluation and Metrics

We ran all trackers across the VMAT dataset for evaluation, for a qualitative sample of these runs, refer to (Fig. 3). We adopt standardized metrics for evaluating state-of-the-art trackers on our benchmark dataset. Namely e consider the (1) *success rate*, (2) *precision*, and *normalized precision* metrics accumulated across all frames in each sequence, and averaged across the entire dataset and over each attribute subset. Many of the trackers are stochastic, and so we average results over 5 runs, as in many other benchmarks (Fan et al., 2020; Müller et al., 2018). These metrics are well-established and described in detail in Müller et al. (2018). For completeness, we give brief descriptions of each here. The success rate, or overlap, for each frame and tracker is computed by taking the intersection-over-union (IoU), which is a value between 0 and 1, of the groundtruth bounding box and the predicted bounding box of the tracker. To generate reasonable visualizations and rankings, the IoU is subject to thresholds spanning 0–1, a frame is considered a "success" if the IoU exceeds the threshold or not. The percentage of "successful" frames can then be computed for each threshold as shown in Fig. 4. Rankings are then taken by estimating the area under the success curve (AUC). Because some trackers may have reasonably predicted overlapping bounding boxes but focus on the incorrect region of an object, it is standard practice to also consider the *precision* metric as well. Precision is computed by taking the distance of the center of the groundtruth bounding box to the center of the predicted bounding box, this is measured in pixels. For precision, a different threshold, measured using number of pixels distance, is applied to generate meaningful visualizations and rankings. For overall precision rankings in Fig. 4, rather than computing an AUC-like metric for precision, it is standard practice to report the precision for a threshold of 20px, which we do here as well. Since distances are measured in pixels, precision is subject to resolution of the target, and so normalized precision accounts for image resolution. More details about these metrics can be found in Müller et al. (2018).

Some benchmarks, such as Fan et al. (2020), do not report speed, which is a critical evaluation metric if these algorithms are to be used in real-world active tracking systems. We thus ensure to provide a baseline measurement of speed in Fig. 7.

The results for the full dataset can be found in Fig. 4, results along standard attributes are in Fig. 5, and results along the unique underwater attributes are in Fig. 6. To enable consistent comparisons, especially related to speed, all results are run on the same desktop with a Nvidia GeForce 1080 GPU, Intel Core i7-6900K CPU, and 64GB RAM.

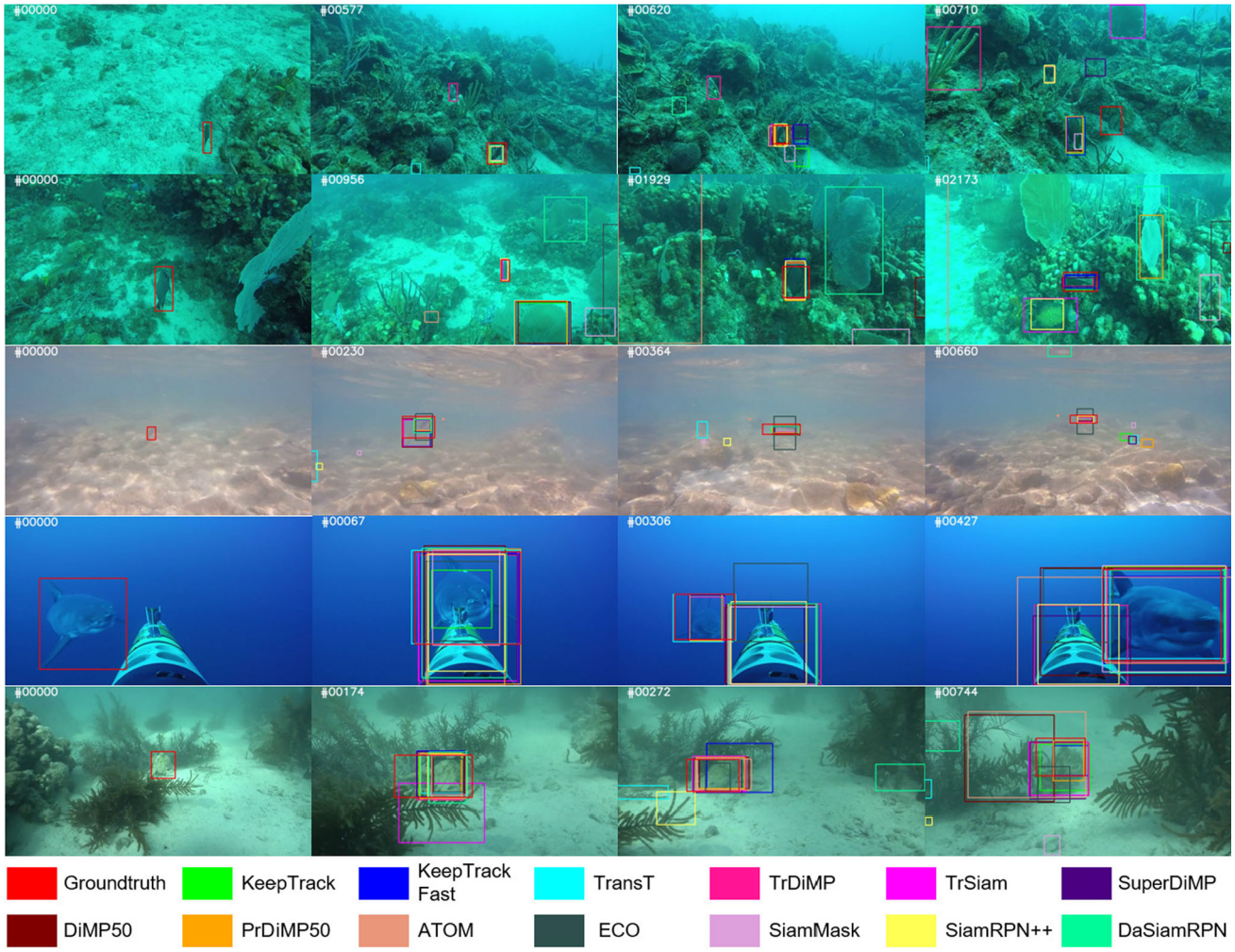

Fig. 3 Sample representative results on difficult sequences of a bluetang, angelfish, squid, shark, and octopus (from top to bottom) (Color figure online)

5 Real-World Experiments: Tracking in the Wild Using an AUV

Solely evaluating performance of semi-supervised trackers on pre-recorded videos is insufficient to fully estimate their benefits and issues when operating in real-world conditions. Pre-recorded videos mask issues where real-time decision-making on the AUVs can cause additional problems. For instance, if the tracker focuses on the wrong object even momentarily, it could cause the entire platform to further point away from the target. This creates a positive feedback loop that likely results in catastrophic failure of the tracker. In pre-recorded videos however, trackers can often re-identify targets because it is likely they will re-center in the video eventually given the nature of the recording. Thus, in order to better understand in-situ performance, we deployed a semi-supervised visual tracker to actively control a real AUV and track a marine animal in the wild. We describe our system implementation and experiment details below and provide some qualitative evaluation of the field trials afterward.

5.1 System Overview

Our goal was to test semi-supervised tracking of a complex organism in-situ and in real-time on an AUV. To do this we used a custom AUV developed at the Woods Hole Oceanographic Institution, known as CUREE (Girdhar et al., 2023), rated to 100 m, equipped with a camera, a high-bandwidth connection to an on-board Nvidia Jetson Xavier which had direct control of the thrusters over a ROS interface (ROS, 2022). All image processing and control action selection was performed on the AUV itself. The vehicle is equipped with 3 wide-angle cameras looking 30° down from the horizon. The middle camera provides a 720p RGB stream at 15 fps, while the two side cameras provide 1080p monochrome streams and can be used for stereo. We used KeepTrackFast because

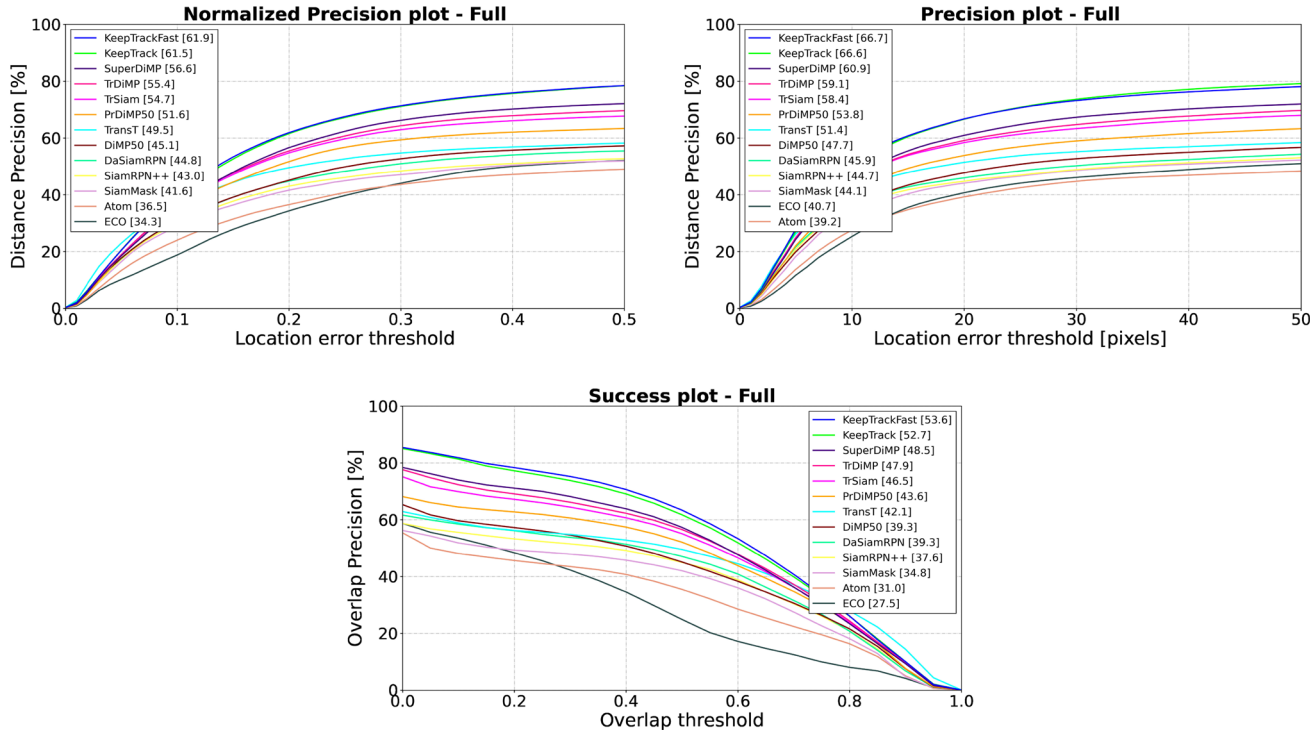

Fig. 4 Results of all trackers on the full dataset using success AUC, precision, and normalized precision metrics (Color figure online)

it was both the best performing algorithm on VMAT while also able to run at roughly 7–9 fps continuously on the edge device Nvidia Jetson Xavier, which is at the bare minimum speed to enable real-time feedback control. Note that this device is significantly less powerful than the system used in Fig. 7, but is necessary for use on smaller autonomous vehicles. This on-board computing limitation is also why the resolution and speed of the video processing was low. The camera is mounted at a 30° downward angle from the horizon in order to better track animals along the sea-floor, which is also typically a more complex environment. The AUV is also equipped with a WaterLinked DVL A50, or a *Dopper Velocity Log*, which provides capabilities to accurately estimate 3D vehicle velocities and altitudes, which can in turn be integrated to estimate positions via *dead-reckoning*. Finally, the AUV has 6-thrusters capable of full 6-DOF control.

5.2 Real-World Deployment Details

We deployed our system in parallel to our dataset collection efforts in the U.S. Virgin Islands in October 2022. A human operator was stationed on a boat and was tethered to the AUV. The operator had a high-bandwidth tether connection to the AUV in order to perform remote control of the vehicle for the initial target search and specification step of the experiment. This connection also provides the operator with a live stream of the video from the AUV. Once a target is identified by the operator, they draw a bounding box over the target using a GUI developed for this task. After the bounding box is specified, the AUV goes into a fully autonomous tracking mode. While this may be considered an ROV setup, once the initial bounding box has been specified, the tether can be disconnected, and the goal is for these algorithms to be adopted on full AUVs.

In fully autonomous mode, the AUV performs closed-loop visual servoing control using the bounding boxes provided by KeepTrackFast. Because of the monocular setting, and because there is no known information about the target a priori, we cannot compute a target size or distance estimate. Consequently, we command the vehicle to maintain constant width of the bounding box. Qualitatively, we found this metric to be more stable than height or area. The resultant control loop guides the robot closer to the target if the width of the bounding box grows, and vice versa. The AUV also yaws accordingly to center the target in its field of view. If the center is above or below, the vehicle rises or sinks in the water column. All these parameters are controlled via a series of PID loops running at least at 9 Hz, which is the speed of KeepTrackFast running on the Jetson Xavier. For additional safety, to prevent damaging coral reefs, using the DVL's measurement of altitude, we disallow the vehicle from moving below 0.5–1 m from the bottom.

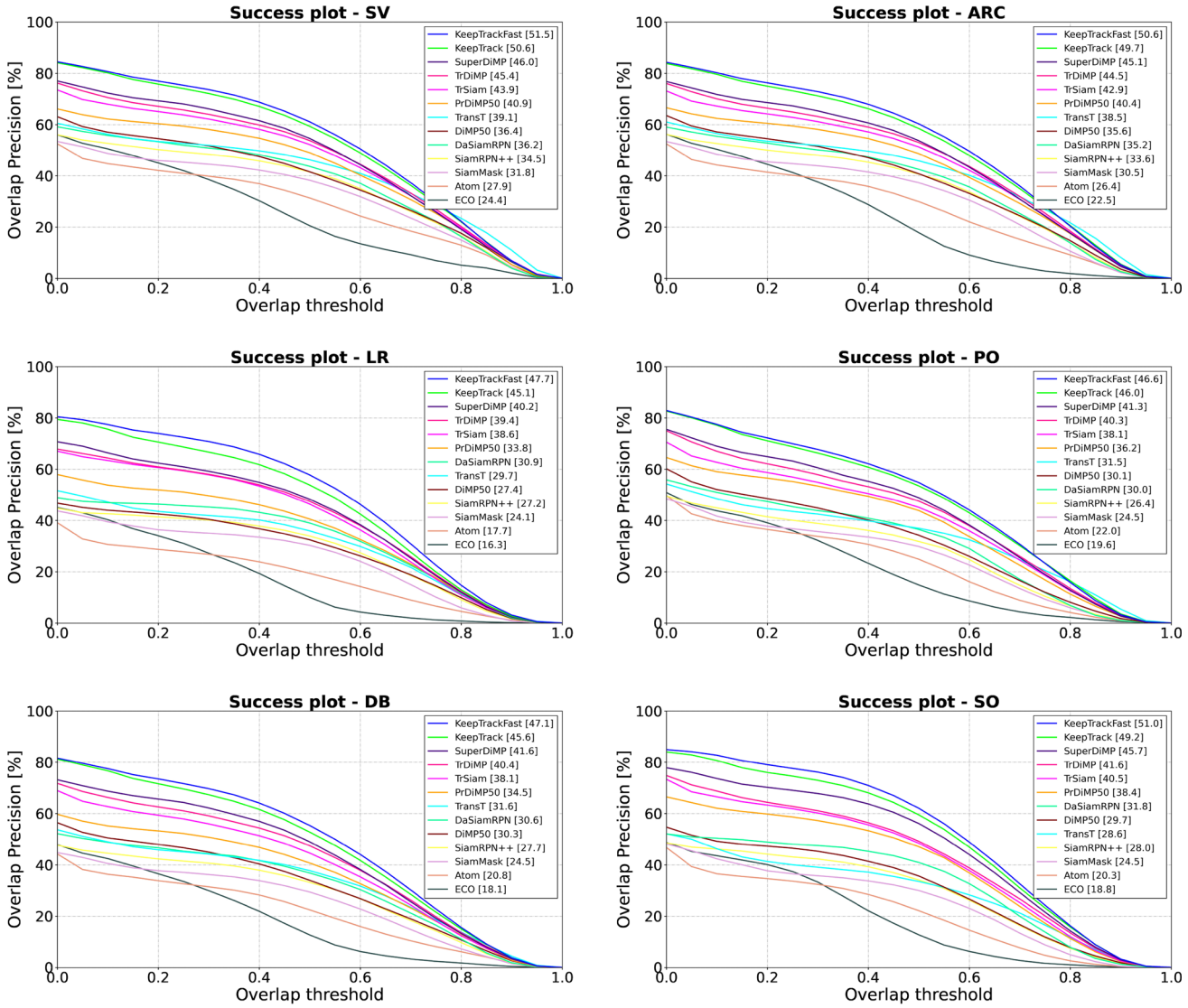

Fig. 5 Success results of all trackers on sequences containing the specified generic attributes (Color figure online)

5.3 Results of Demonstrations

We were able to successfully track several organisms, including a barracuda (Fig. 8), multiple jellyfish (Fig. 10), and some smaller fish like jacks (Fig. 11).

The barracuda was tracked for nearly 10 min across a trajectory of roughly 100 m in length out and back. The barracuda swam from an initial point next to a particular species of coral, known as *dendrogyra*, had multiple encounters with other organisms along its track, and eventually returned to the same coral before the AUV lost track. The AUV was able to autonomously follow the barracuda for the majority of the track, however, there were short instances where the human operator needed to manually control the AUV when KeepTrackFast lost the target. Overall, the tracker was able to maintain track of the animal across a wide range of terrains, as it swam over coral, sand, seagrass, and nearby several other fish, turtles, and other organisms. However, at around 10 min, it swam near coral that also resembled it, and KeepTrackFast and the human operator as well were unable to see it afterwards. Because of the DVL and depth sensor on-board the AUV, we are able to estimate a full 3D trajectory of the vehicle in real-time, which can serve as a proxy of the trajectory of the barracuda. The AUV also has stereovision, so with further processing, we would be able to recover a more accurate 3D trajectory of the barracuda itself by computing the 3D offset of the barracuda from the vehicle, however, that is out of scope for this paper. The resulting trajectory estimates, and time periods when human intervention was required, are shown in Fig. 9. We needed to manually align the vehicle compass, due to a compass miscalibration while

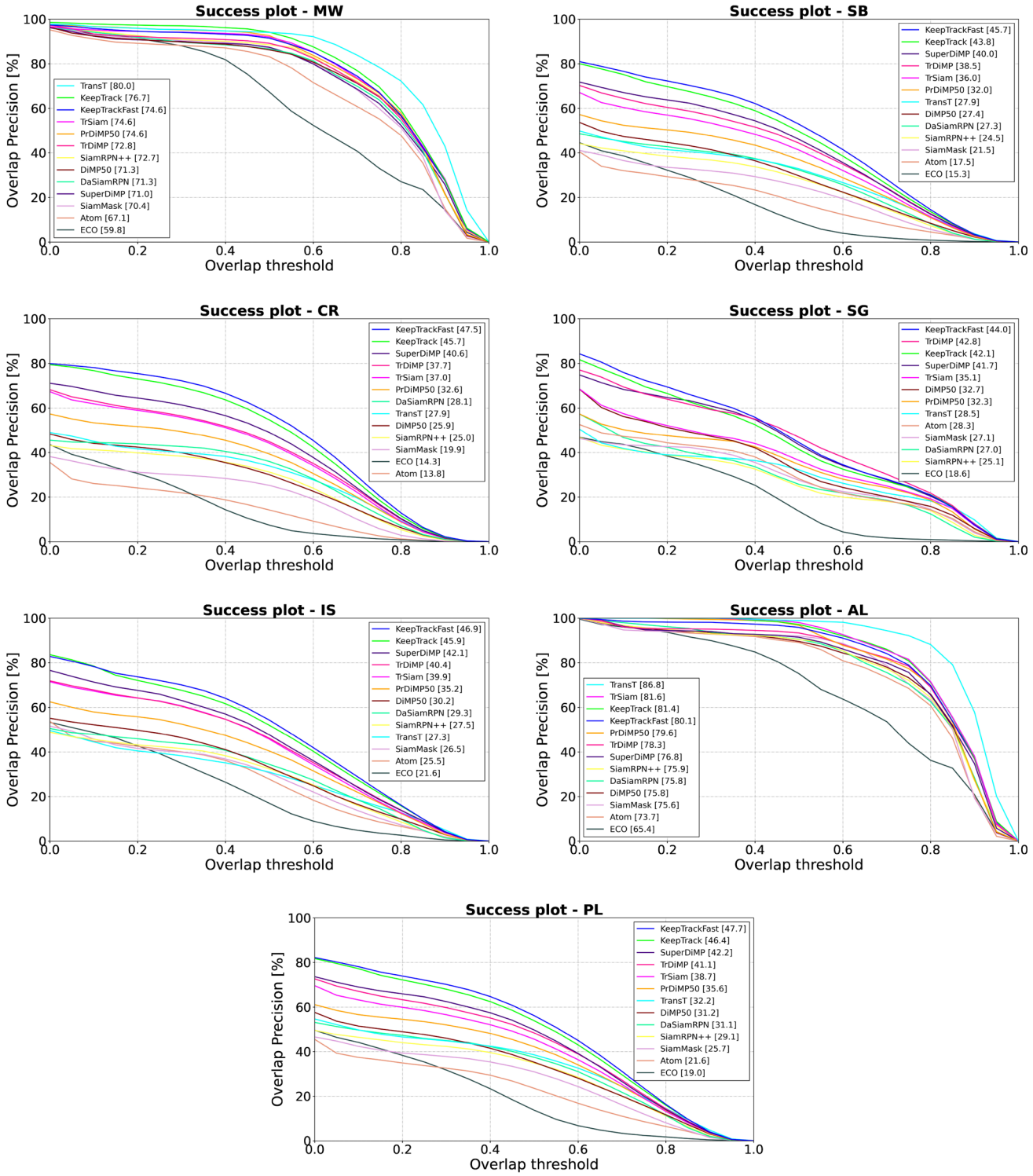

Fig. 6 Success results of all trackers on sequences containing the specified environmental attributes (Color figure online)

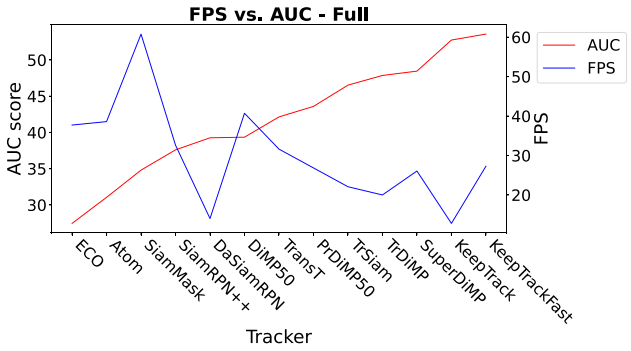

Fig. 7 Frame rate vs. AUC over the full VMAT dataset for each tracker. With the exception of KeepTrackFast, SiamMask, and DiMP, most trackers appear to tradeoff speed for accuracy. Results run using a desktop with Nvidia GeForce 1080 GPU, Intel Core i7-6900K CPU, and 64GB RAM (Color figure online)

we were tracking the barracuda, in order to post-process the map overlay in Fig. 9.

We further initially tested the capability of these algorithms to replicate results such as Yoerger et al. (2021) and Katija et al. (2021) to track jellyfish in more visually complex scenarios. To do this, we tracked multiple jellyfish, the resulting vehicle depth plots, which again, roughly estimate the depth profiles of the jellyfish, are shown Fig. 10. While we were unable to test in situations with large clusters of jellyfish as in Katija et al. (2021), we were still able to track jellyfish in more visually complex scenarios for reasonable time lengths. The tracks ended because the vehicle is tethered to the boat of the human operator, which is in turned moored, and once the jellyfish reached beyond the length of the tether it was no longer possible to continue. However, in the case of the jellyfish tracks, once the initial bounding box was specified, the entire track was fully autonomous.

Finally, we attempted tracks of many smaller organisms, of which we show one demonstration of a track of a jack in a small school of three in Fig. 11. In this track, three jacks were swimming together and we commanded the vehicle to track one. We were able to track at least one jack for several minutes, though the human operator needed to intervene for a few seconds several times in cases where the tracker or the vehicle was unable to maintain the track because either the tracker got confused and tracked a coral, or the vehicle itself was physically slower.

6 Discussion

6.1 Tracker Evaluation

In Fig. 4, we see that the KeepTrack algorithm (KeepTrackFast is the same underlying algorithm, just slightly different parameters to improve speed) the best performing algorithm over the entire dataset and, in Figs. 3 and 6, along most attributes by a fairly wide margin, having a 53.6 and 52.7 success overall and 66.7 and 66.6 precision overall for KeepTrackFast and KeepTrack, respectively. The next best performing is SuperDiMP, achieving 48.5 and 60.9 AUC and precision scores respectively. This makes sense because KeepTrack is based on SuperDiMP, and is specifically designed to handle distracting objects and backgrounds, which make up over 35% and 70% of the dataset. Surprisingly, transformer-based networks (TrDiMP, TrSiam, and TransT), which are mostly based on SuperDiMP, perform marginally worse than SuperDiMP in most cases. The only cases where these seem to exceed performance is in the midwater and active lighting results, where all trackers performed exceptionally well, likely because the backgrounds are much less challenging relative to the targets in our dataset. Also, online discriminative style trackers tend to perform better than Siamese-based ones.

In terms of the standardized attributes shown in Fig. 5, the rankings are mostly stable, suggesting that some of these attributes are either heavily correlated or that certain innovations seem confusing across multiple of these dimensions, making them still difficult to analyse fully. However, it is clear that low-resolution frames are still a significant concern, with the best performing tracker only achieving 47.7 success rate. These are exacerbated in many marine animals, such as several fish species or in the case of darting octopuses, where their bodies become extremely narrow and difficult to track purely from appearance. We believe these situations can be mitigated with a more probabilistic handling of search area, which none of the current algorithms address. These can take approaches from the multi-object tracking community, as in DeepSORT (Wojke et al., 2017), where appearance uncertainty and localization and motion modelling are both taken into account for estimating the next location.

The underwater-specific attributes, in Fig. 6, however are more insightful. In the midwater and active lighting scenarios, which as previously noted are highly correlated in this dataset (that could be mitigated by collecting shallow nighttime data), most trackers perform exceptionally well. This suggests that it may be reasonable to deploy semi-supervised trackers in deeper midwater environments that are actively lit, especially if the animals tend to be solitary. It is also interesting to note that the transformer-based architectures also perform dramatically better under those conditions, perhaps because of better appearance representations.

By contrast, complex environments such as coral reefs and sea grass still pose exceptional challenges, as shown by Fig. 6 in the seabed (SB), coral reef (CR), and seagrass (SG) environments. The results are especially skewed downward in the case of seagrass. Across these attributes, online discriminators, especially KeepTrack and SuperDiMP, still are the most accurate.

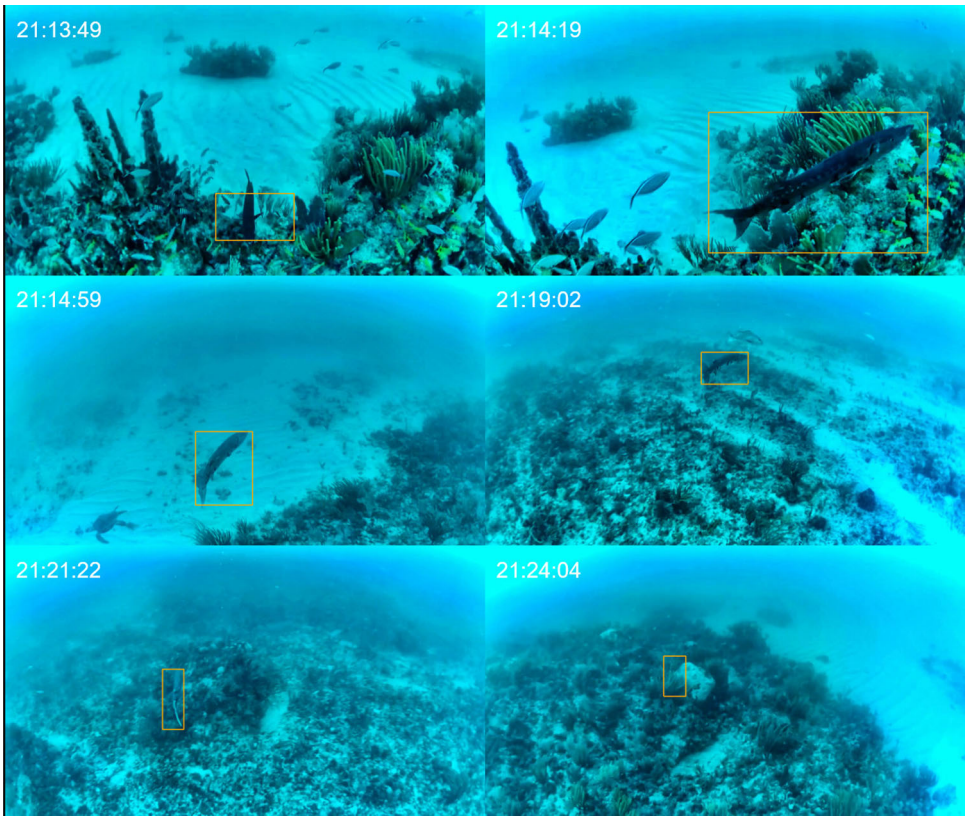

Fig. 8 We spontaneously and mostly autonomously tracked a barracuda at Joel's Shoal, St. John, U.S. Virgin Islands, USA on Nov. 3rd, 2022, for roughly 10 min using a semi-supervised visual tracker on an AUV. Tracking was performed at roughly 7–9 fps at 720p resolution on-board a Nvidia Jetson Xavier using KeepTrackFast (Mayer et al., 2021). Images are taken from the AUV perspective. The UTC time is listed in the upper left of each image. The provided bounding box is shown in the 1st frame, it is not well-centered because the operator had to click-and-drag the box while on-board an oscillating boat, and the rest are automatically generated during tracking and used to visual servo the AUV. In the last frame, the barracuda is actually a few pixels to the upper right of the predicted bounding box, which highlights in extraordinary challenges of this environment, as even a human cannot separate the barracuda from the background, even though it is minimally occluded. The AUV loses the barracuda here and we ended the track. For the full track video, please refer to supplementary video S1 (Color figure online)

In many cases, such as most of the fish species in CR environments, or the squid in rocky terrain, as can be seen in Fig. 3, we found that all trackers were likely to fail when the animal was both in (A) *specific orientations or body configurations (such as compressed body profiles in squid)* and (B) nearby complex backgrounds or similar objects. For example, most fish in the dataset are extremely thin when viewed from the back or top, in these perspectives they lack any distinguishing visual characteristics. The homogeneous colors in these settings, due to color absorption, and similar textures from the backgrounds, in the case of fish in reefs or squids in rocky terrain, tend to cause all the trackers to fail. This can be seen in the bluetang, angelfish, and squid examples in Fig. 3.

One reason that the seagrass examples are difficult is that we have several octopus sequences, as shown in the fifth row of Fig. 3, that were showing some amount of camouflage.

Our experiments suggest that the grand challenge in the domain of marine animal tracking problems corresponds to tracking octopuses, both in fully-supervised or semi-supervised tracking contexts, because of their extreme adversarial nature to tracking. They have evolved stealth behaviors and unequalled rapid adaptive camouflage. On a typical forage, an octopus might change its camouflage more than 170 time/h (Hanlon et al., 1999). Moreover, the octopus can change its body shape dramatically with its 8 malleable arms, and it can also change its 3-D skin texture to match fine or course texture of adjacent corals, sponges, tunicates, algal epiphytes, etc. so that its ability to blend in is exceptional, even very experienced diving biologists are fooled. In our dataset however, the human annotators are still able to pick them out in the lower resolution dataset, so they present a clear target for future innovations.

We believe one major explanation for the success of online discriminative networks is to consider the problem from

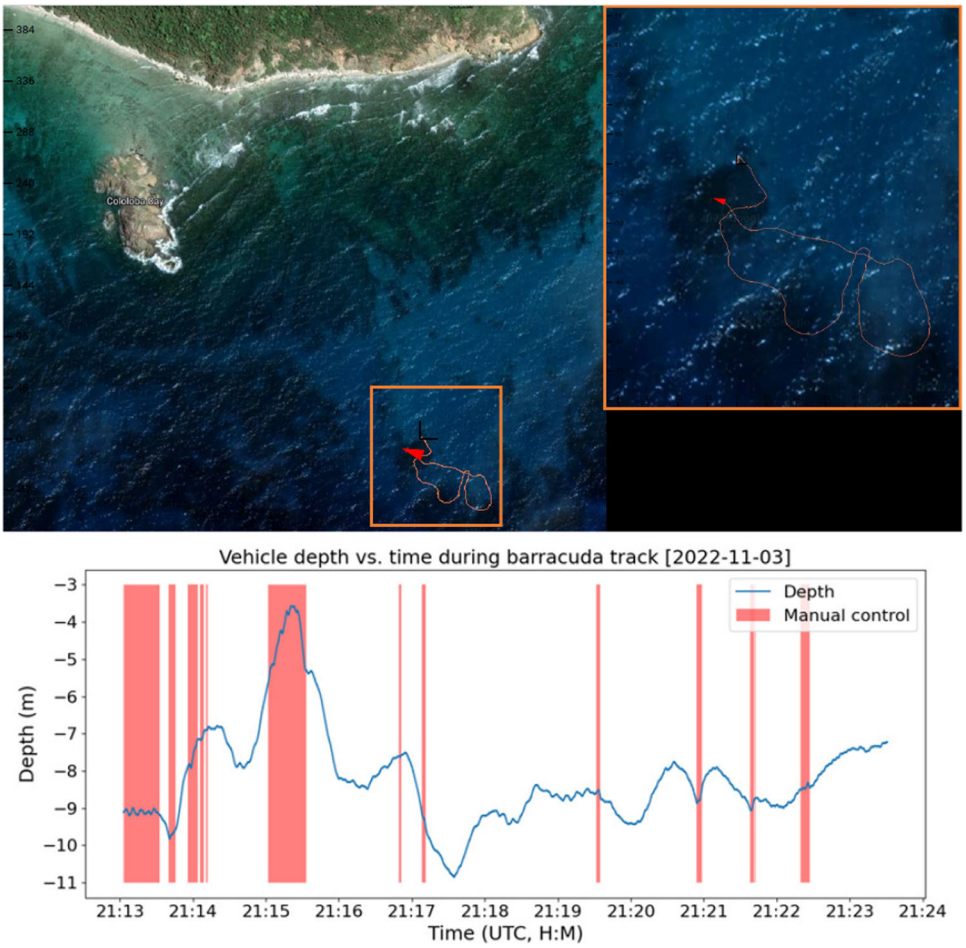

Fig. 9 Here we show the full 3D trajectory estimates of the barracuda track from Fig. 8, which swam roughly 100 m out and back. On top is the AUV real-time estimate of its trajectory, achieved via dead-reckoning by integrating DVL measurements, as it follows the barracuda (which required manual post-processing to align it with the GPS map). Finally, on the bottom, we show the vehicle depths during tracking, along with red highlights indicating when the KeepTrackFast algorithm was unable to stay focussed on the correct target and the operator needed to briefly, manually control the vehicle until KeepTrackFast was able to re-acquire the target. The beginning section required manual control due to the initial training time of KeepTrackFast, and intermittent areas when the barracuda resembled the corals underneath, similar to the last frame in Fig. 8. Note that these position estimates can serve as proxies for animal positions and velocities. Using stereo, we can generate true animal position and velocity estimates, but is out of the scope of this paper (Color figure online)

a perspective on long-term robustness vs. short-term accuracy. In the visual tracking problem, the best predictor of the appearance of an object at frame t is its appearance at $t-1$. And so having a representation of this appearance model increases short-term accuracy. However, because only at time $t=0$ do we have a ground truth label, this is the most robust appearance model. This means that all decisions made on later object appearances are based on weak labels provided from the algorithm itself, this is prone to what is called concept drift in the online learning community (Mittal & Kashyap, 2015). All the online discriminative-based trackers attempt to combat this issue by training not only on the most recent appearance, but also a history of appearances as well as augmentations of the initial image. However, many of these approaches are not well-principled. This trade-off is perhaps most easily seen in the shark track shown in Fig. 3. Here, the Siamese-based networks are able to recover because they are robust to concept drift, whereas most trackers end up learning the appearance of the AUV that partially occludes the shark for a short time, even though the AUV was not specified in the original object appearance.

Though our dataset is relatively small, we hope to continue to contribute to it with more sequences in the future. We believe it provides a necessary and sufficient first step towards understanding performance of semi-supervised trackers in a variety of conditions specific to underwater domains. We can begin to distinguish between simpler and more difficult environments, species, and behaviors, so that AUV practitioners and the semi-supervised tracking community can begin to take more educated directions in their use in the wild.

6.2 Challenges in Real-World Marine Animal Tracking

Overall, the real-world tracking results shown in Figs. 8, 10, and 11 illustrate the exciting potential, but also some of the remaining challenges, of using semi-supervised trackers on vision-capable AUVs to perform longer-term marine animal tracking and monitoring, without the need for extensive labelled datasets. To the authors' knowledge, these demonstrations are among the first attempts to track complex marine animals with an AUV, in a difficult underwater setting, using only a visual semi-supervised tracker. There were *no a-priori*

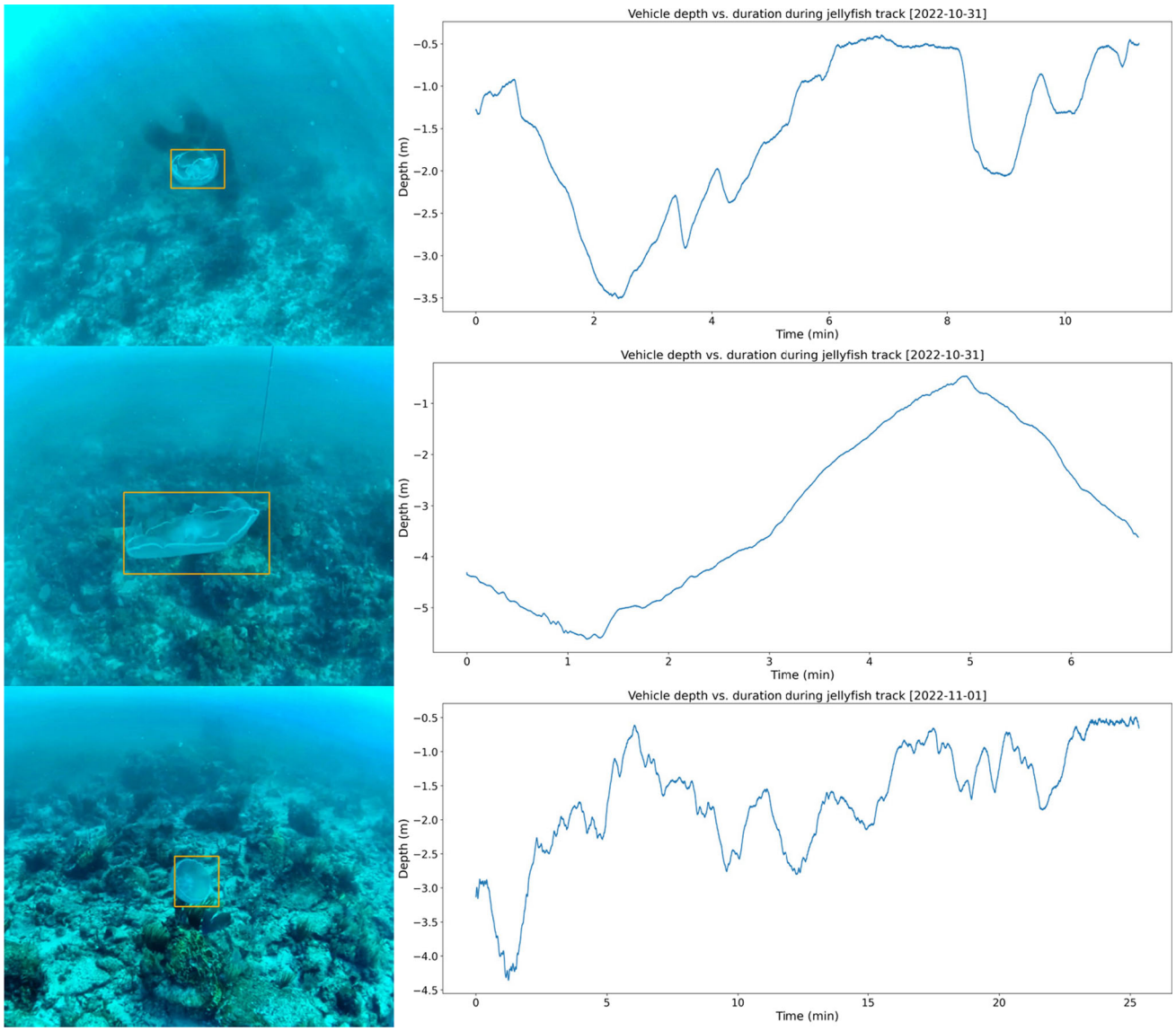

Fig. 10 We tracked 3 separate jellyfish at 2 different reefs in St. John, USVI, USA. From top to bottom they were at Tektite, Booby Rock, and Booby Rock for roughly 12, 7, and 25 min respectively. Once the initial bounding box was specified, the remainder of the tracks were completely autonomous. The tracks ended because the vehicle was tethered and the jellyfish drifted beyond the tether range. The vehicle depth profiles are shown on the right, which can approximate the depth of the jellyfish, and their initial bounding boxes and images are on the left, which have been cropped for space. Videos of the jellyfish tracks are available in the Supplementary materials S2–S4, respectively (Color figure online)

training of the targets or tagging, and most of the tracking was *autonomous*. The results from Figs. 9 and 10 especially highlight the benefits of applying these techniques to generate high-resolution data about animal trajectories in both novel environments and species, and even more so if stereo-vision can be incorporated in the near-term. For future tracks, it would be possible to generate reasonable 3D trajectories all in real-time. With stereo, marine biologists can then use these to estimate velocity and animal size, and in turn quantities such as energetics.

However, many challenges still exist before these can be used more reliably in other scenarios. We discuss some of our findings here based on our tracks of other animals, such as the jack in Fig. 11, and even in some cases for Fig. 8. While in some instances, while the jack was alone and swimming around the axis of the robot, as in Pane 1 in Fig. 11, the tracker had no issues, though the robot was sometimes unable to rotate fast enough. In others, the tracker would either track another jack, as in Pane 2 and 3, because of their very similar appearance. Finally, the tracker would occasionally latch

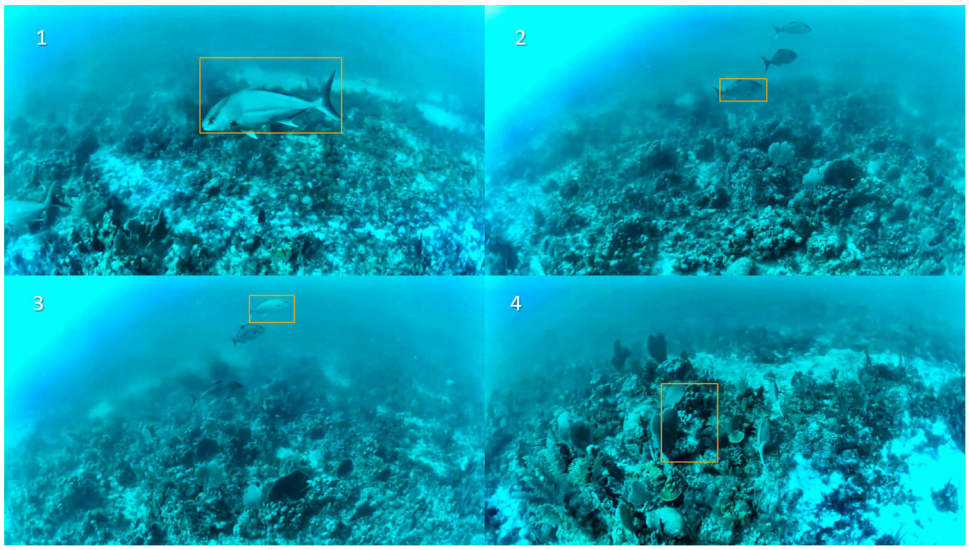

Fig. 11 We autonomously track a type of jack that was swimming with two others. KeepTrackFast oscillated between the three, even though it is designed to minimize these confusions. This also caused the vehicle to jerk between all three, and frequently lose track. The issue is especially highlighted when the fish turn to face away from the camera, as in the 4th frame. The thin profile of the fish and homogeneous colors from color absorption makes the fish difficult to distinguish from the background and surrounding corals. These situations trigger KeepTrackFast to oscillate even more between targets and occasionally background regions. Refer to the Supplementary material S5 for the jack track video (Color figure online)

onto coral reefs, as in Pane 4 of Fig. 11 or the last frame of Fig. 8, when the fish were rotated such that they faced away from the camera, their profile would become thin and resemble much of the fan corals around them. In these cases, the tracker would often, incorrectly, latch onto the corals.

In both the barracuda and jack tracks, we found many problems could be summarized by the following limitations:

- **Limitations of the tracker**
  - *Initialization issues*—importantly, as shown in the manual control plot of Fig. 9, though also present in the jack track as well, we often had to manually control the robot for the first several seconds of the track. This is because KeepTrackFast (and all the KeepTrack and DiMP-based trackers) have an initial training phase that can last several seconds on edge computing devices like the Jetson Xavier. In these cases, it is impossible to track the animal, and for smaller fish which rotate quickly, we often could not continue tracking after initialization. Additionally, initially selecting the bounding box is challenging. The human operator was using a mouse on a non-stationary and small boat and also experiencing latency over the tether, the human operator was unable to perfectly select the target while also maintaining sight of the animal unless it was mostly stationary to start. These issues are only experienced through real-world testing.
  - *Track-time issues*—many of these issues were highlighted previously and discovered through our analysis on the VMAT dataset. Specifically, in cases where the tracked animals rotated or compressed into thin perspectives and were surrounded by either similar objects or complex backgrounds (such as other fish or corals), the trackers were unable to distinguish between the other objects and the target, and often failed in these cases. Extremely fast darting maneuvers, associated with both immediate changes in position and appearance, also tended to cause the tracker to fail. We believe that some investigation into improving speed on edge devices and more probabilistic localization techniques that rely less on appearance in these cases, as proposed in Sect. 6.1, would be helpful.
- *Limitations of vehicle dynamics and other sensing*—the vehicle itself was often too slow to track some species of highly dynamic organisms such as smaller fish or sharks. In addition, many smaller fish spend most of their time close to, or in crevices of, the reef, which the vehicle is unable to approach for coral safety reasons. Many of these issues require hardware changes to the vehicle, or additional sensing techniques to better estimate distance from sensitive boundaries, such as the coral reefs, in order to better avoid them and navigate around them during tracking.

Our analysis of these algorithms on our VMAT dataset also helped both select an appropriate algorithm in terms of both *accuracy* and *speed*. It is important to note that KeepTrack was not useable on our AUV because it ran at only 2–4 fps in our bench tests, which we believe is insufficient for real-world autonomous tracking, compared to 7–9 fps of KeepTrackFast.

7 Conclusion

Behavior of marine animals are inherently hard to study due to lack of availability of datasets characterizing their behavior in their natural environment. In this work we propose the use of a semi-supervised tracker on underwater robots to rapidly collect large datasets with minimal prior knowledge. Our contributions are on three fronts. First, we introduced a marine animal tracking dataset with 33 video sequences captured by mobile camera systems while following a marine animal. Second, we evaluate current state-of-the-art semi-supervised tracking algorithms over this dataset using novel evaluation metrics, specific to the underwater domain, which allow marine practitioners to better distinguish application-specific tradeoffs and capabilities of different trackers. Finally, we have, to the best of our knowledge, demonstrated the first use of a semi-supervised tracker onboard an AUV to track a wide variety of marine animals (barracuda, jellyfish, jacks,..) spontaneously, in visually complex underwater environments. These demonstrations provides encouraging evidence towards the use of these technologies for marine animal tracking.

8 Supplementary information

The full dataset is available at http://warp.whoi.edu/vmat/. We also include the following supplemental videos for the AUV tracks:

- S1 barracuda tracking video.
- S2 jellyfish tracking video, 2022-10-31 at Tektite.
- S3 jellyfish tracking video, 2022-10-31 at Booby Rock.
- S4 jellyfish tracking video, 2022-11-1 at Booby Rock.
- S5 jack tracking video.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11263-023-01762-5.

Acknowledgements Special thanks to Amy Kukulya and Roger Stokey at the Woods Hole Oceanographic Institution for the shark videos, the Mesobot team (Yoerger et al. 2021) for the solmissus and larvacean videos. Thanks to the WHOI WARPLab members Seth McCammon, John San Soucie, Stewart Jamieson, Daniel Yang, and Jessica E. Todd, Ethan Fahnestock for editing, and Cynthia Becker, Prajna Jandial, Nadège Aoki, Nathan Formel, and Sierra Jarriel for species identification and support in the USVI data collection efforts. This work is supported in part by The Investment in Science Fund at WHOI, and NSF NRI awards # 1734400, 2133029. Levi Cai was supported by the NDSEG Fellowship. Also thanks to the NVIDIA Hardware Grant for a GPU for running evaluations.

References

Akkaynak, D., & Treibitz, T. (2019). Sea-thru: A method for removing water from underwater images. In *IEEE CVPR*.

Bateson, M., & Martin, P. (2021). Measuring behaviour: An introductory guide.

Bhat, G., Danelljan, M., Van Gool, L., & Timofte, R. (2019). Learning discriminative model prediction for tracking. In *2019 IEEE/CVF international conference on computer vision (ICCV)* (pp. 6181–6190). IEEE. https://doi.org/10.1109/ICCV.2019.00628. https://ieeexplore.ieee.org/document/9010649/ Accessed 19 April 2021.

Caelles, S., Pont-Tuset, J., Perazzi, F., Montes, A., Maninis, K.-K., & Van Gool, L. (2019). The 2019 DAVIS challenge on VOS: Unsupervised multi-object segmentation. arXiv:1905.00737. Accessed 23 March 2021.

Chatfield, K., Simonyan, K., Vedaldi, A., & Zisserman, A. (2014). Return of the devil in the details: Delving deep into convolutional nets. In *British machine vision conference (BMVC)*. Accessed 29 April 2022.

Chen, X., Yan, B., Zhu, J., Wang, D., Yang, X., & Lu, H. (2021). Transformer tracking. In *2021 IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (pp. 8122–8131). IEEE. https://doi.org/10.1109/CVPR46437.2021.00803. https://ieeexplore.ieee.org/document/9578609/. Accessed 25 April 2022.

Danelljan, M., Bhat, G., Khan, F. S., & Felsberg, M. (2017). ECO: Efficient convolution operators for tracking. arXiv:1611.09224. Accessed 14 December 2020.

Danelljan, M., Bhat, G., Khan, F. S., & Felsberg, M. (2019). ATOM: Accurate tracking by overlap maximization. arXiv:1811.07628. Accessed 20 February 2021.

Danelljan, M., Gool, L. V., & Timofte, R. (2020). Probabilistic regression for visual tracking. In *2020 IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (pp. 7181–7190). https://doi.org/10.1109/CVPR42600.2020.00721. ISSN: 2575-7075.

Danelljan, M., Hager, G., Khan, F. S., & Felsberg, M. (2015). Convolutional features for correlation filter based visual tracking. In *2015 IEEE international conference on computer vision workshop (ICCVW)* (pp. 621–629). IEEE. https://doi.org/10.1109/ICCVW.2015.84. http://ieeexplore.ieee.org/document/7406433/ Accessed 2019-07-22

Dawkins, M., Sherrill, L., Fieldhouse, K., Hoogs, A., Richards, B., Zhang, D., Prasad, L., Williams, K., Lauffenburger, N., & Wang, G. (2017). An open-source platform for underwater image and video analytics. In *2017 IEEE winter conference on applications of computer vision (WACV)* (pp. 898–906). https://doi.org/10.1109/WACV.2017.105

Fan, H., Bai, H., Lin, L., Yang, F., Chu, P., Deng, G., Yu, S., Harshit, Huang, M., Liu, J., Xu, Y., Liao, C., Yuan, L., & Ling, H. (2020). LaSOT: A high-quality large-scale single object tracking benchmark. arXiv:2009.03465. Accessed 25 April 2022.

Galoogahi, H. K., Fagg, A., Huang, C., Ramanan, D., & Lucey, S. (2017). Need for speed: A benchmark for higher frame rate object tracking. In *2017 IEEE international conference on computer vision (ICCV)* (pp. 1134–1143). IEEE. https://doi.org/10.1109/ICCV.2017.128. http://ieeexplore.ieee.org/document/8237390/ Accessed 29 April 2022.

Girdhar, Y., McGuire, N., Cai, L., Jamieson, S., McCammon, S., Claus, B., Soucie, J. E. S., Todd, J. E., & Mooney, T. A. (2023). CUREE: A curious underwater robot for ecosystem exploration. In *IEEE international conference on robotics and automation (ICRA)* [To appear].

Hanlon, R. T., Forsythe, J. W., & Joneschild, D. E. (1999). Crypsis, conspicuousness, mimicry and polyphenism as antipredator defences of foraging octopuses on indo-pacific coral reefs, with a method of quantifying crypsis from video tapes. *Biological Journal of the Linnean Society, 66*(1), 1–22. https://doi.org/10.1006/bijl.1998.0264.

Hanlon, R. T., & McManus, G. (2020). Flamboyant cuttlefish behavior: Camouflage tactics and complex colorful reproductive behavior assessed during field studies at Lembeh Strait, Indonesia. *Journal of Experimental Marine Biology and Ecology, 529,* 151397. https://doi.org/10.1016/j.jembe.2020.151397.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *2016 IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 770–778). https://doi.org/10.1109/CVPR.2016.90. http://ieeexplore.ieee.org/document/7780459/. Accessed 29 April 2022.

Huang, L., Zhao, X., & Huang, K. (2021). GOT-10k: A large high-diversity benchmark for generic object tracking in the wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 43*(5), 1562–1577. https://doi.org/10.1109/TPAMI.2019.2957464.

Katija, K., Orenstein, E., Schlining, B., Lundsten, L., Barnard, K., Sainz, G., Boulais, O., Cromwell, M., Butler, E., Woodward, B., & Bell, K. C. (2022). FathomNet: A global image database for enabling artificial intelligence in the ocean. arXiv:2109.14646. Accessed 29 April 2022.

Katija, K., Roberts, P. L. D., Daniels, J., Lapides, A., Barnard, K., Risi, M., Ranaan, B. Y., Woodward, B. G., & Takahashi, J. (2021). Visual tracking of deepwater animals using machine learning-controlled robotic underwater vehicles. In *IEEE WACV*

Kristan, M., Leonardis, A., Matas, J., Felsberg, M., Pflugfelder, R., Kamarainen, J.-K., Danelljan, M., Zajc, L. C., Lukezic, A., Drbohlav, O., He, L., Zhang, Y., Yan, S., Yang, J., Fernandez, G., Hauptmann, A., Memarmoghadam, A., Garcia-Martin, A., Robinson, A., Varfolomieiev, A., Gebrehiwot, A. H., Uzun, B., Yan, B., Li, B., Qian, C., Tsai, C.-Y., Micheloni, C., Wang, D., Wang, F., Xie, F., Lawin, F. J., Gustafsson, F., Foresti, G. L., Bhat, G., Chen, G., Ling, H., Zhang, H., Cevikalp, H., Zhao, H., Bai, H., Kuchibhotla, H. C., Saribas, H., Fan, H., Ghanei-Yakhdan, H., Li, H., Peng, H., Lu, H., Li, H., Khaghani, J., Bescos, J., Li, J., Fu, J., Yu, J., Xu, J., Kittler, J., Yin, J., Lee, J., Yu, K., Liu, K., Yang, K., Dai, K., Cheng, L., Zhang, L., Wang, L., Wang, L., Van Gool, L., Bertinetto, L., Dunnhofer, M., Cheng, M., Dasari, M. M., Wang, N., Wang, N., Zhang, P., Torr, P.H.S., Wang, Q., Timofte, R., Gorthi, R. K. S., Choi, S., Marvasti-Zadeh, S. M., Zhao, S., Kasaei, S., Qiu, S., Chen, S., Schön, T. B., Xu, T., Lu, W., Hu, W., Zhou, W., Qiu, X., Ke, X., Wu, X.-J., Zhang, X., Yang, X., Zhu, X., Jiang, Y., Wang, Y., Chen, Y., Ye, Y., Li, Y., Yao, Y., Lee, Y., Gu, Y., Wang, Z., Tang, Z., Feng, Z.-H., Mai, Z., Zhang, Z., Wu, Z., & Ma, Z. (2020). The eighth visual object tracking VOT2020 challenge results. In A. Bartoli, & A. Fusiello (Eds.) *Computer vision—ECCV 2020 workshops* (pp. 547–601). Springer. https://doi.org/10.1007/978-3-030-68238-5_39

Kristan, M., Pflugfelder, R., Leonardis, A., Matas, J., Porikli, F., Cehovin, L., et al. (2013). The visual object tracking VOT2013 challenge results. In *IEEE international conference on computer vision workshops* (pp. 98–111). https://doi.org/10.1109/ICCVW.2013.20.

Kukulya, A. L., Stokey, R., Fiester, C., Padilla, E. M. H., & Skomal, G. (2016). Multi-vehicle autonomous tracking and filming of white sharks carcharodon carcharias. In *2016 IEEE/OES autonomous underwater vehicles (AUV)* (pp. 423–430). https://doi.org/10.1109/AUV.2016.7778707. ISSN: 2377-6536.

Kukulya, A. L., Stokey, R., Littlefield, R., Jaffre, F., Padilla, E. M. H., & Skomal, G. (2015). 3d real-time tracking, following and imaging of white sharks with an autonomous underwater vehicle. In *OCEANS 2015–Genova* (pp. 1–6). https://doi.org/10.1109/OCEANS-Genova.2015.7271546.

Labelbox: The leading training data platform for data labeling. https://labelbox.com/. Accessed 29 April 2022.

Li, B., Wu, W., Wang, Q., Zhang, F., Xing, J., & Yan, J. (2019). SiamRPN++: Evolution of siamese visual tracking with very deep networks. In *2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (pp. 4277–4286). IEEE. https://doi.org/10.1109/CVPR.2019.00441. https://ieeexplore.ieee.org/document/8954116/. Accessed 22 March 2021.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., & Dollar, P. (2017). Focal loss for dense object detection. In *IEEE ICCV*.

Mayer, C., Danelljan, M., Pani Paudel, D., & Van Gool, L. (2021). Learning target candidate association to keep track of what not to track. In *2021 IEEE/CVF international conference on computer vision (ICCV)* (pp. 13424–13434). IEEE. https://doi.org/10.1109/ICCV48922.2021.01319. https://ieeexplore.ieee.org/document/9710884/. Accessed 29 April 2022.

Mittal, V., & Kashyap, I. (2015). Online methods of learning in occurrence of concept drift. *International Journal of Computer Applications, 117*(13), 18–22.

Mooney, T. A. (2020). Biologging ecology and oceanography: Integrative approaches to animal-bourne observations in a changing ocean. In *Ocean sciences meeting 2020*

Mueller, M., Smith, N., & Ghanem, B. (2016). A benchmark and simulator for UAV tracking. In B. Leibe, J. Matas, N. Sebe, & M. Welling (Eds.), *Computer vision–ECCV 2016. Lecture notes in computer science* (pp. 445–461). Springer. https://doi.org/10.1007/978-3-319-46448-0_27.

Müller, M., Bibi, A., Giancola, S., Alsubaihi, S., & Ghanem, B. (2018). TrackingNet: A large-scale dataset and benchmark for object tracking in the wild. In *European conference on computer vision (ECCV)* (vol. 11205, pp. 310–327). https://doi.org/10.1007/978-3-030-01246-5_19. Accessed 23 April 2021.

Nam, H., & Han, B. (2016). Learning multi-domain convolutional neural networks for visual tracking. In *2016 IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 4293–4302). IEEE. https://doi.org/10.1109/CVPR.2016.465. http://ieeexplore.ieee.org/document/7780834/. Accessed 29 April 2022.

OzFish Dataset—Machine learning dataset for baited remote underwater video stations. https://apps.aims.gov.au/metadata/view/38c829d4-6b6d-44a1-9476-f9b0955ce0b8. Accessed 29 April 2022.

Priede, I. G., Drazen, J. C., Bailey, D. M., Kuhnz, L. A., & Fabian, D. (2020). Abyssal demersal fishes recorded at station m (34 50n, 123 00w, 4100 m depth) in the northeast pacific ocean: An annotated check list and synthesis. *Deep Sea Research Part II: Topical Stud-*

ies in Oceanography, 173, 104648. https://doi.org/10.1016/j.dsr2.2019.104648.

ROS: Home. https://www.ros.org/. Accessed 29 April 2022.

Saleh, A., Laradji, I. H., Konovalov, D. A., Bradley, M., Vazquez, D., & Sheaves, M. (2020). A realistic fish-habitat dataset to evaluate algorithms for underwater visual analysis. *Nature Scientific Reports, 10*(1), 14671. https://doi.org/10.1038/s41598-020-71639-x.

Schlining, B. M., & Stout, N. J. (2006). MBARI's video annotation and reference system. In *OCEANS 2006* (pp. 1–5). https://doi.org/10.1109/OCEANS.2006.306879. ISSN: 0197-7385

Tao, R., Gavves, E., & Smeulders, A. W. M. (2016). Siamese instance search for tracking. arXiv:1605.05863. Accessed 29 April 2022.

Valmadre, J., Bertinetto, L., Henriques, J. F., Tao, R., Vedaldi, A., Smeulders, A., Torr, P., & Gavves, E. (2018). Long-term tracking in the wild: A benchmark. in *IEEE ECCV* Accessed 22 March 2021.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems* (vol. 30). Curran Associates, Inc. https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html. Accessed 29 April 2022.

Wang, N., Zhou, W., Wang, J., & Li, H. (2021). Transformer meets tracker: Exploiting temporal context for robust visual tracking. In *2021 IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (pp. 1571–1580). IEEE. https://doi.org/10.1109/CVPR46437.2021.00162. https://ieeexplore.ieee.org/document/9578157/ Accessed 25 April 2022.

Wang, Q., Zhang, L., Bertinetto, L., Hu, W., & Torr, P. H. S. (2018). Fast online object tracking and segmentation: A unifying approach. In *IEEE CVPR*. Accessed 22 March 2021.

Wang, Y., Yu, X., An, D., & Wei, Y. (2021). Underwater image enhancement and marine snow removal for fishery based on integrated dual-channel neural network. *Computers and Electronics in Agriculture, 186,* 106182. https://doi.org/10.1016/j.compag.2021.106182.

Williams, S. B., Pizarro, O., How, M., Mercer, D., Powell, G., Marshall, J., & Hanlon, R. (2009). Surveying noctural cuttlefish camouflage behaviour using an AUV. pp. 214–219. https://doi.org/10.1109/ROBOT.2009.5152868. ISSN: 1050-4729.

Wojke, N., Bewley, A., & Paulus, D. (2017). Simple online and real-time tracking with a deep association metric. arXiv:1703.07402. Accessed 09 January 2021.

Wu, Y., Lim, J., & Yang, M.-H. (2015). Object tracking benchmark. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 37*(9), 1834–1848. https://doi.org/10.1109/TPAMI.2014.2388226.

Xu, N., Yang, L., Fan, Y., Yue, D., Liang, Y., Yang, J., & Huang, T. (2018). YouTube-VOS: A large-scale video object segmentation benchmark. arXiv:1809.03327. Accessed 29 April 2022.

Yoerger, D. R., Govindarajan, A. F., Howland, J. C., Llopiz, J. K., Wiebe, P. H., Curran, M., Fujii, J., Gomez-Ibanez, D., Katija, K., Robison, B. H., Hobson, B. W., Risi, M., & Rock, S. M. (2021). A hybrid underwater robot for multidisciplinary investigation of the ocean twilight zone. In *AAAS science robotics*. American Association for the Advancement of Science. https://doi.org/10.1126/scirobotics.abe1901

Zhu, Z., Wang, Q., Li, B., Wu, W., Yan, J., & Hu, W. (2018). Distractor-aware siamese networks for visual object tracking. arXiv:1808.06048. Accessed 15 March 2021.