UTB180: A High-quality Benchmark for Underwater Tracking

Basit Alawode[1], Yuhang Guo[1], Mehnaz Ummar[1], Naoufel Werghi[1], Jorge Dias[1], Ajmal Mian[2], and Sajid Javed[1]

[1]Department of Electrical Engineering and Computer Science, Khalifa University of Science and Technology, Abu Dhabi, UAE .
[2]The University of Western Australia, Stirling Highway, Australia.
{100060517,sajid.javed}@ku.ac.ae

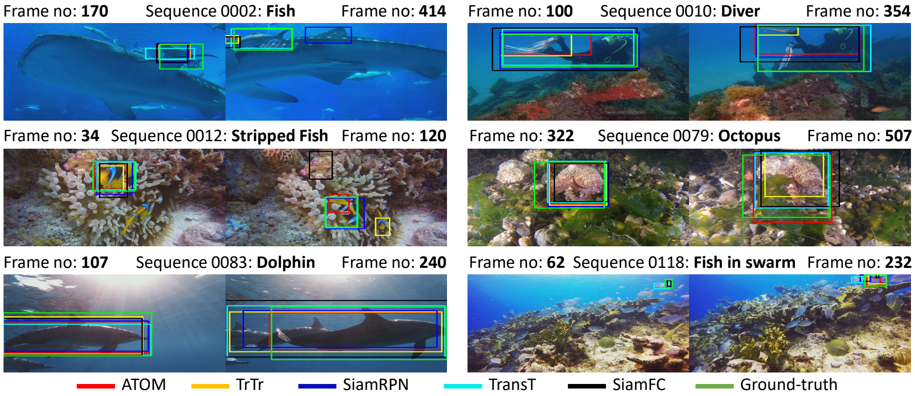


Fig. 1: Sample images of our proposed UTB180 benchmark dataset. The tracking results of some representative State-Of-The-Art (SOTA) trackers including ATOM [8], TrTr [41], SiamRPN [26,25], TransT [5], and SiamFC [1] are shown in terms of bounding boxes. The frame indexes and sequence names are also shown in each row.

Abstract. Deep learning methods have demonstrated encouraging performance on open-air visual object tracking (VOT) benchmarks, however, their strength remains unexplored on underwater video sequences due to the lack of challenging underwater VOT benchmarks. Apart from the open-air tracking challenges, videos captured in underwater environments pose additional challenges for tracking such as low visibility, poor video quality, distortions in sharpness and contrast, reflections from suspended particles, and non-uniform lighting. In the current work, we propose a new Underwater Tracking Benchmark (UTB180) dataset consisting of 180 sequences to facilitate the development of underwater deep trackers. The sequences in UTB180 are selected from both underwater natural and online sources with over 58,000 annotated frames. Video-level attributes are also provided to facilitate the development of robust trackers for specific challenges. We benchmark 15 existing pre-trained

State-Of-The-Art (SOTA) trackers on UTB180 and compare their performance on another publicly available underwater benchmark. The trackers consistently perform worse on UTB180 showing that it poses more challenging scenarios. Moreover, we show that fine-tuning five high-quality SOTA trackers on UTB180 still does not sufficiently boost their tracking performance. Our experiments show that the UTB180 sequences pose a major burden on the SOTA trackers as compared to their open-air tracking performance. The performance gap reveals the need for a dedicated end-to-end underwater deep tracker that takes into account the inherent properties of underwater environments. We believe that our proposed dataset will be of great value to the tracking community in advancing the SOTA in underwater VOT. Our dataset is publicly available on Kaggle.

1 Introduction

Visual Object Tracking (VOT) is the task of estimating the trajectory and state of an arbitrary target object in a video sequence [20]. Given the location of the target in the first frame, the main objective is to learn a robust appearance model to be used when searching for the target object in subsequent frames [1,18]. VOT has numerous open-air applications including autonomous driving, video surveillance, robotics, medical imaging, and sports video analysis [10,16]. In recent years, dominant deep learning trackers such as Siamese [1,26,25], correlation filters [8,2] and transformers [29] have advanced the SOTA performance in tracking. Despite the recent progress, VOT is still an open problem in computer vision because of its challenging nature [16].

Underwater video analysis is an emerging research area where VOT has significant importance in robotics applications including ocean exploration, homeland and maritime security, sea-life monitoring, search and rescue operations to name a few [14,16,4]. Over the years, considerable progress has been made by the tracking community in the development of SOTA end-to-end open-air trackers [16,19,21,22,19,17]. One of the main reasons behind this success is the availability of a variety of large-scale open-air tracking benchmarks such as LASOT [9], GOT-10K [15], and TrackingNet [31] to train, objectively evaluate and compare the different trackers. For instance, as shown in Fig. 2, these datasets exist in small and large scale from a few hundreds of video sequences such as the VOT dataset series [13,23,24], Object Tracking Benchmark (OTB100) [37], Unmanned Aerial Vehicle (UAV) [30], Temple Color (TC) [27] to several thousands of video sequences such as Large-Scale Single Object Tracking (LaSOT) [9], Generic Object Tracking (GOT-10K) [15], and TrackingNet [31]. These datasets provide high quality dense annotations (i.e. per frame) to ensure more accurate evaluations of open-air deep trackers [9,31]. As shown by the average sequence duration (see Fig. 2 and Table 1), they are available for both short-term (average sequence length less than 600 frames) [37,13,23,24] and long-term [30,9] tracking with video specific attributes to further enhance the tracking performance. Furthermore, the large number of video sequences and span variability

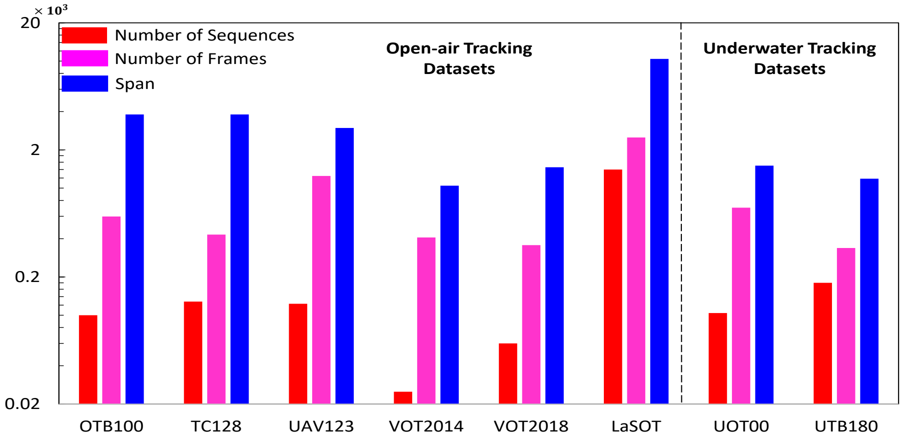


Fig. 2: Summary of open-air and underwater tracking benchmark datasets. Open-air datasets includes OTB100 [37], TC128 [27], UAV123 [30], VOT2014 [13], VOT2018 [24], and LaSOT [9]. Underwater datasets include UOT100 [32] and our proposed UTB180. Span means the difference between the minimum and the maximum number of frames per sequence.

have encouraged the direct training of deep open-air trackers for generic object tracking in these datasets [9,15,31].

All these aforementioned characteristics have immensely contributed towards open-air tracking. However, the same cannot be said for underwater VOT despite its importance. All solutions in this context have simply deployed open-air trackers directly on underwater visual data [32]. One of the major reasons for such stagnation is the unavailability of high-quality benchmarks for underwater tracking exhibiting the challenges of underwater scenes such as of poor visibility, non-uniform lighting conditions, scattering, absorption, blurring of light, flickering of caustic patterns, photometric artifacts, and color variations. To the best of our knowledge, Underwater Object Tracking (UOT100) benchmark is the only available dataset containing 100 underwater sequences covering diverse categories of objects [32]. Our frame-wise evaluation on UOT100 reveals that it belongs to the category of sparsely annotated benchmark datasets using a semi-automatic annotation approach i.e. manual annotations performed every 20 frames and the rest are generated by a tracker. While such an approach speeds up the annotation process, it often yields less accurate ground-truth bounding box predictions due to the propagation and accumulation of the tracker's prediction errors in subsequent frames. All the above motivate us to propose a novel high-quality benchmark dataset UTB180 for the tracking community.

1.1 Contributions

Following are the main contributions of this paper:

1. **Creation of a dense and diversified high-quality Underwater Tracking Benchmark (UTB180) dataset**. Our dataset consists of 180 se-

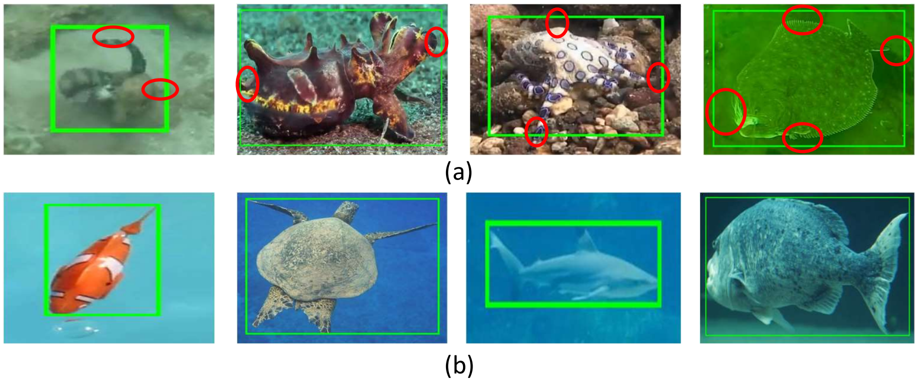


Fig. 3: Zoomed-in samples from the (a) UOT100 underwater benchmark dataset and (b) our proposed UTB180 dataset. The green rectangles show the ground-truth annotations. The red ovals highlight annotation errors in the UOT100 dataset. Our proposed dataset provides more accurate annotations of the target objects compared to UOT100.

quences with over 58,000 annotated frames. Each frame is carefully manually annotated and then visually verified to ensure its correctness. The dataset includes both natural and artificial videos under varying visibility levels sourced from our local marine facilities and several online underwater videos. Additionally, 10 underwater-adapted video-level attributes are also provided to benchmark the tracking performance under various challenges e.g. motion blur and occlusions etc. Although UOT100 has a larger average sequence length and span than UTB180 (Fig. 2), our proposed UTB180 provides more accurate, precise, and reliable annotations (Fig. 3) and a higher number of video sequences.

2. **Benchmarking.** We conducted an extensive benchmarking of 15 high-quality SOTA deep learning trackers on our UTB180 and the UOT100 datasets. Our experiments demonstrate that the majority of the SOTA trackers consistently show lower performance on several underwater challenging attributes revealing the more challenging nature of the proposed UTB180 dataset compared to existing ones (details in Section 4). Visual results comparison of some of the SOTA trackers is shown in Fig. 1 using six sequences captured from our proposed UTB180 dataset.

3. **Fine-tuning recent SOTA trackers on UTB180 benchmark.** We fine-tune five recent SOTA trackers on our dataset and show that performance improvements are small. This experiment demonstrates that there is still a significant performance gap of the existing trackers on underwater data compared to open-air data. This motivates the need to develop specialized end-to-end underwater trackers capable of handling the inherent challenges of underwater scenes.

2 Related Work

In recent years, the tracking community has put significant efforts towards open-air VOT, thanks to the availability of a variety of open-air tracking benchmarks. Since the main objective of the current work relates to underwater benchmarks, we discuss the available underwater datasets in this section. However, we also briefly explain the open-air datasets for comparison and completeness. Table 1 presents the summary of the available open-air and underwater VOT benchmarks. Surprisingly, 9 out of 11 presented datasets are utilized for open-air tracking. This shows the lagging state of underwater VOT.

2.1 Open-Air Tracking Datasets

Several open-air VOT datasets have been proposed in the past decade as shown in Table 1. For instance, Wu *et al.* proposed **OTB100** that consists of 100 videos with 11 tracking attributes with an average frame resolution of 356 × 530 pixels [37]. Liang *et al.* proposed **TC128** to evaluate the significance of color information for tracking [27]. This dataset consists of 128 video sequences with 11 distinct tracking challenges with an average resolution of 461 × 737 pixels. Muller *et al.* proposed **UAV123** dataset for short term tracking [30]. This dataset contains 123 short sequences of nine diverse object categories. The average resolution of each sequences is 1231 × 699 with 12 tracking attributes. The **VOT2014** [13], **VOT2016** [23], and **VOT2018** [24] are the datasets accompanying the VOT challenge competition to benchmark short-term and long-term tracking performance. As described in Table 1, the VOT2014-2018 series contain 25, 60, and 60 sequences and 12 tracking attributes with an average frame resolution of 757×480, 758×465, and 758×465 pixels, respectively. **LaSOT** [9], **GOT-10k** [15], and **TrackingNet** [31] are relatively larger open-air tracking benchmarks. LaSOT contains 1120 training sequences (2.8M frames) and 280 test sequences (685K frames). GOT-10k contains $10,000$ sequences in which $9,340$ are used for training and remaining 420 sequences used for testing purpose. Similarly, TrackingNet contains a total of $30,643$ sequences where $30,130$ sequences are used for training and remaining 511 sequences used for testing. These large-scale datasets also contain 14-16 distinct tracking attributes with average frame resolutions of 632 × 1089, 929 × 1638, and 591 × 1013 respectively. Due to the large diversity in these benchmarks, many SOTA open-air trackers have been entirely trained and tested on these datasets.

2.2 Underwater Tracking Datasets

Compared to open-air tracking benchmarks, underwater tracking datasets are scarcely available. To the best of our knowledge, the UOT100 is the only available underwater tracking benchmark [32]. This dataset comprises 104 underwater sequences selected from YouTube. It contains a total of $74,042$ annotated frames with 702 average number of frames per sequence. The dataset captures a wide variety of underwater distortions and non-uniform lighting conditions. However, this dataset is not sufficiently diverse for generic object tracking in underwater settings. Moreover, it is also sparsely annotated, containing annotation errors that lead to inaccuracies in tracking. In contrast, our proposed UTB180 dataset is more accurate and densely-annotated benchmark for underwater tracking.

Table 1: Summary of the existing open-air and underwater SOTA VOT benchmark datasets and our proposed UTB180 dataset.

Dataset/ Publication	Video Sequences	Attributes	Min Frames	Average Frames	Max Frames	Open- Air	Under- water
OTB100 PAMI2015 [37]	100	11	71	598	3872	✓	
TC128 TIP2015 [27]	128	11	71	431	3872	✓	
UAV123 ECCV2016 [30]	123	12	109	1247	3085	✓	
VOT2014 ECCV-W2014 [13]	25	12	164	409	1210	✓	
VOT2016 ECCV-W2016 [23]	60	12	48		1507	✓	
VOT2018 ICCV-W2018 [24]	60	12	41	356	1500	✓	
LaSOT CVPR2019 [9]	1.4k	14	1000	2506	11397	✓	
GOT-10k PAMI2019 [15]	10k	6	51		920	✓	
TrackingNet ECCV2018 [31]	30.643k	15	96	471	2368	✓	
UOT100 IEEE JOE 2022 [32]	104	3	264	702	1764		✓
Proposed UTB180	**180**	**10**	**40**	**338**	**1226**		✓

3 Proposed High-quality UTB180 Benchmark

In this section, we explain our proposed Underwater Tracking Benchmark (UTB180) dataset in detail including data collection step, bounding box annotation process, and several video-level attributes included with the dataset.

3.1 Dataset

UTB180 consists of 180 videos selected from underwater environments offering dense (i.e. frame by frame), carefully, and manually annotated frames (58K bounding boxes). It spans sequences for both short-term and long-term underwater tracking. The minimum, average, and maximum number of frames per sequence are 40, 338, and 1226, respectively (shown in Table 1). Our dataset also contains a large variety of diverse underwater creature objects including diverse species of fishes (e.g., dwarf lantern shark, jelly fish, juvenile frog fish, cookie cutter shark, bristle mouths, angler fish, viper fish, grass carp, peruvian anchoveta, and silver carp etc.), crab, sea horse, turtle, squid, octopus, and seal. It aims to offer the tracking community a high-quality benchmark for underwater VOT.

3.2 Data Collection

UTB180 has been sourced from several publicly available online sources such as YouTube, pexel [33] and underwater change detection [34]. We also collected sequences from our marine observatory pond, adding thus more diversity to the dataset. The minimum, average, and maximum frame resolution of the sequences are 1520 × 2704, 1080 × 1920, and 959 × 1277 at 30 frames per second.

3.3 Annotation

To annotate target ground-truth bounding boxes in a sequence, each frame undergoes five sequential processes: 1) Rough estimate of the bounding box is done using a Computer Vision Annotation Tool (CVAT) [7], 2) Each bounding box is then manually and carefully examined, afterwards, to ensure accurate and precise bounding box values around each target object, 3) Each bounding box is then further inspected by a validator to ascertain the accurateness. If it fails at this validation step, it is returned to step 2. 4) For each video sequence, its attributes are labeled, and finally, 5) the sequence is validated with the attributes to ascertain the accurateness. Using these steps, we are able to create a high-quality error-free annotated sequences. It should be noted that each bounding box is a rectangle of four values using the format [x, y, w, h], where x and y denotes the top and left coordinates, w and h denotes the width and height of the rectangle, respectively.

3.4 Underwater Tracking Attributes

Attributes, are video content's aspects that are used to better assess the trackers performance on specific challenges. In this work, we have carefully selected 10 underwater-adapted video-level attributes covering most of the essential variations expected in an underwater environment. These attributes are summarized as follows:

- **Unclear Water (UW):** It presents the low visibility tracking challenge indicating if the water is clear or not.
- **Target Scale Variation (SV):** It indicates whether or not the target varies in scale above a certain degree across the frames.

- **Out-of-View (OV):** It indicates that some portion of the target object leaves the scene.
- **Partial Occlusion (PO):** Accounts for partial occlusion of the target by other objects in the scene.
- **Full Occlusion (FO):** Indicates if the target is fully occluded by another object.
- **Deformation (DF):** It tells if the object is deformed probably due to camera angle or view.
- **Low Resolution (LR):** It indicates if a frame is of low resolution typically less than 300 dots per inch (dpi).
- **Fast Motion (FM):** It indicates if the target moves fast across the frames in the sequence.
- **Motion Blur (MB):** This indicates if the target is blurry.
- **Similar Object (SO):** This attribute indicates if there are object(s) similar to the target in the frames.

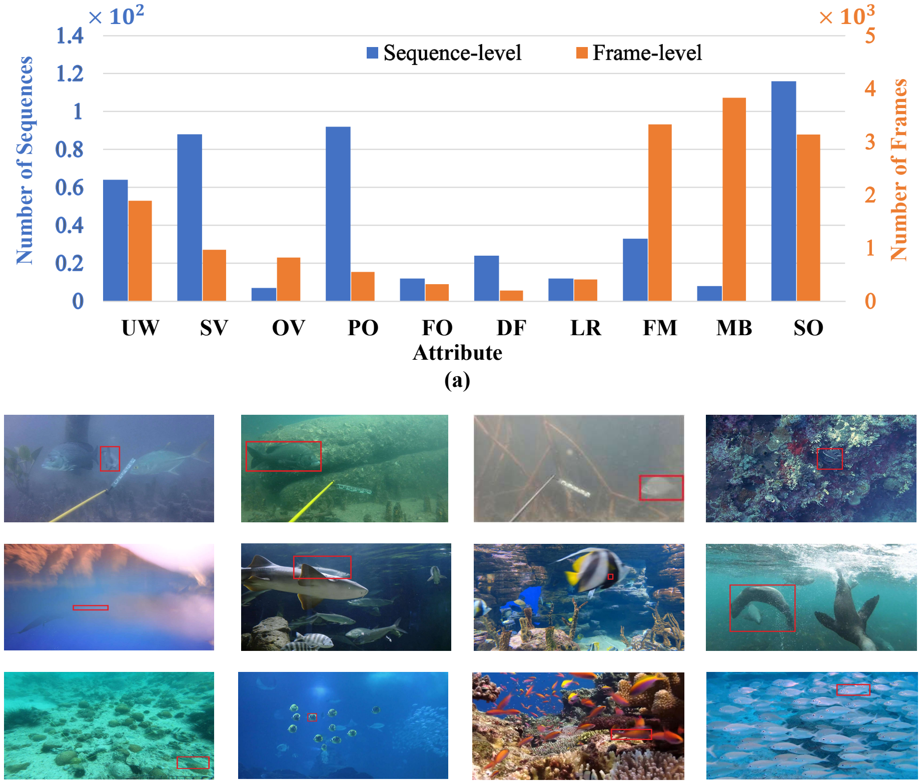
Note that all attributes assume binary values, i.e. 1 (presence) or 0 (absence). An attribute is considered present in a sequence if it is present in at least one frame. The sequence-level and frame-level distributions of the attributes in our proposed UTB180 dataset are shown in Fig. 4(a). Moreover, for each of the attributes, a sample image with red colored ground truth bounding box is also shown in Fig. 4(b) except for the UW attribute which shows three sample images illustrating the diverse and challenging nature of underwater visual data. In the next section, we benchmark and compare several SOTA trackers on the UTB180 dataset.

4 Experimental Evaluations

We evaluate and analyse the performance of existing trackers on our proposed UTB180 dataset and further compare with the publicly available underwater tracking benchmark UOT100 [32]. We also fine-tune 5 high-quality SOTA trackers on a training split of UTB180 dataset to improve their tracking performance. In addition, we analyse the attributes-wise tracking performance to further test the robustness of the trackers on specific challenges. All experiments are conducted on a workstation with a 128 GB of memory, CPU Intel Xeon E5-2698 V4 2.2 Gz (20-cores), and two Tesla V100 GPUs. All the trackers are implemented using the official source codes provided by the respective authors.

4.1 Evaluated Trackers

We evaluated the tracking performance of several popular SOTA deep tracking algorithms. These include end-to-end Discriminative Correlation Filters (DCFs)-based trackers such as ATOM [8], DiMP [2], and KYS [3], deep Siamese-based trackers such as SiamFC [1,38], SiamMask [36], SiamRPN [26,25], SiamCAR [12], DaSiamRPN [40], SiamBAN [6], and SiamGAT [11], and the recently proposed transformer-driven DCFs and Siamese-based trackers such as TrSiam [35], TrDimp [35], TrTr [41], TransT [5], and Stark [39].



(a)

(b)

Fig. 4: Proposed UTB180 dataset statistics and sample images. (a) Statistics of sequence-level and frame-level attributes. (b) Sample images of distinct tracking attributes. From left to right, top row shows UW, UW, UW, and SV tracking challenges. Mid row represents OV, PO, FO, and DF attributes. Bottom row shows sample images involving LR, FM, MB, and SO attributes. A red bounding box shows the ground-truth target object. Three UW attribute samples are shown to illustrate the diverse and challenging nature of underwater visual data.

4.2 Performance Metrics

Following popular tracking protocols developed in open-air tracking datasets e.g. OTB100 [37] and LaSOT [9], we performed the One-Pass Evaluation (OPE) on the benchmarks and measured the precision, normalized precision, and success of different tracking algorithms. The tracking performance metrics are defined as follows:

1. **Precision:** This metric is computed by estimating the distance between a predicted bounding box and a ground-truth bounding box in pixels. Similar to the protocols defined by Wu *et al.* [37], we ranked different trackers using this metric with a threshold of 20 pixels.
2. **Success:** Since the precision metric only measures the localization performance, it does not measure the scale of the predicted bounding boxes in

relation to the ground truth. The success metric takes this into account by employing the intersection over union (IOU) to evaluate the tracker [37]. The IOU is the ratio of the intersection and union of the predicted and the ground truth bounding box. The success plot is then generate by varying the IOU from 0 to 1. The trackers are ranked at a success rate of 0.5.

3. **Normalized Precision:** As the precision metric is sensitive to target size and image resolution, we also used the normalized precision as defined in [31]. With the normalized precision measure, we ranked the SOTA trackers using the area under the curve between 0 to 0.5. More details about this metric can be found in [31].

4.3 Evaluation Protocols

Similar to [9], we used two different protocols and evaluated the SOTA trackers on UTB180 dataset. In **Protocol I**, we used all 180 videos of UTB180 and evaluated the open-air pre-trained models of the SOTA tracking algorithms. This protocol aims to provide large-scale evaluations of the tracking algorithms. In **Protocol II**, we firstly divided the UTB180 into training and testing subsets and then fine-tuned recent SOTA trackers on the training split. Using a 70/30 split, we select 14 out of 20 videos in each category for training and the rest for testing. More specifically, the training subset contains 130 sequences with 41K frames, and the testing subset consists of 50 sequences with 17K frames. The evaluation of SOTA trackers is performed on the testing subset. This protocol aims to provide a large set of underwater videos for training and testing trackers.

4.4 Experiments on Protocol I: Pre-trained Trackers Evaluation

Overall Performance: In this experiment, we benchmark the pre-trained models of the SOTA trackers on the UTB180 and UOT100 [32] datasets. The overall performance in terms of success, normalized precision, and precision is shown in Table 2. Further, the success and precision plots are shown in Fig. 5(first row) and 5(second row) for the UOT100 and UTB180 respectively.

From the results, it can be observed that the Siamese and transformer-driven trackers achieved the best performance on UTB180. Among the compared SOTA trackers, TransT achieved the best results of 58.4% and 51.2% in terms of success and precision rates. In terms of normalized precision rate, SiamBAN achieved the best results of 67.9%. All compared trackers achieved consistently lower performance on all metrics on the UTB180 compared to the UOT100 despite the fact that UTB180 has fewer annotated frames compared to UOT100. The low performance obtained by the SOTA trackers evidenced the novel challenging scenarios in the UTB180 benchmark, and therefore, the need for the development of more powerful underwater trackers.

4.5 Experiments on Protocol II: Fine-tuned Trackers Evaluation

Overall Performance: In this experiment, we investigated the ability of the open-air pre-trained trackers to generalize to underwater dataset. For this pur-

Table 2: Comparative results of pre-trained trackers on UTB180 and UOT100 benchmarks under protocol I. The best three trackers are shown in red, green, and blue colors, respectively.

Tracker	**Sucess ↑**		**Norm Precision ↑**		**Precision ↑**	
	UOT100	**UTB180**	UOT100	**UTB180**	UOT100	**UTB180**
SiamFC [1]	0.438	**0.350**	0.534	**0.412**	0.304	**0.228**
SiamRPN [26,25]	0.597	**0.534**	0.748	**0.635**	0.487	**0.419**
SiamBAN [6]	0.570	**0.562**	0.749	**0.679**	**0.522**	**0.462**
SiamMASK [36]	0.547	**0.523**	0.723	**0.640**	0.467	**0.418**
SiamCAR [12]	0.528	**0.461**	0.665	**0.549**	0.450	**0.389**
DaSiamRPN [40]	0.364	**0.355**	0.411	**0.370**	0.184	**0.180**
ATOM [8]	0.545	**0.477**	0.692	**0.555**	0.444	**0.348**
DiMP [2]	0.568	**0.467**	0.698	**0.529**	0.449	**0.332**
KYS [3]	0.585	**0.529**	0.729	**0.613**	0.480	**0.401**
KeepTrack [28]	**0.609**	**0.543**	**0.779**	**0.637**	0.515	**0.421**
Stark [39]	**0.614**	**0.482**	0.757	**0.542**	**0.532**	**0.400**
TrDiMP [35]	0.599	**0.580**	**0.759**	**0.676**	0.503	**0.455**
TrSiam [35]	0.598	**0.566**	0.752	**0.656**	0.492	**0.438**
TrTr [41]	0.535	**0.500**	0.713	**0.601**	0.486	**0.406**
TransT [5]	**0.624**	**0.584**	**0.789**	**0.672**	**0.555**	**0.512**

pose, we fine-tuned five SOTA trackers including SiamFC, SiamRPN, ATOM, TrTr, and TransT using the training split (130 videos) of UTB180 dataset. We froze the backbone of each pre-trained tracker for feature extraction and fine-tuned their prediction heads. For the most part during fine-tuning, the default training parameters were unchanged except for the learning rate which was reduced. The pre-trained and fine-tuned trackers performance evaluated on the testing split (50 videos) are presented in Table 3. The success and precision plots are also presented in the Fig. 6(first row) and 6(second row) respectively.

Table 3: Comparative results of the pre-trained and fine-tuned trackers on UTB180 benchmark under protocol II. The best two trackers are shown in red and green colors, respectively.

Tracker	**Pretrained ↑**			**Finetuned ↑**		
	Success	**Norm**	**Precision**	**Success**	**Norm**	**Precision**
SiamFC [1]	0.308	0.355	0.287	0.315	0.368	0.294
SiamRPN [26,25]	0.486	**0.568**	0.450	0.491	**0.596**	0.459
ATOM [8]	0.451	0.532	0.460	**0.500**	0.600	**0.516**
TrTr [41]	**0.490**	**0.597**	**0.486**	0.490	**0.605**	0.499
TransT [5]	**0.492**	0.562	**0.508**	**0.494**	0.570	**0.510**

The following conclusions are drawn from this experiment:

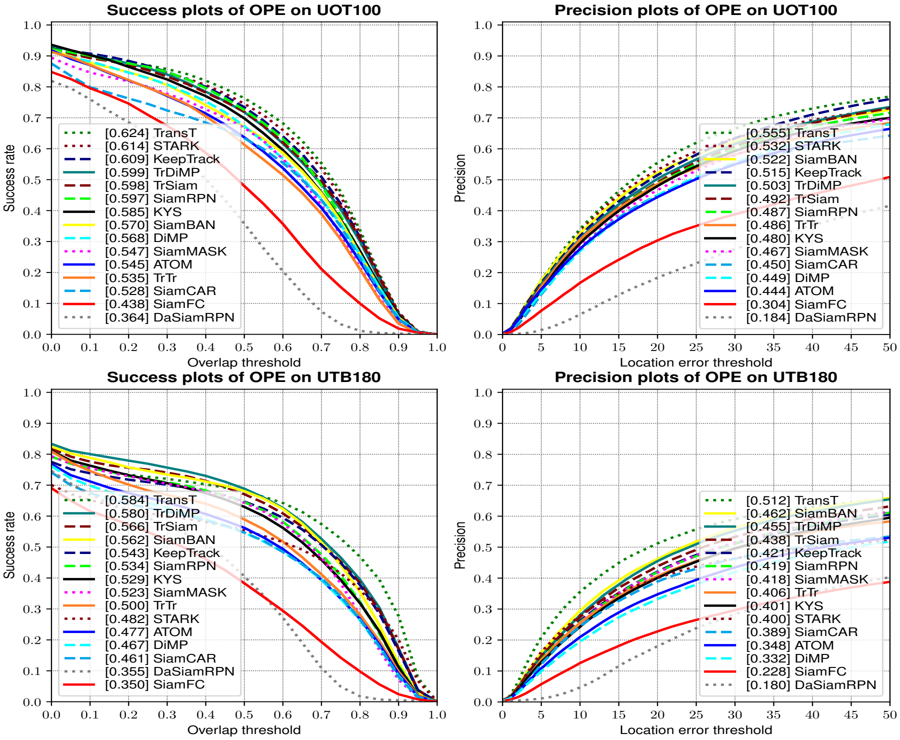


Fig. 5: Evaluation results on UOT100 and UTB180 benchmarks under protocol I using success and precision measures. The legend of precision plot contains threshold scores at 20 pixels, while the legend of success rate contains area-under-the-curve score for each tracker. Overall, the TransT tracker performs better against the SOTA trackers.

1. While fine-tuning the trackers on underwater data slightly improved the tracking performance, it is still not comparable with the performance on open-air data. This suggests that specialized trackers are needed to be developed for underwater applications.
2. While the recent transformer-based trackers such as TrTr and TransT perform better, other trackers benefited more from the fine-tuning. This suggests that with the availability of enough data, trackers can be trained longer to achieve better performance.

4.6 Attribute-wise Evaluation

We also investigated the attribute-wise performance on the UTB180 dataset. We selected a recently proposed TransT tracker since it achieved the best performance shown in Tables 3- 2. We benchmark the TransT on sequences belonging to each of the attributes discussed in section 3.4. Table 4 shows the tracking performance in terms of success, normalized precision, and precision. The attribute-wise performance plots can be found in our supplementary material. It can be

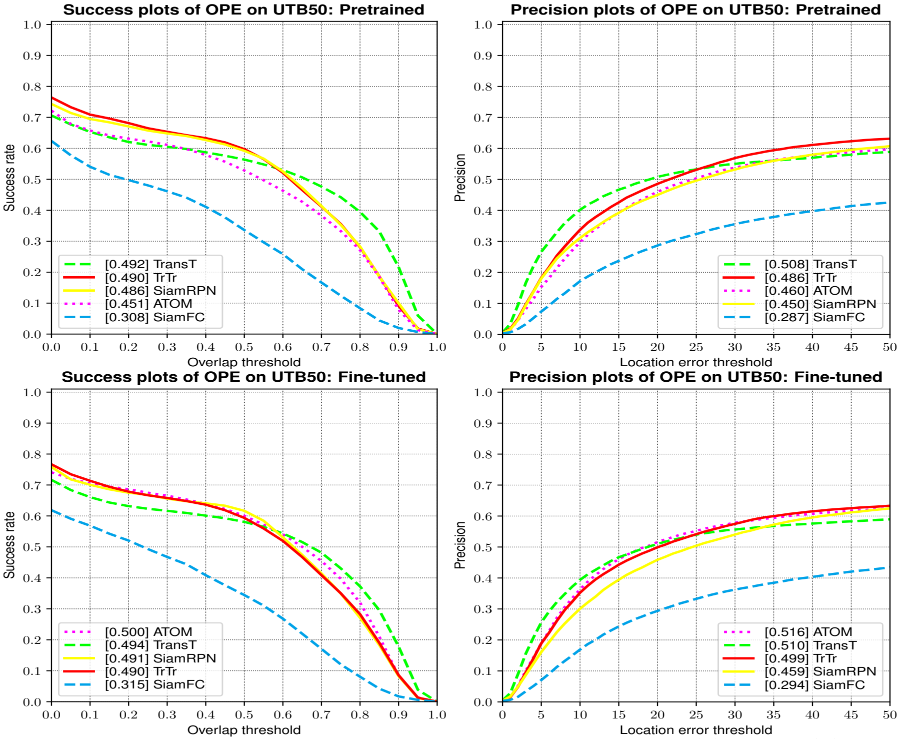


Fig. 6: Evaluation results on UTB50 dataset under protocol I (a) and II (b) using success and precision measure. The legend also contains area under the curve score for each tracker..

observed from the results that each of the attributes tend to degrade the performance when compared to the whole dataset. Overall, TransT achieved the best performance on UW attribute while it could hardly achieve 50% tracking performance on other attributes.

Conclusively, their is still a significant performance gap that needs to be filled for reliable and robust underwater tracking. The difficult target state estimation and several other environmental variations such as low visibility condition make the field of underwater VOT challenging.

5 Conclusion and Future Research Directions

5.1 Conclusion

In this work, we proposed a new VOT dataset dedicated to underwater scenes. It is a dense and diversified high-quality underwater VOT dataset with 180 video sequences and over $58,000$ carefully and manually annotated frames. We benchmarked and fine-tuned existing SOTA Simaese and transformer trackers on the proposed dataset. Our results demonstrate that there is still a significant performance gap between open-air and underwater VOT. We showed that UTB180

Table 4: Attribute-wise performance of pre-trained TransT tracker on UTB180 dataset. The best three trackers are shown in red, green, and blue colors, respectively. More details can be found in the supplementary material.

Attribute	**Acronym**	**Number of Videos**	**Success ↑**	**Norm Precision ↑**	**Precision ↑**
	UTB180	**180**	**0.584**	**0.672**	**0.512**
Unclear Water	UW	64	**0.636**	**0.743**	**0.586**
Scale Variation	SV	88	0.559	0.640	**0.478**
Out of View	OV	7	**0.566**	**0.660**	0.475
Partial Occlusion	PO	92	0.475	0.542	0.409
Full Occlusion	FO	12	0.342	0.375	0.330
Deformation	DF	24	**0.564**	**0.657**	0.402
Low Resolution	LR	12	0.489	0.583	0.390
Fast Motion	FM	33	0.515	0.593	**0.486**
Motion Blur	MB	8	0.485	0.540	0.417
Similar Objects	SO	116	0.513	0.583	0.472

presents more challenging sequences compared to the publicly available UOT100 dataset. It is expected that UTB180 will play an instrumental role in boosting the underwater VOT research.

5.2 Future Research Directions

When compared to open-air, the available underwater datasets are still insignificant. At the moment, the available underwater datasets can only be utilized for benchmarking and fine-tuning of the designed trackers. They are insufficient for the direct training of deep trackers. As such, we intend to extend this work to enable not only fine-tuning but also the training and testing of deep underwater trackers with underwater datasets.

From our experiments, we showed that recent transformer-based trackers consistently performed better than their DCFs and Siamese-based counterparts. While this performance still lags compared to the open-air performance, it suggests that variants of transformer-based trackers could pave the way for the development of better underwater trackers. Improved backbone feature extraction, sophisticated target state estimation, and the role of implicit or explicit underwater video denoising approaches are required for robust end-to-end underwater VOT. Such extensions could lead to more generic algorithms suited for both open-air and underwater VOT.

Acknowledgments This publication acknowledges the support provided by the Khalifa University of Science and Technology under Faculty Start Up grants FSU-2022-003 Award No. 84740 0 0401.

References

1. Bertinetto, L., Valmadre, J., Henriques, J.F., Vedaldi, A., Torr, P.H.: Fully-convolutional siamese networks for object tracking. Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics) **9914 LNCS**, 850–865 (2016). https://doi.org/10.1007/978-3-319-48881-3_56
2. Bhat, G., Danelljan, M., Van Gool, L., Timofte, R.: Learning discriminative model prediction for tracking. Proc. IEEE Int. Conf. Comput. Vis. **2019-Octob**(Iccv), 6181–6190 (2019). https://doi.org/10.1109/ICCV.2019.00628
3. Bhat, G., Danelljan, M., Van Gool, L., Timofte, R.: Know Your Surroundings: Exploiting Scene Information for Object Tracking. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.M. (eds.) Comput. Vis. – ECCV 2020. pp. 205–221. Springer International Publishing, Cham (2020)
4. Boudiaf, A., Guo, Y., Ghimire, A., Werghi, N., De Masi, G., Javed, S., Dias, J.: Underwater image enhancement using pre-trained transformer. In: International Conference on Image Analysis and Processing. pp. 480–488. Springer (2022)
5. Chen, X., Yan, B., Zhu, J., Wang, D., Yang, X., Lu, H.: Transformer Tracking pp. 8122–8131 (2021). https://doi.org/10.1109/cvpr46437.2021.00803
6. Chen, Z., Zhong, B., Li, G., Zhang, S., Ji, R.: Siamese Box Adaptive Network for Visual Tracking. Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. pp. 6667–6676 (2020). https://doi.org/10.1109/CVPR42600.2020.00670
7. CVAT: Computer Vision Annotation Tool, https://cvat.org
8. Danelljan, M., Van Gool, L., Timofte, R.: Probabilistic regression for visual tracking. Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. pp. 7181–7190 (2020). https://doi.org/10.1109/CVPR42600.2020.00721
9. 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.: LaSOT: A High-quality Large-scale Single Object Tracking Benchmark. Int. J. Comput. Vis. **129**(2), 439–461 (2021). https://doi.org/10.1007/s11263-020-01387-y, https://doi.org/10.1007/s11263-020-01387-y
10. Giraldo, J.H., Javed, S., Bouwmans, T.: Graph moving object segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence (2020)
11. Guo, D., Shao, Y., Cui, Y., Wang, Z., Zhang, L., Shen, C.: Graph Attention Tracking. Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. pp. 9538–9547 (2021). https://doi.org/10.1109/CVPR46437.2021.00942
12. Guo, D., Wang, J., Cui, Y., Wang, Z., Chen, S.: SiamCAR: Siamese Fully Convolutional Classification and Regression for Visual Tracking. Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. pp. 6268–6276 (2020). https://doi.org/10.1109/CVPR42600.2020.00630
13. Hadfield, S.J., Lebeda, K., Bowden, R.: The Visual Object Tracking VOT2014 challenge results (2014)
14. Han, M., Lyu, Z., Qiu, T., Xu, M.: A Review on Intelligence Dehazing and Color Restoration for Underwater Images. IEEE Trans. Syst. Man, Cybern. Syst. **50**(5), 1820–1832 (may 2020). https://doi.org/10.1109/TSMC.2017.2788902
15. Huang, L., Zhao, X., Huang, K.: Got-10k: A large high-diversity benchmark for generic object tracking in the wild. IEEE Trans. Pattern Anal. Mach. Intell. **43**(5), 1562–1577 (2021). https://doi.org/10.1109/TPAMI.2019.2957464, https://www.scopus.com/inward/record.uri?eid=2-s2.0-85103976016{&}doi=10.1109{%}2FTPAMI.2019.2957464{&}partnerID=40{&}md5=3fd7d1e870e60df363a83a52a092c544

16. Javed, S., Danelljan, M., Khan, F.S., Khan, M.H., Felsberg, M., Matas, J.: Visual Object Tracking with Discriminative Filters and Siamese Networks: A Survey and Outlook **14**(8), 1–20 (2021), http://arxiv.org/abs/2112.02838
17. Javed, S., Dias, J., Werghi, N.: Low-rank tensor tracking. In: 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW). pp. 605–614 (2019). https://doi.org/10.1109/ICCVW.2019.00074
18. Javed, S., Mahmood, A., Dias, J., Seneviratne, L., Werghi, N.: Hierarchical spatiotemporal graph regularized discriminative correlation filter for visual object tracking. IEEE Transactions on Cybernetics (2021)
19. Javed, S., Mahmood, A., Dias, J., Werghi, N.: Robust structural low-rank tracking. IEEE Transactions on Image Processing **29**, 4390–4405 (2020)
20. Javed, S., Mahmood, A., Ullah, I., Bouwmans, T., Khonji, M., Dias, J.M.M., Werghi, N.: A novel algorithm based on a common subspace fusion for visual object tracking. IEEE Access **10**, 24690–24703 (2022)
21. Javed, S., Zhang, X., Dias, J., Seneviratne, L., Werghi, N.: Spatial graph regularized correlation filters for visual object tracking. In: International Conference on Soft Computing and Pattern Recognition. pp. 186–195. Springer (2020)
22. Javed, S., Zhang, X., Seneviratne, L., Dias, J., Werghi, N.: Deep bidirectional correlation filters for visual object tracking. In: 2020 IEEE 23rd International Conference on Information Fusion (FUSION). pp. 1–8. IEEE (2020)
23. Kristan, M., Leonardis, A., Matas, e.a.: The Visual Object Tracking VOT2016 Challenge Results. pp. 777–823. Springer International Publishing (2016)
24. Kristan, M., Leonardis, A., Matas, e.: The Sixth Visual Object Tracking VOT2018 Challenge Results. In: Leal-Taixé, L., Roth, S. (eds.) Comput. Vis. – ECCV 2018 Work. pp. 3–53. Springer International Publishing, Cham (2019)
25. Li, B., Wu, W., Wang, Q., Zhang, F., Xing, J., Yan, J.: SIAMRPN++: Evolution of siamese visual tracking with very deep networks. Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. **2019-June**, 4277–4286 (2019). https://doi.org/10.1109/CVPR.2019.00441
26. Li, B., Yan, J., Wu, W., Zhu, Z., Hu, X.: High Performance Visual Tracking with Siamese Region Proposal Network. IEEE Conf. Comput. Vis. Pattern Recognit. pp. 8971–8980 (2018)
27. Liang, P., Blasch, E., Ling, H.: Encoding color information for visual tracking: Algorithms and benchmark. IEEE transactions on image processing **24**(12), 5630–5644 (2015)
28. Mayer, C., Danelljan, M., Pani Paudel, D., Van Gool, L.: Learning Target Candidate Association to Keep Track of What Not to Track (Iccv), 13424–13434 (2022). https://doi.org/10.1109/iccv48922.2021.01319
29. Meinhardt, T., Kirillov, A., Leal-Taixé, L., Feichtenhofer, C.: Trackformer: Multi-object tracking with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8844–8854 (June 2022)
30. Mueller, M., Smith, N., Ghanem, B.: A benchmark and simulator for UAV tracking. Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics) **9905 LNCS**, 445–461 (2016). https://doi.org/10.1007/978-3-319-46448-0_27, https://www.scopus.com/inward/record.uri?eid=2-s2.0-84990050293{&}doi=10.1007{%}2F978-3-319-46448-0{_}27{&}partnerID=40{&}md5=8114acc8d0b92dd954d3703034a3fac7
31. Müller, M., Bibi, A., Giancola, S., Alsubaihi, S., Ghanem, B.: TrackingNet: A Large-Scale Dataset and Benchmark for Object Tracking in the Wild. Lect. Notes Comput. Sci. (including Subser. Lect.

Notes Artif. Intell. Lect. Notes Bioinformatics) **11205 LNCS**, 310–327 (2018). https://doi.org/10.1007/978-3-030-01246-5_19, https://www.scopus.com/inward/record.uri?eid=2-s2.0-85055120586{&}doi=10.1007{%}2F978-3-030-01246-5{_}19{&}partnerID=40{&}md5=aa046aa39b63637f50ed76bf2016ce3d

32. Panetta, K., Kezebou, L., Oludare, V., Agaian, S.: Comprehensive Underwater Object Tracking Benchmark Dataset and Underwater Image Enhancement with GAN. IEEE J. Ocean. Eng. **47**(1), 59–75 (2022). https://doi.org/10.1109/JOE.2021.3086907
33. Pexel: 1,363+ Best Free Underwater 4K Stock Video Footage & Royalty-Free HD Video Clips, https://www.pexels.com/search/videos/underwater/
34. Underwaterchangedetection: Videos - Underwaterchangedetection, http://underwaterchangedetection.eu/Videos.html
35. Wang, N., Zhou, W., Wang, J., Li, H.: Transformer Meets Tracker: Exploiting Temporal Context for Robust Visual Tracking. Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. pp. 1571–1580 (2021). https://doi.org/10.1109/CVPR46437.2021.00162
36. Wang, Q., Zhang, L., Bertinetto, L., Hu, W., Torr, P.H.: Fast online object tracking and segmentation: A unifying approach. Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. **2019-June**, 1328–1338 (2019). https://doi.org/10.1109/CVPR.2019.00142
37. Wu, Y., Lim, J., Yang, M.H.: Object tracking benchmark. IEEE Trans. Pattern Anal. Mach. Intell. **37**(9), 1834–1848 (2015). https://doi.org/10.1109/TPAMI.2014.2388226
38. Xu, Y., Wang, Z., Li, Z., Yuan, Y., Yu, G.: SiamFC++: Towards robust and accurate visual tracking with target estimation guidelines. AAAI 2020 - 34th AAAI Conf. Artif. Intell. pp. 12549–12556 (2020). https://doi.org/10.1609/aaai.v34i07.6944
39. Yan, B., Peng, H., Fu, J., Wang, D., Lu, H.: Learning Spatio-Temporal Transformer for Visual Tracking pp. 10428–10437 (2022). https://doi.org/10.1109/iccv48922.2021.01028
40. Yu, Y., Xiong, Y., Huang, W., Scott, M.R.: Deformable Siamese Attention Networks for Visual Object Tracking. Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. pp. 6727–6736 (2020). https://doi.org/10.1109/CVPR42600.2020.00676
41. Zhao, M., Okada, K., Inaba, M.: TrTr: Visual Tracking with Transformer (2021), http://arxiv.org/abs/2105.03817