CDTB: A Color and Depth Visual Object Tracking Dataset and Benchmark

Alan Lukežič[1], Ugur Kart[2], Jani Käpylä[2], Ahmed Durmush[2], Joni-Kristian Kämäräinen[2], Jiří Matas[3] and Matej Kristan[1]

[1]Faculty of Computer and Information Science, University of Ljubljana, Slovenia

[2]Computing Sciences, Tampere University, Finland

[3]Faculty of Electrical Engineering, Czech Technical University in Prague, Czech Republic

alan.lukezic@fri.uni-lj.si

Abstract

We propose a new color-and-depth general visual object tracking benchmark (CDTB). CDTB is recorded by several passive and active RGB-D setups and contains indoor as well as outdoor sequences acquired in direct sunlight. The CDTB dataset is the largest and most diverse dataset for RGB-D tracking, with an order of magnitude larger number of frames than related datasets. The sequences have been carefully recorded to contain significant object pose change, clutter, occlusion, and periods of long-term target absence to enable tracker evaluation under realistic conditions. Sequences are per-frame annotated with 13 visual attributes for detailed analysis. Experiments with RGB and RGB-D trackers show that CDTB is more challenging than previous datasets. State-of-the-art RGB trackers outperform the recent RGB-D trackers, indicating a large gap between the two fields, which has not been detected by the prior benchmarks. Based on the results of the analysis we point out opportunities for future research in RGB-D tracker design.

1. Introduction

Visual object tracking has been enjoying a significant interest of the research community for over several decades due to scientific challenges it presents and its large practical potential. In its most general formulation, it addresses localization of an arbitrary object in all frames of a video, given a single annotation specified in one frame. This is a challenging task of self-supervised learning, since a tracker has to localize and carefully adapt to significant target appearance changes, cope with ambient changes, clutter, and detect occlusion and target disappearance. As such, general object trackers cater a range of applications and research challenges like surveillance systems, video editing, sports analytics and autonomous robotics.

Fuelled by emergence of tracking benchmarks [40, 44, 27, 25, 37, 36] that facilitate objective comparison of different approaches, the field has substantially advanced in the last decade. Due to a wide adoption of RGB cameras, the benchmarks have primarily focused on color (RGB) trackers and trackers that combine color and thermal (infrared) modalities [28, 26, 23, 24].

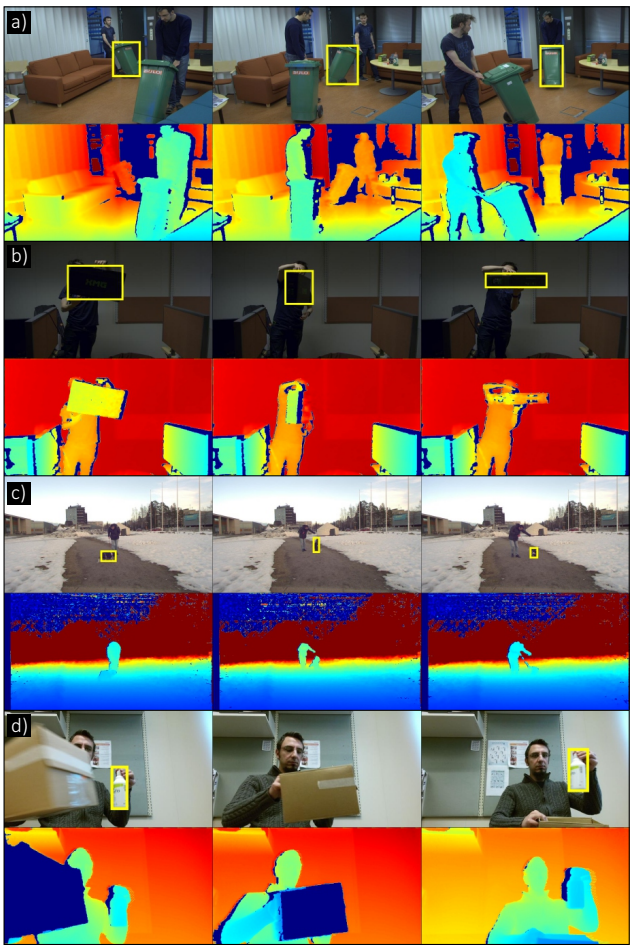

Figure 1. RGB and depth sequences from CDTB. Depth offers a complementary information to color: two identical objects are easier to distinguish in depth (a), low illumination scenes (b) are less challenging for trackers if depth information is available, tracking a deformable object in depth simplifies the problem (c) and a sudden significant change in depth is a strong clue for occlusion (d). Sequences (a,b) are captured by a ToF-RGB pair of cameras, (c) by s tereo-camera sensor and (d) by a Kinect sensor.

Only recently various depth sensors like RGB-D, time-

of-flight (ToF) and LiDAR have become widely accessible. Depth provides an important cue for tracking since it simplifies reasoning about occlusion and offers a better object-to-background separation compared to only color. In addition, depth is a strong cue to acquire object 3D structure and 3D pose without a prior 3D model, which is crucial in research areas like robotic manipulation [5]. The progress in RGB-D tracking has been boosted by the emergence of RGB-D benchmarks [41, 45], but the field significantly lags behind the advancements made in RGB-only tracking.

One reason for the RGB – RGB-D general object tracking performance gap is that existing RGB-D benchmarks [41, 45] are less challenging than their RGB counterparts. The sequences are relatively short from the perspective of practical applications, the objects never leave and re-enter the field of view, they undergo only short-term occlusions and rarely significantly rotate away from the camera. The datasets are recorded indoor only with Kinect-like sensors which prohibits generalization of the results to general outdoor setups. These constraints were crucial for early development of the field, but further boosts require a more challenging benchmark, which is the topic of this paper.

In this work we propose a new color-and-depth tracking benchmark (CDTB) that makes several contributions to the field of general object RGB-D tracking. (i) The CDBT dataset is recorded by several color-and-depth sensors to capture a wide range of depth signals. (ii) The sequences are recorded indoor as well as outdoor to extend the domain of tracking setups. (iii) The dataset contains significant object pose changes to encompass depth appearance variability from real-world tracking environment. (iv) The objects are occluded or leave the field of view for longer duration to emphasize the importance of trackers being able to report target loss and perform re-detection. (v) We compare several state-of-the-art RGB-D trackers as well as state-of-the-art RGB trackers and their RGB-D extensions. Examples of CDTB dataset are shown in Figure 1.

The reminder of the paper is structured as follows. Section 2 summarizes the related work, Section 3 details the acquisition and properties of the dataset, Section 4 summarizes the performance measures, Section 5 reports experimental results and Section 6 concludes the paper.

2. Related work

RGB-D Benchmarks. The diversity of the RGB-D datasets is limited compared to those in RGB tracking. Many of the datasets are application specific, e.g., *pedestrian tracking* or *hand tracking*. For example, Ess *et al*. [11] provide five 3D bounding box annotated sequences captured by a calibrated stereo-pair, the RGB-D People Dataset [42] contains a single sequence of pedestrians in a hallway captured by a static RGB-D camera and Stanford Office [8] contains 17 sequences with a static and one with a moving Kinect. Garcia-Hernando *et al*. [13] introduce an RGB-D dataset for hand tracking and action recognition. Another important application field for RGB-D cameras is robotics, but here datasets are often small and the main objective is real-time model-based 3D pose estimation. For example, the RGB-D Object Pose Tracking Dataset [7] contains 4 synthetic and 2 real RGB-D image sequences to benchmark visual tracking and 6-DoF pose estimation. Generating synthetic data has become popular due to requirements of large training sets for deep methods [39], but it is unclear how well these predict real world performance.

Only two datasets are dedicated to general object tracking. The most popular is Princeton Tracking Benchmark (PTB) [41], which contains 100 RGB-D video sequences of rigid and nonrigid objects recorded with Kinect. The choice of sensor constrains the dataset to only indoor scenarios. The dataset diversity is further reduced since many sequences share the same tracked objects and the background. More than half of the sequences are people tracking. The sequences are annotated by five global attributes. The RGB and depth channels are poorly calibrated. In approximately 14% of sequences the RGB and D channels are not synchronized and approximately 8% are miss-aligned. The calibration issues were addressed by Bibi et al [3] who published a corrected dataset. PTB addresses long-term tracking, in which the tracker has to detect target loss and perform re-detection. The dataset thus contains several full occlusions, but the target never leaves and re-enters the field of view, thus limiting the evaluation capabilities of re-detecting trackers. Performance is evaluated as the percentage of frames in which the bounding box predicted by tracker exceeds a 0.5 overlap with the ground truth. The overlap is artificially set to 1 when the tracker accurately predicts target absence. Recent work in long-term tracker performance evaluation [43, 33] argue against using a single threshold and [33] further show reduced interpretation strength of the measure used in PTB.

The Spatio-Temporal Consistency dataset (STC) [45] was recently proposed to address the drawbacks of PTB. The dataset is recorded by Asus Xtion RGB-D sensor, which also constrains the dataset to only indoor scenarios and a few low-light outside scenarios, but care has been taken to increase the sequence diversity. The dataset is smaller than PTB, containing only 36 sequences, but annotated by thirteen global attributes. STC addresses short-term tracking scenario, i.e., trackers are not required to perform re-detection. Thus the sequences are relatively short and the short-term performance evaluation methodology is used. This makes the dataset inappropriate for evaluating trackers useful in many practical setups, in which target loss detection and redetection are crucial capabilities.

RGB Trackers. Recent years have seen a surge in Short-term Trackers (ST) and especially Discriminative Correla-

tion Filter (DCF) based approaches have been popular due to their mathematical simplicity and elegance. In their seminal paper, Bolme *et al*. [4] proposed using DCF for visual object tracking. Henriques *et al*. [16] proposed an efficient training method by exploiting the properties of circular convolution. Lukezic *et al*. [32] and Galoogahi *et al*. [12] proposed a mechanism to handle boundary problems and segmentation-based DCF constraints have been introduced in [32]. Danelljan *et al*. [10] used a factorized convolution operator and achieved excellent scores on well-known benchmarks.

As a natural extension of the ST, Long-term Trackers (LT) have been proposed [18] where the tracking is decomposed into short-term tracking and long-term detection. Lukezic *et al*. proposed a fully-correlational LT [31] by storing multiple correlation filters that are trained at different time scales. Zhang *et al*. [46] used deep regression and verification networks and they achieved the top rank in VOT-LT 2018 [25]. Despite being published as an ST, MDNet [38] has proven itself as an efficient LT. MDNet uses discriminatively trained Convolutional Neural Networks(CNN) and won the VOT 2015 challenge [26].

RGB-D Trackers. Compared to RGB trackers, the body of literature on RGB-D trackers is rather limited which can be attributed to the lack of available datasets until recently. In 2013, the publication of PTB [41] ignited the interest in the field and there have been numerous attempts by adopting different approaches. The authors of PTB have proposed multiple baseline trackers which use different combinations of HOG [9], optical flow and point clouds. As a part of particle filter tracker family, Meshgi *et al*. [34] proposed a particle filter framework with occlusion awareness using a latent occlusion flag. They pre-emptively predict the occlusions, expand the search area in case of occlusions. Bibi *et al*. [3] represented the target by sparse, part-based 3-D cuboids while adopting particle filter as their motion model. Hannuna *et al*. [14], An *et al*. [1] and Camplani *et al*. [6] extended the Kernelized Correlation Filter (KCF) RGB tracker [16] by adding the depth channel. Hannuna *et al*. and Camplani *et al*. proposed a fast depth image segmentation which is later used for scale, shape analysis and occlusion handling. An *et al*. proposed a framework where the tracking problem is divided into detection, learning and segmentation. To use depth inherently in DCF formulation, Kart *et al*. [20] adopted Gaussian foreground masks on depth images in CSRDCF [32] training. They later extended their work by using a graph cut method with color and depth priors for the foreground mask segmentation [19] and more recently proposed a view-specific DCF using object's 3D structure based masks [21]. Liu *et al*. [30] proposed a 3D mean-shift tracker with occlusion handling. Xiao *et al*. [45] introduced a two-layered representation of the target by adopting a spatio-temporal consistency constraints.

3. Color and depth tracking dataset

We used several RGB-D acquisition setups to increase the dataset diversity in terms of acquisition hardware. This allowed unconstrained indoor as well as outdoor sequence acquisition, thus diversifying the dataset and broaden the scope of scenarios from real-word tracking environment. The following three acquisition setups were used: (i) RGB-D sensor (Kinect), (ii) time-of-flight (ToF)-RGB pair and (iii) stereo cameras pair. The setups are described in the following.

RGB-D Sensor sequences were captured with a Kinect v2 that outputs 24-bit 1920×1080 RGB images (8-bit per color channel) and 512×424 32-bit floating point depth images with an average frame rate of 30 fps. JPEG compression is applied to RGB frames while depth data is converted into 16-bit unsigned integer and saved in PNG format. The RGB and depth images are synchronized internally and no further synchronization was required.

ToF-RGB pair consists of Basler tof640-20gm time-of-flight and Basler acA1920-50gc color cameras. The ToF camera has 640x480pix resolution and maximum 20 fps frame rate whereas color camera has 1920x1200pix resolution and 50 fps maximum frame rate at full resolution. Both cameras can be triggered externally using the I/O's of the cameras for external synchronisation. The cameras were mounted on a high precision CNC-machined aluminium base in a way that the baseline of the cameras are 75.2mm and camera sensor center points are on the same level. The TOF camera has built in optics with $57° \times 43°$ (HxV) field-of-view. The color camera was equipped with a 12mm focal length lens (VS-1214H1), which has $56.9° \times 44°$ (HxV) field-of-view for 1" sensors, to match the field-of-view of the ToF camera. The cameras were synchronised by an external triggering device at the rate of 20 fps. The color camera output was 8-bit raw Bayer images whereas ToF camera output was 16-bit depth images. The raw Bayer images were later debayered to 24-bit RGB images (8-bit per color channel).

Stereo-cameras pair is composed of two Basler acA1920-50gc color cameras which are mounted on a high precision machined aluminium base with 70mm baseline. The cameras were equipped with 6mm focal length lenses (VS-0618H1) with $98.5° \times 77.9°$ (HxV) field-of-view for 1" sensors. The cameras were synchronised by an external triggering device at the rate of 40 fps at full resolution. The camera outputs were 8-bit raw Bayer images which were later Bayer demosaiced to 24-bit RGB images (8-bit per color channel). A semi-global block matching algorithm [17] was applied to the rectified stereo images and converted to metric depth values using the camera calibration parameters.

3.1. RGB and Depth Image Alignment

All three acquisition setups were calibrated using the Caltech Camera Calibration Toolbox[1] with standard modifications to cope with image pairs of different resolution for the RGB-D sensor and ToF-RGB-pair setups. The calibration provides the external camera parameters, *rotation matrix* $\boldsymbol{R}_{3\times3}$ and *translation vector* $\boldsymbol{t}_{3\times1}$, and the intrinsic camera parameters, *focal length* $\boldsymbol{f}_{2\times1}$, *principal point* $\boldsymbol{c}_{2\times1}$, *skew* α and lens distortion coefficients $\boldsymbol{k}_{5\times1}$. The forward projection is defined by [15]

$$\boldsymbol{m} = \mathcal{P}(\boldsymbol{x}) = (\mathcal{P}_c \circ \mathcal{R})(d), \tag{1}$$

where $\boldsymbol{x} = (x, y, z)^T$ is the scene point in world coordinates, $\boldsymbol{m}$ is the projected point in image coordinates and $d = I_{depth}(\boldsymbol{m})$ is the depth. $\mathcal{R}$ is a rigid Euclidean transformation, $\boldsymbol{x}_c = \mathcal{R}(\boldsymbol{x})$, defined by $\boldsymbol{R}$ and $\boldsymbol{t}$, and $\mathcal{P}_c$ is the intrinsic operation $\mathcal{P}_c(\boldsymbol{x}_c) = (\mathcal{K} \circ \mathcal{D} \circ \hat{\nu})(\boldsymbol{x}_c)$ of the perspective division operation $\hat{\nu}$, distortion operation $\mathcal{D}$ using $\boldsymbol{k}$ and the affine mapping $\mathcal{K}$ of $\boldsymbol{f}$ and α.

The depth images of RGB-D Sensor and ToF-RGB pair were per-pixel aligned to the RGB images as follows. A 3D point corresponding to each pixel in the calibrated depth image was computed using the inverse of (1) as $\boldsymbol{x} = \mathcal{P}^{-1}(\boldsymbol{m}, d)$. These points were projected to the RGB image and a linear interpolation model was used to estimate missing per-pixel-aligned re-projected depth values. For further studies we provide the original data and calibration parameters upon request.

3.2. Sequence Annotation

The VOT Aibu image sequence annotator[2] was used to manually annotate the targets by axis-aligned bounding boxes. The bounding boxes were placed following the VOT [28] definition by maximizing the number of target pixels within the bounding box and minimizing their number outside the bounding box. All bounding boxes were checked by several annotators for quality control. In case of a disagreement the authors consolidated and reached an agreement on annotation.

All sequences were annotated per-frame with thirteen attributes. We selected standard attributes for short-term tracking (partial occlusion, deformable target, similar targets, out-of-plane rotation, fast motion and target size change) and for the long-term tracking (target out-of-view and full occlusion). We additionally included RGBD tracking-specific attributes (reflective target, dark scene and depth change). The following attributes were manually annotated: (i) target out-of-view, (ii) full occlusion, (iii) partial occlusion, (iv) out-of-plane rotation, (v) similar objects, (vi) deformable target, (vii) reflective target and (viii) dark scene. The attribute (ix) fast motion was assigned to a frame in which the target center moves by at least 30% of its size in consecutive frames, (x) target size change was assigned when the ratio between maximum and minimum target size in 21 consecutive frames[3] was larger than 1.5 and (xi) aspect ratio change was assigned when the ratio between the maximum and minimum aspect (i.e., width / height) within 21 consecutive frames was larger than 1.5. The attribute (xii) depth change was assigned when the ratio between maximum and minimum of median of depth within target region in 21 consecutive frames was larger than 1.5. Frames not annotated with any of the first twelve attributes were annotated as (xiii) unassigned.

[1] http://www.vision.caltech.edu/bouguetj/calib_doc

[2] https://github.com/votchallenge/aibu

4. Performance Evaluation Measures

Tracker evaluation in a long-term tracking scenario in which targets may disappear/re-appear, requires measuring the localization accuracy, as well as re-detection capability and ability to report that target is not visible. To this end we adopt the recently proposed long-term tracking evaluation protocol from [33], which is used in the VOT2018 long-term challenge [25]. The tracker is initialized in the first frame and left to run until the end of the sequence without intervention.

The implemented performance measures are tracking precision (Pr) and recall (Re) from [33]. Tracking precision measures the accuracy of target localization when deemed visible, while tracking recall measures the accuracy of classifying frames with target visible. The two measures are combined into F-measure, which is the primary measure. In the following we briefly present how the measures are calculated. For details and derivation we refer the reader to [33].

We denote G_t as a ground-truth target pose and $A_t(\tau_\theta)$ as a pose prediction given by a tracker at frame t. The evaluation protocol requires that the tracker reports a confidence value besides the pose prediction. The confidence of the tracker in frame t is denoted as θ_t while confidence threshold is denoted as τ_θ. If the target is not visible in frame t, then ground-truth is an empty set i.e., $G_t = \emptyset$. Similarly, if tracker does not report the prediction or if confidence score is below the confidence threshold, i.e., $\theta_t < \tau_\theta$, then the output is an empty set $A_t(\tau_\theta) = \emptyset$.

From the object detection literature, when intersection-over-union between the tracker prediction and ground-truth $\Omega(A_t(\tau_\theta), G_t)$, exceeds overlap threshold τ_Ω, the prediction is considered as correct. This definition of correct prediction highly depends on the minimal overlap threshold τ_Ω. The problem is in [33] addressed by integrating tracking precision and recall over all possible overlap thresholds

[3] We observed that target size and aspect ratio change are reliably detected differentiating values at 10 frames before and after the current timestep - thus the discrete temporal derivative considers 21 frames.

which results in the following measures

$$Pr(\tau_\theta) = \frac{1}{N_p} \sum_{t \in \{t: A_t(\tau_\theta) \neq \emptyset\}} \Omega(A_t(\tau_\theta), G_t), \quad (2)$$

$$Re(\tau_\theta) = \frac{1}{N_g} \sum_{t \in \{t: G_t \neq \emptyset\}} \Omega(A_t(\tau_\theta), G_t), \quad (3)$$

where N_g is number of frames where target is visible, i.e., $G_t \neq 0$ and N_p is number of frames where tracker made a prediction, i.e., $A_t(\tau_\theta) \neq \emptyset$. Tracking precision and recall are combined into a single score by computing tracking F-measure $F(\tau_\theta) = \Big(2Re(\tau_\theta)Pr(\tau_\theta)\Big) / \Big(Re(\tau_\theta) + Pr(\tau_\theta)\Big)$. Tracking performance is visualized on precision-recall and F-measure plots by computing scores for all confidence thresholds τ_θ. The highest F-measure on the F-measure plot represents the optimal confidence threshold and it is used for ranking trackers. This process also does not require manual threshold setting for each tracker separately.

The performance measures are directly extended to per-attribute analysis. In particular, the tracking Precision, Recall and F-measure are computed from predictions on the frames corresponding to a particular attribute.

5. Experiments

This section presents experimental results on the CDTB dataset. Section 5.1 summarizes the list of tested trackers, Section 5.2 compares the CDTB dataset with most related datasets, Section 5.3 reports overall tracking performance and Section 5.4 reports per-attribute performance.

5.1. Tested Trackers

The following 16 trackers were chosen for evaluation. We tested (i) RGB baseline and state-of-the-art short-term correlation and deep trackers (KCF [16], NCC [29], BACF [22], CSRDCF [32], SiamFC [2], ECOhc [10], ECO [10] and MDNet [38]), (ii) RGB state-of-the-art long-term trackers (TLD [18], FuCoLoT [31] and MBMD [46]) and (iii) RGB-D state-of-the-art trackers (OTR [21] and Ca3dMS [30]). Additionally, the following RGB trackers have been modified to use depth information: ECOhc-D [19], CSRDCF-D [19] and KCF-D[4].

5.2. Comparison with Existing Benchmarks

Table 1 compares the properties of CDTB with the two currently available datasets, PTB [41] and STC [45]. CDTB is the only dataset that contains sequences captured with several devices in indoor and outdoor tracking scenes. STC [45] does in fact contain a few outdoor sequences, but these are confined to scenes without direct sunlight due to infra-red-based depth acquisition. The number of attributes is comparable to STC and much higher than PTB. The number of sequences (N_{seq}) is comparable to the currently largest dataset PTB, but CDTB exceeds the related datasets by an order of magnitude in the number of frames (N_{frm}). In fact, the average sequence of CDTB is approximately six times longer than in related datasets (N_{avg}), which affords a more accurate evaluation of long-term tracking properties.

A crucial tracker property required in many practical applications is target absence detection and target re-detection. STC lacks these events. The number of target disappearances followed by re-appearance in CDTB is comparable to PTB, but the disappearance periods (N_{out}) are much longer in CDTB. The average period of target absent (N_{avgout}) in PTB is approximately 6 frames, which means that only short-term occlusions are present. The average period of target absent in CDTB is nearly ten times larger, which allows tracker evaluation under much more challenging and realistic conditions.

Pose changes are much more frequent in CDTB than in the other two datasets. For example, the target undergoes a 180 degree out-of-plane rotation less than once per sequence in PTB and STC (N_{seqrot}). Since CDTB captures more dynamic scenarios, the target undergoes such pose change nearly 5 times per sequence.

The level of appearance change, realism, disappearances and sequence lengths result in a much more challenging dataset that allows performance evaluation more similar to the real-world tracking environment than STC and PTB. To quantify this, we evaluated trackers Ca3dMS, CSR-D and OTR on the three datasets and averaged their results. The trackers were evaluated on STC and CDTB using the PTB performance measure, since PTB does not provide ground truth bounding boxes for public evaluation.

Table 1 shows that the trackers achieve the highest performance on PTB, making it least challenging. The performance drops on STC, which supports the challenging small dataset diversity paradigm promoted in [45]. The performance further significantly drops on CDTB, which confirms that this dataset is the most challenging among the three.

5.3. Overall Tracking Performance

Figure 2 shows trackers ranked according to the F-measure, while tracking Precision-Recall plots are visualized for additional insights. A striking result is that the overall top-performing trackers are pure RGB trackers, which do not use depth information at all. MDNet and MBMD achieve comparable F-score, while FuCoLoT ranks third. It is worth mentioning that all three trackers are long-term with strong re-detection capability [33]. Even though MDNet was originally published as a short-term tracker, it has been shown that it performs well in a long-term scenario [33, 35, 43] due to its powerful CNN-based classi-

[4]KCF-D is modified by using depth as a feature channel in a correlation filter.

Table 1. Comparison of CDTB with related benchmarks in the number of RGB-D devices used for acquisition (N_{HW}), presence of indoor and outdoor sequences (In/Out), per-frame attribute annotation (Per-frame), number of attributes (N_{atr}), number of sequences (N_{seq}), total number of frames (N_{frm}) average sequence length (N_{avg}), number of frames with target not visible (N_{out}), number of target disappearances (N_{dis}), average length of target absence period (N_{avgout}), number of times a target rotates away from the camera by at least 180°(N_{rot}), average number of target rotations per sequence (N_{seqrot}) and tracking performance under the PTB protocol ($\Omega_{0.5}$).

Dataset	N_{HW}	In	Out	Per-frame	N_{atr}	N_{seq}	N_{frm}	N_{avg}	N_{out}	N_{avgout}	N_{dis}	N_{rot}	N_{seqrot}	$\Omega_{0.5}$
CDTB	3	✓	✓	✓	13	80	101,956	1,274	10,656	56.4	189	358	4.5	0.316
STC [45]	1	✓	✓	✓	12	36	9,195	255	0	0	0	30	0.8	0.530
PTB [41]	1	✓	✗	✗	5	95	20,332	214	846	6.3	134	83	0.9	0.749

fier with selective update and hard negative mining. Another long-term tracker, TLD, is ranked very low despite its re-detection capability, due to a fairly simplistic visual model which is unable to capture complex target appearance changes.

State-of-the-art RGB-D trackers, OTR and CSRDCF-D, using only hand-crafted features, achieve a comparable performance to complex deep-features-based short-term RGB trackers ECO and SiamFC. This implies that modern RGB deep features may compensate for the lack of depth information to some extent. On the other hand, state-of-the-art RGB trackers show improvements when extended by depth channel (CSRDCF-D, ECOhc-D and KCF-D). This means that existing RGB-D trackers lag behind the state-of-the-art RGB trackers which is a large opportunity for improvement by utilizing deep features combined with depth information.

Overall, both state-of-the-art RGB and RGB-D trackers exhibit a relatively low performance. For example, tracking Recall can be interpreted as the average overlap with ground truth on frames in which the target is visible. This value is below 0.5 for all trackers, which implies the dataset is particularly challenging for all trackers and offers significant potential for tracker improvement. Furthermore, we calculated tracking F-measure on sequences captured with each depth sensor. The results are comparable – 0.30 (ToF), 0.33 (Kinect) and 0.39 (stereo) – but they also imply that ToF is the most challenging and stereo is the least challenging sensor.

Precision-recall analysis. For further performance insights, we visualize the tracking Precision and Recall at the optimal tracking point, i.e., at the highest F-measure, in Figure 3. Precision and Recall are similarly low for most trackers, implying that trackers need to improve in target detection as well as localization accuracy. FuCoLoT, CSRDCF-D and TLD obtain significantly higher Precision than Recall, which means that mechanism for reporting loss of target is rather conservative in these trackers – a typical property we observed in all long-term trackers. The NCC tracker achieves significantly higher precision than recall, but this is a degenerated case since the target is reported as lost for most part of the sequence (very low Recall).

Another interesting observation is that tracking precision of the FuCoLoT is comparable to the top-performing MDNet and MBMD which shows that predictions made by FuColoT are similarly accurate to those made by top-performing trackers. On the other hand, top-performing MDNet and MBMD have a much higher recall, which shows that they are able to correctly track much more

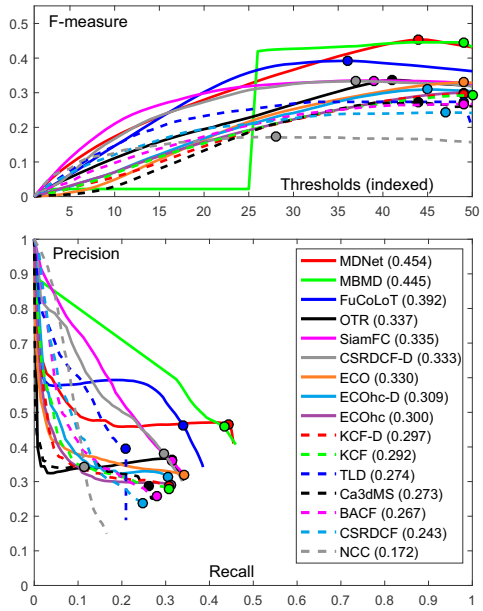

Figure 2. The overall tracking performance is presented as tracking F-measure (top) and tracking Precision-Recall (bottom). Trackers are ranked by their optimal tracking performance (maximum F-measure).

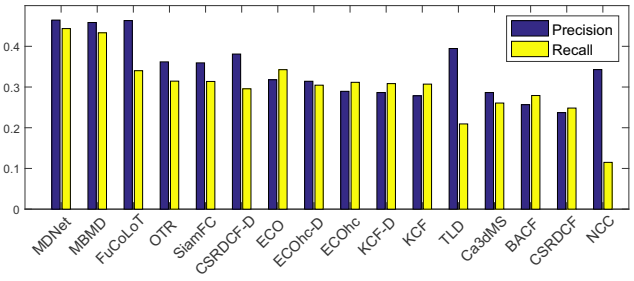

Figure 3. Tracking precision and recall calculated at the optimal point (maximum F-measure).

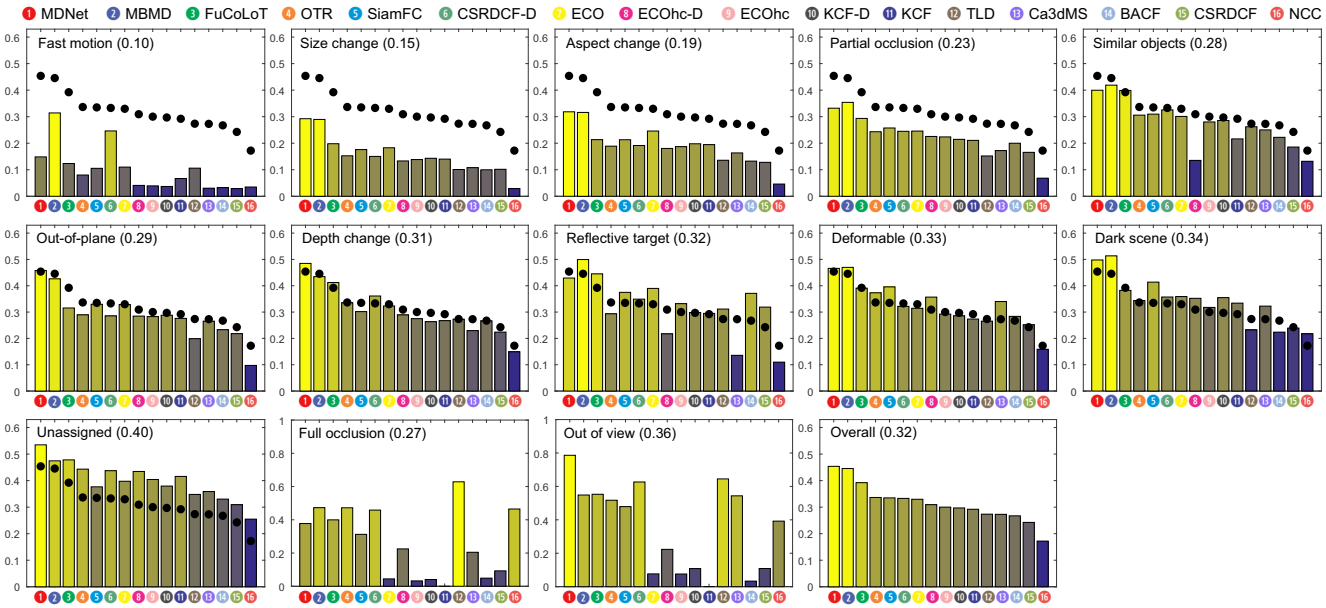

Figure 4. Tracking performance w.r.t. visual attributes. The first eleven attributes correspond to scenarios with a visible target (showing F-measure). The overall tracking performance is shown in each graph with black dots. The attributes *full occlusion* and *out of view* represent periods when the target is not visible and true negative rate is used to measure the performance.

frames where the target is visible, which might again be attributed to the use of deep features.

Overall findings. We can identify several good practices in the tracking architectures that look promising according to the overall results. Methods based on *deep features* show promise in capturing complex target appearance changes. We believe that *training* deep features on depth offers an opportunity for performance boost. A reliable *failure detection mechanism* is an important property for RGB-D tracking. Depth offers a convenient cue for detection of such events and combined with image-wide *re-detection* some of the RGB-D trackers address the long-term tracking scenario well. Finally, we believe that depth offers a rich *information complementary* to RGB for 3D target appearance modeling and depth-based target separation from the background, which can contribute in target localization. None of the existing RGB-D trackers incorporates all of these architectural elements, which opens a lot of new research opportunities.

5.4. Per-attribute Tracking Performance

The trackers were also evaluated on thirteen visual attributes (Section 3.2) in Figure 4. Performance on the attributes with visible target is quantified by the average F-measure, while true-negative rate (TNR [43]) is used to quantify the performance under full occlusion and out-of-view target disappearance.

Performance of all trackers is very low on *fast-motion*, making it the most challenging attribute. The reason for performance degradation is most likely the relatively small frame-to-frame target search range. Some of the long-term RGB-D and RGB trackers, e.g., MBMD and CSRDCF-D, stand out from the other trackers due to a well-designed image-wide re-detection mechanism, which compensates for a small frame-to-frame receptive field.

The next most challenging attributes are target *size change* and *aspect change*. MDNet and MBMD significantly outperform the other trackers since they explicitly estimate the target aspect. Size change is related to depth change, but the RGB-D tracker do not exploit this, which opens an opportunity for further research in depth-based robust scale adaptation.

Partial occlusion is particularly challenging for both RGB and RGB-D trackers. Failing to detect occlusion can lead to adaptation of the visual model to the occluding object and eventual tracking drift. In addition, too small frame-to-frame target search region leads to failure of target re-detection after the occlusion.

The attributes *similar objects*, *out-of-plane rotation*, *deformable*, *depth-change* and *dark scene* do not significantly degrade the performance compared to the overall performance. Nevertheless, the overall performance of trackers is rather low, which leaves plenty of room for improvements. We observe a particularly large drop in ECOhc-D on the similar-objects attribute which indicates that the tracker locks on to the incorrect/similar object at re-detection stage.

The *reflective target* attribute, unique for objects such as metal cups, mostly affects RGB-D trackers. The reason is that objects of this class are fairly well distinguished from the background in RGB, while their depth image is consistently unreliable. This means that more effort should be put

in information fusion part of the RGB-D trackers.

The attributes *deformable* and *dark-scene* are very well addressed by deep trackers (MDNet, MBMD, SiamFC and ECO), which makes them the most promising for coping with such situations. It seems that normalization, non-linearity and pooling in CNNs make deep features sufficiently invariant to image intensity changes and object deformations observed in practice.

Full occlusions are usually short-lasting events. On average, the trackers detect full a occlusion with some delay, thus a large percentage of occlusion frames are mistaken for the target visible. This implies poor ability to distinguish the appearance change due to occlusion from other appearance changes. The best target absence prediction at full occlusion is achieved by TLD, which is the most conservative in predicting target presence.

Situations when the target leaves the field of view (*out-of-view* attribute) are better predictable than full occlusions, due to longer target absence periods. Long-term trackers are performing very well in these situations and conservative visual model update seems to be beneficial.

A no-redetection experiment from [33] was performed to measure target re-detection capability in the considered trackers (Figure 5). In this experiment the standard tracking Recall (Re) is compared to a recall (Re_0) computed on modified tracker output – all overlaps are set to zero after the first occurrence of the zero overlap (i.e., the first target loss). Large difference between the recalls ($Re - Re_0$) indicates a good re-detection capability of a tracker. The trackers with the largest re-detection capability are MBMD, FuCoLoT (RGB trackers) and CSRDCF-D (RGB-D extension of CSRDCF) followed by OTR (RGB-D tracker) and two RGB trackers MDNet and SiamFc.

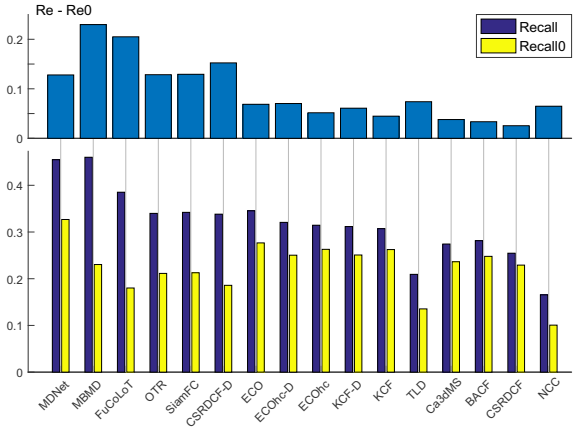

Figure 5. No redetection experiment. Tracking recall is shown on the bottom graph as dark blue bars. Modified tracking recall ($Re0$) is shown as yellow bars and it is calculated by setting the per-frame overlaps to zero after the first tracking failure. The difference between both recalls is shown on top. A large difference indicates good re-detection capability of the tracker.

6. Conclusion

We proposed a color-and-depth general visual object tracking benchmark (CDTB) that goes beyond the existing benchmarks in several ways. CDTB is the only benchmark with RGB-D dataset recorded by several color-and-depth sensors, which allows inclusion of indoor and outdoor sequences captured under unconstrained conditions (e.g., direct sun light) and covers a wide range of challenging depth signals. Empirical comparison to related datasets shows that CDTB contains a much higher level of object pose change and exceeds the other datasets in the number of frames by an order of magnitude. The objects disappear and reappear far more often, with disappearance periods ten times longer than in other benchmarks. Performance of trackers is lower on CDTB than related datasets. CDTB is thus currently the most challenging dataset, which allows RGB-D general object tracking evaluation under various realistic conditions involving target disappearance and re-appearance.

We evaluated recent state-of-the-art (SotA) RGB-D and RGB trackers on CDTB. Results show that SotA RGB trackers outperform SotA RGB-D trackers, which means that the architectures of RGB-D trackers could benefit from adopting (and adapting) elements of the recent RGB SotA. Nevertheless, the performance of all RGB and RGB-D trackers is rather low, leaving a significant room for improvements.

Detailed performance analysis showed several insights. Performance of baseline RGB trackers improved already from straightforward addition of the depth information. Current mechanisms for color and depth fusion in RGB-D trackers are inefficient and perhaps deep features trained on RGB-D data should be considered. RGB-D trackers do not fully exploit the depth information for robust object scale estimation. Fast motion is particularly challenging for all trackers indicating that short-term target search ranges should be increased. Target detection and mechanisms for detecting target loss have to be improved as well.

We believe these insights in combination with the presented benchmark will spark further advancements in RGB-D tracking and contribute to closing the gap between RGB and RGB-D state-of-the-art. Since the CDTB is a testing-only dataset we will work on constructing a large 6DOF dataset which could be used for training deep models for RGB-D tracking in the future.

Acknowledgements. This work is supported by Business Finland under Grant 1848/31/2015 and Slovenian research agency program P2-0214 and projects J2-8175 and J2-9433. J. Matas is supported by the Technology Agency of the Czech Republic project TE01020415 – V3C Visual Computing Competence Center.

References

[1] Ning An, Xiao-Guang Zhao, and Zeng-Guang Hou. Online RGB-D Tracking via Detection-Learning-Segmentation. In *ICPR*, 2016. 3

[2] Luca Bertinetto, Jack Valmadre, João F Henriques, Andrea Vedaldi, and Philip H S Torr. Fully-Convolutional Siamese Networks for Object Tracking. In *ECCV Workshops*, 2016. 5

[3] Adel Bibi, Tianzhu Zhang, and Bernard Ghanem. 3D Part-Based Sparse Tracker with Automatic Synchronization and Registration. In *CVPR*, 2016. 2, 3

[4] David S. Bolme, J.Ross Beveridge, Bruce A. Draper, and Yui-Man Lui. Visual Object Tracking using Adaptive Correlation Filters. In *CVPR*, 2010. 3

[5] Anders Glent Buch, Dirk Kraft, Joni-Kristian Kamarainen, Henrik Gordon Petersen, and Norbert Krüger. Pose estimation using local structure-specific shape and appearance context. In *ICRA*, 2013. 2

[6] Massimo Camplani, Sion Hannuna, Majid Mirmehdi, Dima Damen, Adeline Paiement, Lili Tao, and Tilo Burghardt. Real-time RGB-D Tracking with Depth Scaling Kernelised Correlation Filters and Occlusion Handling. In *BMVC*, 2015. 3

[7] Changhyun Choi and Henrik Iskov Christensen. RGB-D object tracking: A particle filter approach on GPU. In *IROS*, 2013. 2

[8] Wongun Choi, Caroline Pantofaru, and Silvio Savarese. A General Framework for Tracking Multiple People from a Moving Camera. *IEEE PAMI*, 2013. 2

[9] Navneet Dalal and Bill Triggs. Histograms of Oriented Gradients for Human Detection. In *CVPR*, 2005. 3

[10] Martin Danelljan, Goutam Bhat, Fahad Shahbaz Khan, and Michael Felsberg. ECO: Efficient Convolution Operators for Tracking. In *CVPR*, 2017. 3, 5

[11] Andreas Ess, Bastian Leibe, Konrad Schindler, and Luc van Gool. A Mobile Vision System for Robust Multi-Person Tracking. In *CVPR*, 2008. 2

[12] Hamed Kiani Galoogahi, Terence Sim, and Simon Lucey. Correlation Filters with Limited Boundaries. In *CVPR*, 2015. 3

[13] Guillermo Garcia-Hernando, Shanxin Yuan, Seungryul Baek, and Tae-Kyun Kim. First-Person Hand Action Benchmark with RGB-D Videos and 3D Hand Pose Annotations. In *CVPR*, 2018. 2

[14] Sion Hannuna, Massimo Camplani, Jake Hall, Majid Mirmehdi, Dima Damen, Tilo Burghardt, Adeline Paiement, and Lili Tao. DS-KCF: A Real-time Tracker for RGB-D Data. *Journal of Real-Time Image Processing*, 2016. 3

[15] Richard Hartley and Andrew Zisserman. *Multiple View Geometry in Computer Vision*. Second edition, 2004. 4

[16] Joao F. Henriques, Rui Caseiro, Pedro Martins, and Jorge Batista. High-Speed Tracking with Kernelized Correlation Filters. *IEEE PAMI*, 37(3):583–596, 2015. 3, 5

[17] Heiko Hirschmuller. Accurate and Efficient Stereo Processing by Semi-Global Matching and Mutual Information. In *CVPR*, 2005. 3

[18] Zdenek Kalal, Krystian Mikolajczyk, and Jiri Matas. Tracking-Learning-Detection. *IEEE PAMI*, 34(7):1409–1422, 2011. 3, 5

[19] Ugur Kart, Joni-Kristian Kämäräinen, and Jiri Matas. How to Make an RGBD Tracker ? In *ECCV Workshops*, 2018. 3, 5

[20] Ugur Kart, Joni-Kristian Kämäräinen, Jiri Matas, Lixin Fan, and Francesco Cricri. Depth Masked Discriminative Correlation Filter. In *ICPR*, 2018. 3

[21] Ugur Kart, Alan Lukežič, Matej Kristan, J.-K. Kämäräinen, and J. Matas. Object Tracking by Reconstruction with View-Specific Discriminative Correlation Filters. In *CVPR*, 2019. 3, 5

[22] Hamed Kiani Galoogahi, Ashton Fagg, and Simon Lucey. Learning Background-Aware Correlation Filters for Visual Tracking. In *ICCV*, 2017. 5

[23] Matej Kristan, Ales Leonardis, Jiri Matas, Michael Felsberg, Roman Pflugfelder, Luka Čehovin, Tomas Vojír, and et al. The Visual Object Tracking VOT2016 Challenge Results. In *ECCV Workshops*, 2016. 1

[24] Matej Kristan, Ales Leonardis, Jiri Matas, Michael Felsberg, Roman Pflugfelder, and et al. The Visual Object Tracking VOT2017 Challenge Results. In *ICCV Workshops*, 2017. 1

[25] Matej Kristan, Ales Leonardis, Jiri Matas, Michael Felsberg, Roman Pfugfelder, Luka Cehovin Zajc, and Tomas Vojir et al. The sixth Visual Object Tracking VOT2018 challenge results. In *ECCV Workshops*, 2018. 1, 3, 4

[26] Matej Kristan, Jiri Matas, Ales Leonardis, Michael Felsberg, and Luka et al. Čehovin Zajc. The Visual Object Tracking VOT2015 Challenge Results. In *ICCV Workshops*, 2015. 1, 3

[27] Matej Kristan, Jiri Matas, Georg Nebehay, Fatih Porikli, and Luka Čehovin. A Novel Performance Evaluation Methodology for Single-Target Trackers. *IEEE PAMI*, 38(11):2137–2155, 2016. 1

[28] Matej Kristan, Roman Pflugfelder, Ales Leonardis, Jiri Matas, Luka Čehovin, Georg Nebehay, Tomas Vojír, and et al. The Visual Object Tracking VOT2014 Challenge Results. In *ECCV Workshops*, 2014. 1, 4

[29] Matej Kristan, Roman Pflugfelder, Ales Leonardis, Jiri Matas, Fatih Porikli, and et al. The Visual Object Tracking VOT2013 Challenge Results. In *CVPR Workshops*, 2013. 5

[30] Ye Liu, Xiao-Yuan Jing, Jianhui Nie, Hao Gao, Jun Liu, and Guo-Ping Jiang. Context-aware 3-D Mean-shift with Occlusion Handling for Robust Object Tracking in RGB-D Videos. *IEEE TMM*, 2018. 3, 5

[31] Alan Lukežič, Luka Čehovin Zajc, Tom'aš Vojiř, Jiř'i Matas, and Matej Kristan. FuCoLoT - A Fully-Correlational Long-Term Tracker. In *ACCV*, 2018. 3, 5

[32] Alan Lukežič, Tomas Vojír, Luka Čehovin, Jiri Matas, and Matej Kristan. Discriminative Correlation Filter with Channel and Spatial Reliability. In *CVPR*, 2017. 3, 5

[33] Alan Lukezic, Luka Cehovin Zajc, Tomás Vojír, Jiri Matas, and Matej Kristan. Now you see me: evaluating performance in long-term visual tracking. *CoRR*, abs/1804.07056, 2018. 2, 4, 5, 8

[34] Kourosh Meshgi, Shin ichi Maeda, Shigeyuki Oba, Henrik Skibbe, Yu zhe Li, and Shin Ishii. An Occlusion-aware Particle Filter Tracker to Handle Complex and Persistent Occlusions. *CVIU*, 150:81 – 94, 2016. 3

[35] Abhinav Moudgil and Vineet Gandhi. Long-Term Visual Object Tracking Benchmark. In *ACCV*, 2018. 5

[36] Matthias Mueller, Neil Smith, and Bernard Ghanem. A Benchmark and Simulator for UAV Tracking. In *ECCV*, 2016. 1

[37] Matthias Muller, Adel Bibi, Silvio Giancola, Salman Alsubaihi, and Bernard Ghanem. TrackingNet: A Large-Scale Dataset and Benchmark for Object Tracking in the Wild. In *ECCV*, 2018. 1

[38] Hyeonseob Nam and Bohyung Han. Learning Multi-Domain Convolutional Neural Networks for Visual Tracking. In *CVPR*, 2016. 3, 5

[39] Stephan R. Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for Data: Ground Truth from Computer Games. In *ECCV*, 2016. 2

[40] Arnold W. M. Smeulders, Dung Manh Chu, Rita Cucchiara, Simone Calderara, Afshin Dehghan, and Mubarak Shah. Visual Tracking: An Experimental Survey. *IEEE PAMI*, 36(7):1442–1468, 2014. 1

[41] Shuran Song and Jianxiong Xiao. Tracking Revisited Using RGBD Camera: Unified Benchmark and Baselines. In *ICCV*, 2013. 2, 3, 5, 6

[42] Luciano Spinello and Kai Oliver Arras. People detection in RGB-D data. In *IROS*, 2011. 2

[43] Jack Valmadre, Luca Bertinetto, João F. Henriques, Ran Tao, Andrea Vedaldi, Arnold W. M. Smeulders, Philip H. S. Torr, and Efstratios Gavves. Long-term Tracking in the Wild: A Benchmark. In *ECCV*, 2018. 2, 5, 7

[44] Yi Wu, Jongwoo Lim, and Yang Ming-Hsuan. Object Tracking Benchmark. *IEEE PAMI*, 37:1834 – 1848, 2015. 1

[45] Jingjing Xiao, Rustam Stolkin, Yuqing Gao, and Ales Leonardis. Robust Fusion of Color and Depth Data for RGB-D Target Tracking Using Adaptive Range-Invariant Depth Models and Spatio-Temporal Consistency Constraints. *IEEE Transactions on Cybernetics*, 48:2485 – 2499, 2018. 2, 3, 5, 6

[46] Yunhua Zhang, Dong Wang, Lijun Wang, Jinqing Qi, and Huchuan Lu. Learning Regression and Verification Networks for Long-term Visual Tracking. *CoRR*, abs/1809.04320, 2018. 3, 5