DAL: A Deep Depth-Aware Long-term Tracker

Yanlin Qian[§*], Song Yan[§*], Alan Lukežič[†], Matej Kristan[†], Joni-Kristian Kämäräinen[*] and Jiří Matas[‡]
[*]Computing Sciences, Tampere University, Finland
[†]Faculty of Computer and Information Science, University of Ljubljana, Slovenia
[‡]Faculty of Electrical Engineering, Czech Technical University in Prague, Czech Republic

***Abstract*—The best RGBD trackers provide high accuracy but are slow to run. On the other hand, the best RGB trackers are fast but clearly inferior on the RGBD datasets. In this work, we propose a deep depth-aware long-term tracker that achieves state-of-the-art RGBD tracking performance and is fast to run.**

We reformulate deep discriminative correlation filter (DCF) to embed the depth information into deep features. Moreover, the same depth-aware correlation filter is used for target re-detection. Comprehensive evaluations show that the proposed tracker achieves state-of-the-art performance on the Princeton RGBD, STC, and the newly-released CDTB benchmarks and runs 20 fps.

I. Introduction

Visual object tracking has progressed significantly largely thanks to the series of increasingly challenging visual object tracking benchmarks [1], [2], [3], [4], [5], [6]. In the most general formulation, a tracker is initialized in the first frame and is required to output the target position in all remaining frames. In many practical applications, such as surveillance systems, trackers need to cope with occlusions and the target leaving the camera view which are essential properties for *long-term trackers* [2].

A vast majority of the works have focused on RGB tracking, but recently RGBD (RGB+Depth) tracking has gained momentum. Depth is a particularly strong cue for object's 3D localization, potentially simplifies foreground-background separation for occlusion handling and even helps to construct a 3D model of the tracked object [7]. Moreover, a number of RGBD datasets have been introduced in increasing pace [8], [9], [10].

Recent works [11], [7] have demonstrated improved tracking performance by adopting depth based occlusion handling. However, a recent long-term RGBD tracking benchmark [10] revealed that the best performance is achieved with the state-of-the-art RGB trackers that omit the depth input. In the most recent RGBD track of the VOT challenge [3] the best RGBD trackers outperformed RGB trackers by a clear margin. These trackers, however, are complicated architectures using deep object detectors, segmentation and deep feature based tracker pipelines. Their complex structure makes them unacceptably slow (∼2fps) for many real-time applications and it is difficult to improve their computation without sacrificing accuracy. Speed-wise the best RGB trackers outperform RGBD trackers, but the speed-accuracy trade-off gap between the best RGB and RGBD trackers remains an open problem.

[§]These authors contributed equally

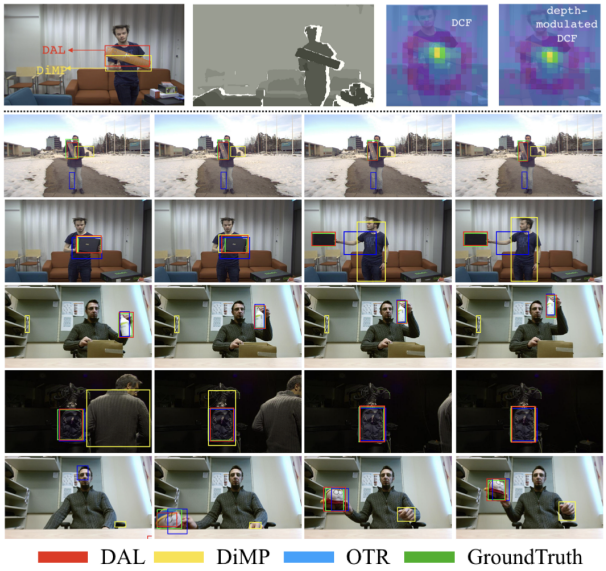

Fig. 1. Qualitative comparison of our tracker DAL and SoTA RGB and RGBD trackers. The top row illustrates the activation maps from a base DCF and a depth-modulated DCF (better zoom-in to see), generating slightly different shifts of target center and resulting in different bounding boxes (red and yellow). The two videos in the 1st and 2nd rows are non-occlusion scenarios, where our tracker, based on non-stationary DCF, localize the target well while the original DiMP fails, after multiple times of applying DCF operations. In the bottom three rows, the target appears from occlusion and are re-detected and tracked by our long-term tracker in a fast and accurate way.

This paper addresses the aforementioned issues and contributes by closing the performance gap in the terms of accuracy and speed between the RGB and RGBD trackers. We propose a new RGBD tracker of a streamlined architecture, namely DAL, which exploits depth information at all levels of processing and obtains performance comparable to the best slow RGBD trackers [3] with the speed comparable to real-time RGB trackers. The target appearance is modeled by adopting the state-of-the-art deep discriminative correlation filter (DCF) architecture [12]. However, the deep DCF is modulated using the depth information such that a large change in the depth suppresses discriminative features in these regions. The proposed "depth modulated" DCF model performs well both in short-time frame-to-frame tracking and target re-detection and therefore makes complex object detection unnecessary and provides significant speed-up. Tracking examples are shown in Figure 1.

The proposed long-term RGBD tracker achieves state-of-

the-art performance on all three available RGBD tracking benchmarks, PTB [8], STC [9] and CDTB benchmark [10], and runs an order of magnitude faster than the recent state-of-the-art RGBD tracker [7] or the winner of the recent VOT-RGBD challenge [3]. We also provide an ablation study that confirms the effectiveness of the depth modulated DCF formulation and other components of the proposed tracker.

II. Related Work

RGB trackers. Generic visual tracking with RGB input can be roughly divided into two tracks –discriminative correlation filter-based family (DCF) and Siamese-based family. Bolme *et al.* [13] inspired the visual tracking community of how visual tracking is addressed by DCF in a mathematical-sound way. DCF was extended by Henriques [14] with fourier-transform-based training, and later augmented with segmentation constraints in CSR-DCF [15]. Recently, DCF tends to be merged into an end-to-end deep network [16], [17]. The representative work is ATOM [18] that allows large-scale training for bounding box estimation and learning discriminative filter on the fly.

Siamese networks present the end-to-end trainable ability and relatively high tracking accuracy [19]. Li *et al.* [20] adopts a region proposal network for better predicted bounding boxes. Zhu *et al.* [21] suppresses the effect of background distractors by controlling the quality of learned target model. The most advanced siamese-based tracker is SiamRPN++[22], utilizing ResNet-50 for feature representation.

RGBD trackers. There are much less RGB-D trackers, compared to RGB ones. PTB [8] opened this research topic by presenting a hybrid RGBD tracker composed of HOG feature, optical flow and 3D point clouds. Under particle filter framework, Meshgi *et al.* [23] addresses RGBD tracker with occlusion awareness and Bibi *et al.* [24] further models a target using sparse 3D cuboids. Based on KCF, Hannuna *et al.* [25] uses depth for occlusion detection and An *et al.* [26] extends KCF with depth channel. Liu *et al.* [27] presents a 3D mean-shift-based tracker. Kart *et al.* [11] applies graph cut segmentation on color and depth information, generating better foreground mask for training CSR-DCF [15]. They then extend the idea with building an object-based 3D model [7], relying on a SLAM system Co-Fusion [28]. At the moment of writing this paper, OTR [7] leads the leaderboard of two RGBD benchmarks.

Benchmarks. Till now, we briefly introduced the most representative and well-performing RGBD trackers. The reason of their performance lag compared to RGB trackers is obvious – none of them has access to semantic target-based prior knowledge, which can be obtained via heavy off-line CNN training. Tracking benchmarks are crucial for the development of trackers.

It is obvious that RGBD benchmarks are much smaller than the RGB counterparts by orders of magnitude, for example, the biggest RGB tracking benchmark, TrackingNet [5] contains up to 14 million samples while the biggest RGBD tracking dataset CDTB [10] 100 thousand samples. Among RGBD datasets, only PTB [8] provide a tiny subset (hundreds of images) for training, which is far from enough for training or fine-tuning a deep net. The shortage of RGBD training set explains why off-line training has not been adopted for RGBD tracking. To narrow the performance gap between RGB and RGBD trackers, it is beneficial to use deep features from deep nets pretrained on massive RGB training set.

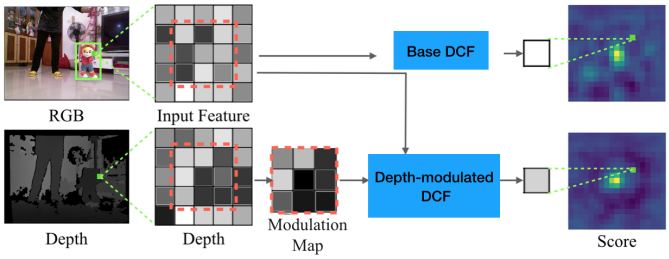

Fig. 2. Visualization of depth-modulated DCF. Depth modulates the DCF by re-weighting the DCF kernels according to the depth similarity with the tested target position. Top: the confidence score map of the target object resulting from base DCF; Bottom: the corresponding score map obtained by our depth-modulated DCF.

III. Method

In Sec. III-A, we first introduce the base RGB tracker briefly. In same section, we also describe the design of depth-aware convolution layer and show its application on DCF-based tracking. In Sec. III-B we briefly describe the bounding box regressor – IoUNet. We overview the long-term RGBD tracking architecture in Sec. III-C , with emphasis on the interaction conditions of switching between short-term tracking and re-detection mode in Sec. III-D.

A. Robust localization

Robust localization is the most crucial element of long-term tracking. We thus formulate the target model as a deep discriminative correlation filter (DCF), which is trained by the efficient deep training algorithm proposed recently [12]. Given a set of labelled training samples $\mathbf{S}_{train}$, the filter $\mathbf{f}$ is optimized by steepest descent on the following loss L_{cls}:

$$L_{\text{cls}} = \frac{1}{N_{\text{iter}}} \sum_{i=0}^{N_{\text{iter}}} \sum_{(x,c)\in S_{\text{train}}} \left\| \ell\big(\mathbf{x} * \mathbf{f}^{(i)}, \mathbf{z}_c\big) \right\|^2 . \qquad (1)$$

where $*$ is the convolution operation and $\mathbf{z}_c$ refers to the corresponding Gaussian function centered on the target location c of the training sample $\mathbf{x}$ and N_{iter} is the number of steepest descent iterations. The loss applies a nonlinear regression error $\ell(s,z) = s - z$ for $z > T$ and $\ell(s,z) = \max(0, s)$ for $z \leq T$, where T is a threshold on the error. The training samples $\mathbf{x} \in S_{train}$ are extracted from the image patch 5 times larger than the target size using a common backbone [29], which is fine-tuned for localization task [12].

The target is localized on a new frame by extracting deep features within a patch 5 times the target size and correlated

by the trained filter $\mathbf{f}$. Position of the maximum correlation response is the new target position estimate.

However, using a stationary filter (i.e., the same filter) on all locations is sub-optimal since certain regions might contain occlusion and are thus less reliable than other regions [30]. Furthermore, certain targets are poorly approximated by a rectangular convolution window and therefore a mechanism for background suppression is required. To solve all these problems simultaneously, we propose a non-stationary deep DCF that utilizes depth to modulate the DCF content with respect to the filter position. Specifically, we define the new depth-modulated DCF as

$$\tilde{\mathbf{f}}(x, y) = \mathbf{f} \odot \mathbf{\Theta}(x, y), \tag{2}$$

where $\mathbf{f}$ is a stationary base filter, $\mathbf{\Theta}(x, y)$ is a non-stationary 2D modulation map, and $\odot$ is a Hamadarad product, that multiplies all channels of the base filter with the same modulation map. The purpose of the modulation map is to give more weight to the pixels with depth values similar to the tested target position, thus reducing the effect of the background and occlusion. Let $\mathbf{D}(x, y)$ be the depth at the tested position and let $\mathbf{D}(x + m, y + n)$ be the depth of the neighboring pixel. The modulation map is then defined as

$$\mathbf{\Theta}_{mn}(x, y) = \exp(-\alpha|\mathbf{D}(x, y) - \mathbf{D}(x + m, y + n)|), \tag{3}$$

where α is a hyper parameter that controls the modulation strength. Figure 2 illustrates the modulation map construction and usage in non-stationary DCF correlation. The loss for training the non-stationary DCF becomes

$$L_{\text{cls}} = \frac{1}{N_{\text{iter}}} \sum_{i=0}^{N_{\text{iter}}} \sum_{(x,c)\in S_{\text{train}}} \|\ell\big(\mathbf{x} * \tilde{\mathbf{f}}^{(i)}(x, y), z_c\big)\|^2 \quad . \tag{4}$$

The loss is optimized using the steepest decent algorithm from [12] within a region five times the target size to harvest a sufficient amount of negative examples. The non-stationary DCF learns to take into account the target-background discontinuities induced by depth and therefore provides improved foreground-background discrimination.

B. Accurate localization

The non-stationary depth-modulated DCF described in Section III-A robustly localizes the target even in presence of clutter. For accurate bounding box prediction i.e., width and height of the target, we follow the recent IoUNet [31] bounding box regression introduced in [18].

The IoUNet is trained offline on image pairs of the same target using a large number of video sequences. First image and the corresponding bounding box are used as a training example. A modulation vector is extracted from this image and used with the second image (test example) to refine the given test bounding box and to predict its intersection over union with the ground-truth bounding box.

During tracking, after the target is approximately localized by the depth-modulated DCF (Section III-A), N^{BB} positions are sampled around the predicted position and IoUNet is applied to produce refined bounding boxes with predicted IoU scores. N^{TOP} bounding boxes with the highest predicted score are averaged to produce the final bounding box.

TABLE I
SUMMARY OF THE TRACKING STATE TRIGGERS IN SECTION III-D.

State	**Conditions**
Target lost	$cond_1 : 1 - \beta_{\text{DCF}}(\tau_l)$ $cond_2 : \ 1 - \beta_{\text{DCF}}(\tau)$ & $1 - \beta_{\text{dep}}(\tau_D)$
Target re-detected	$cond_1 : \beta_{\text{DCF}}(\tau_h)$ $cond_2 : \beta_{\text{DCF}}(\tau)$ & $\beta_{\text{dep}}(\tau_D)$
Update model	$cond_1$: $\beta_{\text{DCF}}(\tau_u)$ & $\beta_{\text{dep}}(\tau_D)$

C. Long-term tracker architecture

A long-term tracker is required to address situation in which the target disappears for a long duration and re-appears later on. Target loss prediction and re-detection play a crucial role in these scenarios. We build on a single-model long-term tracking architecture [32]. In our case, the short-term tracker is composed of a robust localizer i.e., a deep non-stationary DCF (Section III-A) and an accurate bounding box refinement module (Section III-B), and is used for continuous, short-term, target localization. Periods of unreliable target localization are detected by a depth-aware target presence classifier (Section III-D). Once the target is deemed lost, the target search range progressively increases over the consecutive frames. Target is re-detected by applying the depth-modulated DCF from III-A within the enlarged search region. Once the target is re-detected, the search range reduces back to that of short-term tracking.

Since the same model is used for short-term tracking and detection, care has to be taken to prevent model contamination and irrecoverable drift caused by updating from the background. We thus apply target presence indicators (Section III-D) to switch between target presence/absence states and identify periods during which it is safe to update the target model.

D. Depth-aware target presence indicators

The similarity between the model and the detected target is quantified by the maximum of the depth-modulated DCF correlation response, i.e., ρ_{DCF}. Low value indicates a low target presence likelihood. Thus the correlation-based target presence indicator is defined as $\beta_{\text{DCF}}(\tau) = \{1 : \rho_{\text{DCF}} > \tau; 0 : \text{ otherwise}\}$.

Temporal depth consistency is used as another indicator. The target is represented by the set of depth histograms $\mathcal{G}_i \in G, i = 1, ..., N_G$, extracted from the depth images from predicted bounding box region in the previous time-steps. A histogram extracted in the current time-step $\mathcal{H}$ is compared to these histograms by Bhattacharyya similarity

$$\rho^i_{\text{dep}} = \sum_j^{n_B} \sqrt{\mathcal{H}_j \mathcal{G}^i_j}, \tag{5}$$

where n_B is number of the histogram bins. Low values indicate target occlusion or disappearance. The

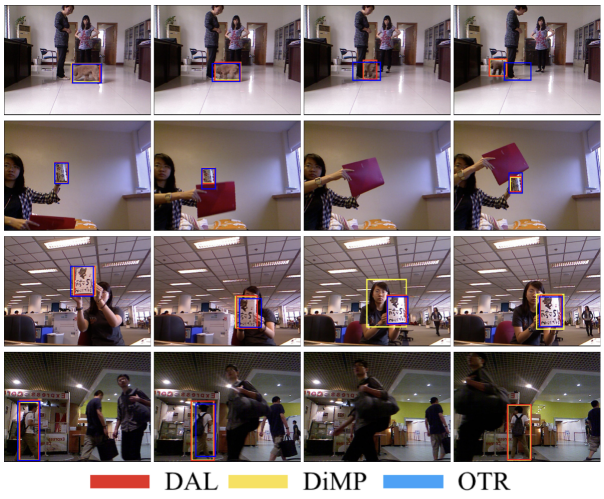

Fig. 3. Qualitative comparison of DAL, DiMP and OTR on the PTB. All trackers localize the target and give precise bounding boxes (the first two columns). With depth-modulated DCF, our tracker shows better discriminative ability when strong distractor appears (human face in the third row and human legs in the first row). Compared to DiMP and OTR, with conservative long-term tracking design, our tracker reports target disappearance more accurately.

depth consistency indicator is therefore defined as $\beta_{\text{dep}}(\tau) = \{1 : \rho^i_{\text{dep}} > \tau \ \forall \ i; 0 : \text{ otherwise}\}$. The set of depth histograms is refreshed each time a target model is updated by first-in-first-out mechanism.

The correlation and depth consistency indicators are applied to construct conditions to trigger (i) target lost state, (ii) target re-detected state, and (iii) to decide whether it is safe to update the target model without background contamination. The triggers are summarized in Table I.

IV. Experiments

A. Implementation Details

The backbones for deep DCF and the IoUNet are pre-trained for localization task on RGB sequences and the filter update parameters are kept as in [12]. The depth modulation hyper-parameter is set to $\alpha = 0.1$. The binds in depth histograms are constrained to 8 meters at resolution of 0.1m per bin. The search region enlargement rate factor during re-detection is set to $r = 1.05$.The target presence indicator thresholds (Table I) are set to $\beta_l = 0.2$, $\beta = 0.25$, $\beta_h = 0.3$ and $\beta_D = 0.8$. The number of depth histograms in the depth temporal consistency model is set to $N_G = 3$. The preliminary study showed that the method is not sensitive to these parameters and we use the same values in all experiments.

B. State-of-the-art Comparison

The proposed depth-aware long-term (DAL) tracker is evaluated on three major RGB-D benchmarks: Princeton tracking benchmark [8] (PTB), STC tracking benchmark [9] and Color-and-depth tracking benchmark [10] (CDTB). In the following we discuss the tracking performance on these benchmarks.

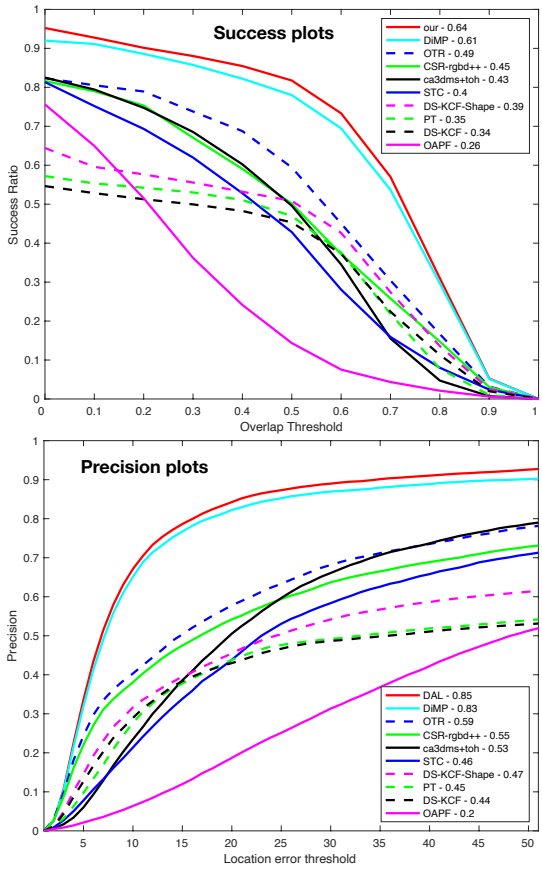

Fig. 4. Success and precision plots on STC benchmark [9].

1) Princeton Tracking Benchmark: Princeton Tracking Benchmark [8] (PTB) is the most popular benchmark in RGBD tracking. The dataset consists of 95 video sequences without publicly available ground-truth annotations to prevent over-fitting. The sequences are annotated with 11 visual attributes for a thorough analysis of tracking performance (Table II). A per-frame tracking performance is measured by a modified overlap measure that sets the overlap to 1 on frames where the target is correctly predicted to be missing. The overall tracking performance is measured by the *success rate*, i.e., the percentage of frames where overlap between ground truth and the predicted bounding box exceeds 0.5.

All state-of-the-art RGBD trackers from PTB are included in the analysis: OTR [7], ca3dms+toh [27], CSR-rgbd++ [11], 3D-T [24], PT [8], OAPF [23], DM-DCF [33], DS-KCF-Shape [25], DS-KCF [34], DS-KCF-CPP [25], hiob_lc2 [35], STC [9] and DLST [26]. We additionally include the state-of-the-art short-term RGB tracker DiMP [12], which ranks top on the most short-term tracking benchmarks.

DAL achieves the average success rate higher than 0.8, outperforming all RGBD trackers and outperform the sota RGBD tracker OTR and the sota RGB tracker DiMP by 5%. On most attributes except “Passive motion”, DAL ranks the first or the second, showing its robustness under various tracking conditions. Compared to OTR, the success rates on

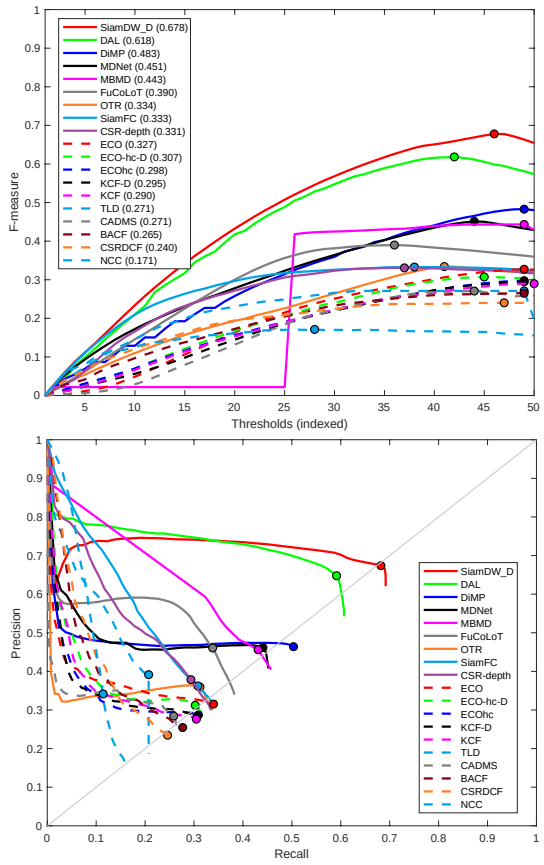

Fig. 5. The overall tracking performance is presented as tracking F-measure (top) and tracking Precision-Recall (bottom) on the CTDB dataset. Trackers are ranked by their optimal tracking performance (maximum F-measure).

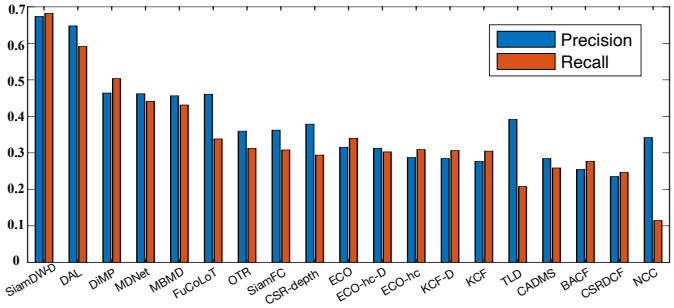

Fig. 6. Tracking precision and recall calculated at the optimal point (maximum F-measure). Evaluated on the CDTB dataset.

"Animal, Small Object, No-Occlusion, Active Motion" are significantly better, verifying the improved utilization in depth for tracking. The per-attribute results also show that DAL deals better with occlusion than DiMP, which may be attributed to the non-stationary DCF modulated by the depth map. See Figure 3 for qualitative comparison on PTB.

2) STC Tracking Benchmark: The STC benchmark [9] is complementary to the PTB in terms of visual attributes and contains 36 sequences annotated per-frame with 10 attributes: *Illumination variation* (IV), *Depth variation* (DV), *Scale variation* (SV), *Color distribution variation* (CDV), *Depth distribution variation* (DDV), *Surrounding depth clutter* (SDC), *Surrounding color clutter* (SCC), *Background color camouflages* (BCC), *Background shape camouflages* (BSC), *Partial occlusion* (PO).

Since the targets in STC dataset are always visible, the standard short-term tracking evaluation methodology is used [36]. Tracking performance is evaluated by the success and precision plots. The success plot shows percentage of frames where overlap of the predicted bounding box is larger than a threshold, for a set of overlap thresholds. Trackers are ranked according to the area under the success rate curve. The precision plot shows percentage of frames where distance between the predicted bounding box center and the ground-truth bounding box center is smaller than the threshold, for a set of center error thresholds. Trackers are ranked according to the performance at the threshold of 20 pixels.

The proposed tracker is compared to the following state-of-the-art RGBD trackers: OTR [7], CSR-rgbd++ [11], ca3dms+toh [27], STC [9], DS-KCF-Shape [25], PT [8], DS-KCF [34] and OAPF [23]. The most recent sota RGB short-term tracker (DiMP [12]) is included as well.

Results are reported in Figure 4. DAL outperforms top-performing RGBD trackers by a large margin. The top RGBD tracker OTR is outperformed significantly by 30.6%, while DiMP is outperformed by 4.9%. The improved performance is consistent across all the attributes, except SCC (Surrounding Color Clutter).

C. Color and depth tracking benchmark

The CDTB dataset [10] is the most recent and the most challenging RGBD dataset. The sequences are captured in the long-term tracking scenario, which means that the target is often fully occluded or that it disappears from the field of view. The most important aspects in long-term tracking are therefore ability to predict target absence and target re-detection. Tracking performance is measured as tracking recall (Re, average overlap on frames where the target is visible) and tracking precision (Pr, average overlap on frames where tracker makes a prediction). Trackers are ranked according to the tracking F-measure, which is combination of Pr and Re.

The proposed tracker is compared to all top trackers from the CDTB benchmark: (i) sota short-term RGB trackers (KCF [14], NCC [1], BACF [37], CSRDCF [15], SiamFC [19], ECOhc [16], ECO [16] and MDNet [38]), (ii) sota long-term RGB trackers (TLD [39], FuCoLoT [32] and MBMD [40]) and (iii) sota RGBD trackers (OTR [7], Ca3dMS [27], ECOhc-D [11] and CSRDCF-D [11]). We also include the most recent short-term RGB tracker DiMP [12] which is the top-performer on the most of the short-term datasets and the winner of the recent VOT2019 RGBD challenge [3] (SiamDW-D [3]).

Tracking results are presented in Figure 5. The proposed tracker outperforms the top-performer in CDTB [10], MDNet, by a large margin of 37% mostly due to the powerful re-detection module and the non-stationary DCF. The OTR, which is the sota RGBD tracker, is outperformed by 85%

TABLE II
RESULTS AND RANKS (PARENTHESIS) RETRIEVED FROM THE PTB ONLINE SERVER. THE TOP THREE RESULTS FOR THE EACH ATTRIBUTE ARE ANNOTATED RESPECTIVELY.

Method	*Avg.Success*	*Human*	*Animal*	*Rigid*	*Large*	*Small*	*Slow*	*Fast*	*Occ.*	*No-Occ.*	*Passive*	*Active*
DAL (**ours**)	0.807(1)	0.78(2)	0.86(1)	0.81(2)	0.76	0.84(1)	0.83(2)	0.80(1)	0.72(2)	0.93(1)	0.78	0.82(1)
OTR [7]	0.769(2)	0.77(3)	0.68	0.81(2)	0.76	0.77(3)	0.81	0.75(2)	0.71	0.85	0.85(1)	0.74
DiMP [12]	0.765(3)	0.67	0.86(1)	0.79	0.67	0.81(2)	0.82(3)	0.73	0.63	0.93(1)	0.74	0.76(2)
ca3dms+toh [27]	0.737	0.66	0.74	0.82(1)	0.73	0.74	0.80	0.71	0.63	0.88(3)	0.83(2)	0.70
CSR-rgbd++ [11]	0.740	0.77	0.65	0.76	0.75	0.73	0.80	0.72	0.70	0.79	0.79	0.72
3D-T [24]	0.750	0.81(1)	0.64	0.73	0.80(1)	0.71	0.75	0.75(2)	0.73(1)	0.78	0.79	0.73
PT [8]	0.733	0.74	0.63	0.78	0.78(3)	0.70	0.76	0.72	0.72(2)	0.75	0.82(3)	0.70
OAPF [23]	0.731	0.64	0.85(3)	0.77	0.73	0.73	0.85(1)	0.68	0.64	0.85	0.78	0.71
DLST [26]	0.740	0.77	0.69	0.73	0.80(1)	0.70	0.73	0.74	0.66	0.85	0.72	0.75(3)
DM-DCF [33]	0.726	0.76	0.58	0.77	0.72	0.73	0.75	0.72	0.69	0.78	0.82	0.69
DS-KCF-Shape [25]	0.719	0.71	0.71	0.74	0.74	0.70	0.76	0.70	0.65	0.81	0.77	0.70
DS-KCF [34]	0.693	0.67	0.61	0.76	0.69	0.70	0.75	0.67	0.63	0.78	0.79	0.66
DS-KCF-CPP [25]	0.681	0.65	0.64	0.74	0.66	0.69	0.76	0.65	0.60	0.79	0.80	0.64
hiob-lc2 [35]	0.662	0.53	0.72	0.78	0.61	0.70	0.72	0.64	0.53	0.85	0.77	0.62
STC [9]	0.698	0.65	0.67	0.74	0.68	0.69	0.72	0.68	0.61	0.80	0.78	0.66

TABLE III
THE NORMALIZED AREA UNDER THE CURVE (AUC) SCORES COMPUTED FROM ONE-PASS EVALUATION ON THE STC BENCHMARK [9]. THE TOP THREE RESULTS FOR THE EACH ATTRIBUTE ARE ANNOTATED.

Method \ **Attributes**	*AUC*	*IV*	*DV*	*SV*	*CDV*	*DDV*	*SDC*	*SCC*	*BCC*	*BSC*	*PO*
DAL (**ours**)	0.64(1)	0.51(1)	0.63(1)	0.50(1)	0.60(1)	0.62(1)	0.64(1)	0.63(2)	0.57(1)	0.58(1)	0.58(1)
DiMP [12]	0.61(2)	0.50(2)	0.62(2)	0.48(2)	0.57(2)	0.58(2)	0.61(2)	0.65(1)	0.52(2)	0.55(2)	0.58(1)
OTR [7]	0.49(3)	0.39(3)	0.48(3)	0.31(3)	0.19	0.45(3)	0.44(3)	0.46	0.42(3)	0.42(3)	0.50(3)
CSR-rgbd++ [11]	0.45	0.35	0.43	0.30	0.14	0.39	0.40	0.43	0.38	0.40	0.46
ca3dms+toh [27]	0.43	0.25	0.39	0.29	0.17	0.33	0.41	0.48(3)	0.35	0.39	0.44
STC [9]	0.40	0.28	0.36	0.24	0.24(3)	0.36	0.38	0.45	0.32	0.34	0.37
DS-KCF-Shape [25]	0.39	0.29	0.38	0.21	0.04	0.25	0.38	0.47	0.27	0.31	0.37
PT [8]	0.35	0.20	0.32	0.13	0.02	0.17	0.32	0.39	0.27	0.27	0.30
DS-KCF [34]	0.34	0.26	0.34	0.16	0.07	0.20	0.38	0.39	0.23	0.25	0.29
OAPF [23]	0.26	0.15	0.21	0.15	0.15	0.18	0.24	0.29	0.18	0.23	0.28

mostly due to the better target representation including deep features and the deep non-stationary DCF. The proposed tracker outperforms sota RBG short-term DiMP *w.r.t.* the all three measures: tracking F-measure by 28%, precision by 40% and recall by 18%, which demonstrates the impact of the re-detection component and the non-stationary DCF. The top-performing tracker in VOT2019 RGBD challenge, SiamDW-D slightly outperforms DAL. SiamDW-D is a complex combination of multiple short-term tracking methods and general object detectors from the off-the-shelf toolbox [41]. This complicated architecture does prevents significant incorporation of depth in the tracker. In fact, depth is used only for target loss identification. Due to computational complexity, SiamDW-D performs at very low frame rate (2 fps as we test) and has large memory footprint due to several network branches. On the other hand, DAL has a very streamlined trainable architecture and runs $10\times$ faster thanks to efficient use of depth, while attaining comparable tracking accuracy.

D. Ablation Study

An ablation study was conducted on the most challenging RGBD dataset [10] to demonstrate the contribution of each component of DAL.

The following variants of DAL are evaluated: (i) DAL, the proposed method uses the non-stationary DCF and target presence classifier using DCF confidence and depth, described in Section III-D, to activate the re-detection, is denoted as +α+LT$^{\beta,\beta_\text{D}}$. (ii) DAL$^{-LT(\beta_D)}$, β_D is not considered in computing target presence. (iii) DAL$^{-\alpha,-LT(\beta_D)}$, β_D is not considered in computing target presence and only base DCF is used. (iv) DAL^{-LT}, long-term design is turned off, depth-modulated DCF is used. (v) DAL$^{-\alpha,-LT}$, long-term design is turned off and only base DCF is used, which equals to the base short-term tracker.

TABLE IV
DAL ABLATION STUDY ON THE CDTB SHOWING TRACKING F-MEASURE.

DAL$^{-\alpha,-LT}$	DAL^{-LT}	DAL$^{-\alpha,-LT(\beta_D)}$	DAL$^{-LT(\beta_D)}$	DAL
0.48	0.51	0.55	0.58	0.62

The results are shown in the Table IV. The short-term version of DAL using a stationary DCF, DAL$^{-\alpha,-LT}$, achieves 0.48 F-measure. Adding non-stationary formulation of DCF in DAL^{-LT} improves the results for 6.3%. The result shows that correlation-based trackers can benefit from the non-stationary DCF formulation. The baseline tracker with a stationary DCF extended to the long-term scenario (DAL$^{-\alpha,-LT(\beta_D)}$) improves the results for 14.6%, which shows the importance of re-detection in long-term tracking scenario. Combining both, non-stationary DCF formulation and target re-detection (DAL$^{-LT(\beta_D)}$) improves the results for 20.8%. Finally, the performance boost of 29.1% is achieved by the non-stationary

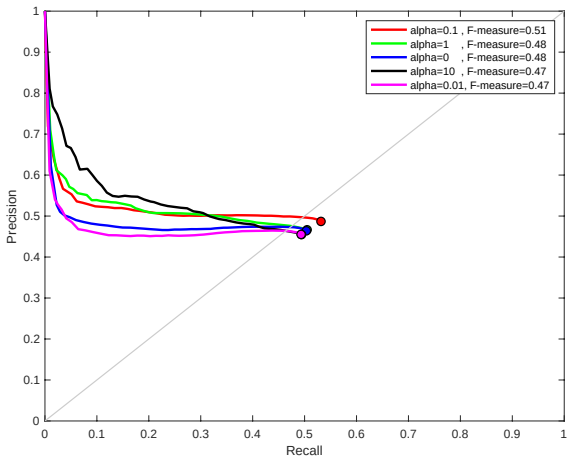

Fig. 7. Precision-Recall curves and F-measure as function of varying α. for depth-modulated DCF. Evaluated on CDTB dataset.

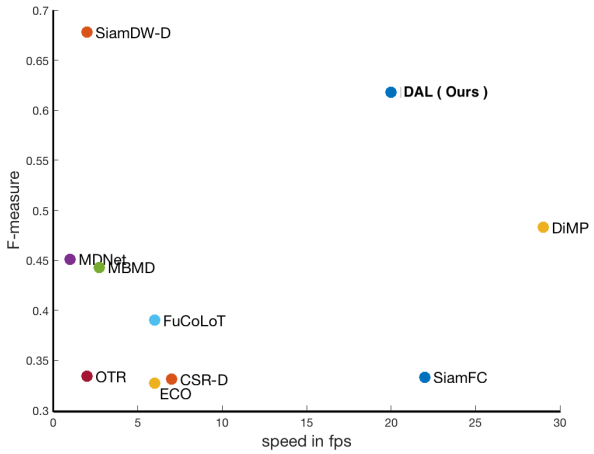

Fig. 8. Tracker practicality evaluation with respect to F-measure and Speed (in frames-per-second) on the CDTB dataset

DCF formulation and target re-detection activated by the multiple conditions using DCF confidence and depth (DAL). This result shows that depth can significantly improve tracking performance.

We additionally performed a sensitivity study of the parameter α from (3), which controls the modulation strength in the DCF depth modulation map. The baseline tracker (i.e., without depth modulation, $\alpha = 0$) was extended by the non-stationary DCF formulation (without target re-detection) and the following values of α were tested: 0.01, 0.1, 1 and 10. The results in Figure 7 show that the highest performance is obtained at $\alpha = 0.1$. Lower α push tracking performance to the baseline tracker. Increasing α amplifies the depth modulation too much, causing a slight performance drop.

Tracking speed. We measure the speed of ten top-performing trackers on the CDTB dataset to evaluate the performance in the context of practical applications that require accurate tracking at high speeds, i.e., robotics and real-time systems. Results are shown in Figure 8. DAL runs close to the real-time, at 20 frames per second, while most of the other trackers (MDNet, MBMD, OTR, ECO, CSR-D) are much slower and achieve significantly lower tracking accuracy. SiamFC runs similarly fast to DAL, but it achieves 46.1% lower tracking performance, DiMP is 45.0% faster, but it achieves 21.8% lower F-measure. The top-performing SiamDW-D achieves 9.7% higher F-measure, but it is 10-times slower. Thus DAL is the top-performing tracker in accuracy among the close-to-realtime tracking.

V. Conclusions

We propose a novel deep DCF formulation for RGBD tracking. The formulation embeds depth information into the correlation filter optimization and provides a strong short-term RGBD tracker, improving the performance from 5% to 6% on all RGBD tracking benchmarks. We also propose a long-term tracking architecture where the same deep DCF is used in target re-detection and depth based tests effectively trigger between the short-term tracking, re-detection and model update modes. The long-term tracker consistently achieves superior performance over the state-of-the-art RGB and RGBD trackers (DiMP and OTR) on all three available RGBD tracking benchmarks (PTB, STC and CDTB) and runs significantly faster than the best RGBD competitor (20 fps vs. 2 fps).

References

[1] M. Kristan, A. Leonardis, J. Matas, M. Felsberg, R. Pflugfelder, L. Cehovin Zajc, T. Vojir, G. Hager, A. Lukezic, A. Eldesokey *et al.*, "The visual object tracking vot2017 challenge results," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 1949–1972.

[2] M. Kristan, A. Leonardis, J. Matas, M. Felsberg, R. Pflugfelder, L. Cehovin Zajc, T. Vojir, G. Bhat, A. Lukezic, A. Eldesokey *et al.*, "The sixth visual object tracking vot2018 challenge results," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 0–0.

[3] M. Kristan, J. Matas, A. Leonardis, M. Felsberg, R. Pflugfelder, J.-K. Kamarainen, L. Cehovin Zajc, O. Drbohlav, A. Lukezic, A. Berg, A. Eldesokey, J. Kapyla, G. Fernandez, A. Gonzalez-Garcia, A. Memarmoghadam, A. Lu, A. He, A. Varfolomieiev, A. Chan, A. Shekhar Tripathi, A. Smeulders, B. Suraj Pedasingu, B. Xin Chen, B. Zhang, B. Wu, B. Li, B. He, B. Yan, B. Bai, B. Li, B. Li, B. Hak Kim, and B. Hak Ki, "The seventh visual object tracking vot2019 challenge results," in *The IEEE International Conference on Computer Vision (ICCV) Workshops*, Oct.

[4] L. Huang, X. Zhao, and K. Huang, "Got-10k: A large high-diversity benchmark for generic object tracking in the wild." in *arXiv preprint arXiv:1810.11981*, 2018.

[5] M. Muller, A. Bibi, S. Giancola, S. Alsubaihi, and B. Ghanem, "Trackingnet: A large-scale dataset and benchmark for object," in *ECCV*, 2018.

[6] H. Fan, L. Lin, F. Yang, P. Chu, G. Deng, S. Yu, H. Bai, Y. Xu, C. Liao, and H. Ling, "Lasot: A high-quality benchmark for large-scale single object tracking." in *CVPR*, 2019.

[7] U. Kart, A. Lukezic, M. Kristan, J.-K. Kamarainen, and J. Matas, "Object tracking by reconstruction with view-specific discriminative correlation filters," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1339–1348.

[8] S. Song and J. Xiao, "Tracking revisited using rgbd camera: Unified benchmark and baselines," in *ICCV*, 2013.

[9] J. Xiao, R. Stolkin, Y. Gao, and A. 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*, vol. 48, no. 8, pp. 2485–2499, 2017.

[10] A. Lukežič, U. Kart, J. Käpylä, A. Durmush, J.-K. Kämäräinen, J. Matas, and M. Kristan, "Cdtb: A color and depth visual object tracking dataset and benchmark," *ICCV*, 2019.

[11] U. Kart, J.-K. Kamarainen, and J. Matas, "How to make an rgbd tracker ?" in *ECCV Workshops*, 2018.

[12] G. Bhat, M. Danelljan, L. Van Gool, and R. Timofte, "Learning discriminative model prediction for tracking," *ICCV*, 2019.

[13] D. S. Bolme, J. R. Beveridge, B. A. Draper, and Y. M. Lui, "Visual object tracking using adaptive correlation filters." in *CVPR*, 2010.

[14] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista, "High-speed tracking with kernelized correlation filters," *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 3, pp. 583–596, 2014.

[15] A. Lukezic, T. Vojir, L. Cehovin Zajc, J. Matas, and M. Kristan, "Discriminative correlation filter with channel and spatial reliability," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6309–6318.

[16] M. Danelljan, G. Bhat, F. Shahbaz Khan, and M. Felsberg, "ECO: efficient convolution operators for tracking." in *CVPR*, 2017.

[17] J. Valmadre, L. Bertinetto, J. Henriques, A. Vedaldi, and P. H. Torr, "End-to-end representation learning for correlation filter based tracking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2805–2813.

[18] M. Danelljan, G. Bhat, F. S. Khan, and M. Felsberg, "Atom: Accurate tracking by overlap maximization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4660–4669.

[19] L. Bertinetto, J. Valmadre, J. F. Henriques, A. Vedaldi, and P. H. Torr, "Fully-convolutional siamese networks for object tracking," in *European conference on computer vision*. Springer, 2016, pp. 850–865.

[20] B. Li, J. Yan, W. Wu, Z. Zhu, and X. Hu, "High performance visual tracking with siamese region proposal network." in *CVPR*, 2018.

[21] Z. Zhu, Q. Wang, B. Li, W. Wu, J. Yan, and W. Hu, "Distractor-aware siamese networks for visual object tracking." in *ECCV*, 2018.

[22] B. Li, W. Wu, Q. Wang, F. Zhang, J. Xing, and J. Yan, "Siamrpn++: Evolution of siamese visual tracking with very deep networks." in *CVPR*, 2019.

[23] K. Meshgi, S.-i. Maeda, S. Oba, H. Skibbe, Y.-z. Li, and S. Ishii, "An occlusion-aware particle filter tracker to handle complex and persistent occlusions," in *CVIU*, 2016, pp. 150:81 – 94.

[24] A. Bibi, T. Zhang, and B. Ghanem, "3d part-based sparse tracker with automatic synchronization and registration," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1439–1448.

[25] S. Hannuna, M. Camplani, J. Hall, M. Mirmehdi, D. Damen, T. Burghardt, A. Paiement, and L. Tao, "Ds-kcf: a real-time tracker for rgb-d data," *Journal of Real-Time Image Processing*, pp. 1–20, 2016.

[26] N. An, X.-G. Zhao, and Z.-G. Hou, "Online rgb-d tracking via detection-learning-segmentation," in *ICPR*, 2016.

[27] Y. Liu, X.-Y. Jing, J. Nie, H. Gao, J. Liu, and G.-P. Jiang, "Context-aware three-dimensional mean-shift with occlusion handling for robust object tracking in rgb-d videos," *IEEE Transactions on Multimedia*, vol. 21, no. 3, pp. 664–677, 2018.

[28] M. Rünz and L. Agapito, "Co-fusion: Real-time segmentation, tracking and fusion of multiple objects," in *2017 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2017, pp. 4471–4478.

[29] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[30] W. Wang and U. Neumann, "Depth-aware cnn for rgb-d segmentation," in *ECCV*, 2018.

[31] B. Jiang, R. Luo, J. Mao, T. Xiao, and Y. Jiang, "Acquisition of localization confidence for accurate object detection," 2018, pp. 816–832.

[32] A. Lukežič, L. Č. Zajc, T. Vojíř, J. Matas, and M. Kristan, "Fucolot–a fully-correlational long-term tracker," in *Asian Conference on Computer Vision*. Springer, 2018, pp. 595–611.

[33] U. Kart, J.-K. Kämäräinen, J. Matas, L. Fan, and F. Cricri, "Depth masked discriminative correlation filter," in *ICPR*, 2018.

[34] M. Camplani, S. L. Hannuna, M. Mirmehdi, D. Damen, A. Paiement, L. Tao, and T. Burghardt, "Real-time rgb-d tracking with depth scaling kernelised correlation filters and occlusion handling," in *BMVC*, 2015.

[35] P. Springstübe, S. Heinrich, and S. Wermter, "Continuous convolutional object tracking," in *ESANN*, 2018.

[36] Y. Wu, J. Lim, and M.-H. Yang, "Object tracking benchmark," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 9, pp. 1834–1848, 2015.

[37] H. Kiani Galoogahi, A. Fagg, and S. Lucey, "Learning background-aware correlation filters for visual tracking," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 1135–1143.

[38] H. Nam and B. Han, "Learning multi-domain convolutional neural networks for visual tracking," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[39] Z. Kalal, K. Mikolajczyk, and J. Matas, "Tracking-learning-detection," *IEEE transactions on pattern analysis and machine intelligence*, vol. 34, no. 7, pp. 1409–1422, 2011.

[40] Y. Zhang, D. Wang, L. Wang, J. Qi, and H. Lu, "Learning regression and verification networks for long-term visual tracking," *arXiv preprint arXiv:1809.04320*, 2018.

[41] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2117–2125.