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*—In this work, we propose a long-term RGBD tracker that achieves state-of-the-art performance on the Princeton RGBD, STC and CDTB datasets and runs nearly real-time (20fps). The tracker is based on non-stationary formulation of the deep Discriminative Correlation Filter (DCF). The DCF feature channels are modulated by the depth context in each location which makes the tracker robust against occlusion and background clutter. Moreover, the non-stationary DCF also performs well in target re-detection enabling long-term tracking.**

I. Introduction

A vast majority of the works focus on RGB tracking, but recently RGBD (RGB + Depth) tracking has gained momentum. Depth is a particularly strong cue for object's 3D localization and simplifies foreground-background separation for occlusion handling. A number of RGBD tracking datasets have recently become available [1], [2], [3] for which recent works [4], [5] have demonstrated improved tracking performance by adopting depth based occlusion handling. However, a recent long-term RGBD tracking benchmark [3] revealed that the best performance is achieved with the state-of-the-art RGB trackers that omit the depth input. In the RGBD track of the VOT 2019 challenge [6] the best RGBD trackers outperformed RGB ones by a clear margin. However, complicated architectures of these RGBD trackers make them too slow (∼2fps) for many online applications.

This paper contributes by closing the performance gap in the terms of accuracy and speed between the RGB and RGBD trackers. We propose a non-stationary RGBD tracker that exploits depth information at all levels of processing and obtains superior performance to the best RGBD trackers [6] 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 [7]. The deep DCF is modulated using the depth information such that large changes in the depth suppress discriminative features in these regions. The proposed non-stationary deep DCF performs well both in short-time frame-to-frame tracking and target re-detection and therefore makes complex target re-detection procedures unnecessary (see examples in Figure 1).

The proposed long-term RGBD tracker achieves state-of-the-art performance on the three available RGBD tracking benchmarks, PTB [1], STC [2] and CDTB [3], and runs an order of magnitude faster than the recent state-of-the-art RGBD tracker [5] or the winner of the recent VOT-RGBD challenge [6]. We also provide an ablation study that confirms the effectiveness of the non-stationary DCF formulation and the proposed components of the DAL tracker.

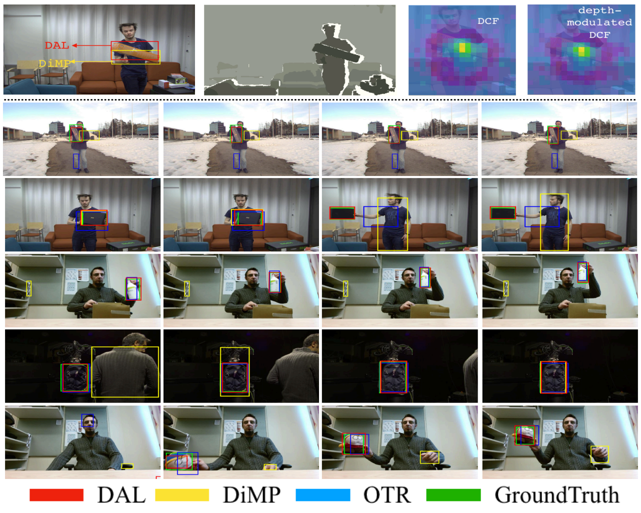

Fig. 1. Top: the activation maps from a base DCF and the proposed depth-modulated DCF, generating slightly different shifts of target center and resulting in different bounding boxes (red and yellow). The 1st and 2nd rows are non-occlusion scenarios, where our tracker DAL localizes the target well while the original DiMP fails. In the last three rows, the target appears from occlusion and are re-detected and tracked. [best viewed by zooming in]

II. Related Work

RGB trackers. There are two dominating tracks at the moment: DCF and the Siamese architecture based. Bolme *et al.* [8] introduced an effective and efficient Discriminative Correlation Filter (DCF) tracker. DCF was extended by Henriques [9] with fourier-transform-based training, and later augmented with segmentation constraints in CSR-DCF [10].

On the other hand, the end-to-end trained Siamese network based trackers have reported high tracking accuracy [11]. Li *et al.* [12] adopts a region proposal network for better predicted bounding boxes. Zhu *et al.* [13] suppresses the effect of background distractors by controlling the quality of learned target model. The most advanced siamese-based tracker is SiamRPN++[14], utilizing ResNet-50 for feature representation. Recently, DCF inspired CNN-based methods have been proposed. A representative work is the ATOM tracker [15] that allows large-scale training for bounding box estimation and learning discriminative filters on the fly.

RGBD trackers. A number of RGBD trackers have been proposed recently. PTB [1] was one of the first by presenting a hybrid RGBD tracker composed of HOG features, optical flow and 3D point clouds. Under the particle filter framework, Meshgi [16] introduce an RGBD tracker with occlusion awareness and Bibi [17] further models targets using sparse 3D cuboids. Based on KCF, Hannuna *et al.* [18] uses depth for occlusion detection and An *et al.* [19] extends KCF with the depth channel. Liu *et al.* [20] present a 3D mean-shift-based tracker. Kart *et al.* [4] apply graph cut segmentation on color and depth information, generating better foreground mask for training a DCF [10]. They extend the method by online 3D model estimation [5] that uses Co-Fusion [21] for SLAM. At the moment of writing this paper, OTR [5] leads the leaderboards of two RGBD benchmarks.

Benchmarks. RGBD benchmarks are smaller than the most recent RGB benchmarks. For example, TrackingNet [22] contains 14 million samples while the biggest RGBD dataset CDTB [3] only 100 sequences. PTB [1] provides a tiny subset (hundreds of images) for training, but this is insufficient to train or fine-tune large deep nets. STC [2] contains only 36 sequences for the short-term tracking setting. The shortage of RGBD training data for tracking explains why off-line training has not been adopted for RGBD tracking. RGBD trackers can however benefit from the large RGB tracking datasets.

III. Method

A. Non-stationary deep DCF

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 optimized by the recently proposed deep learning algorithm [7]. Given a set of labelled training samples $\mathbf{S}_{test}$, the filter $\mathbf{f}$ is optimized by steepest descent on the following loss L_{cls}:

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

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 $\mathbf{x} \in S_{test}$ are extracted from the image patch 5 times larger than the target size using a backbone [23] which is fine-tuned for the localization task [7].

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 $\mathbf{f}$. The 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 [24]. Furthermore, certain targets are poorly approximated by a rectangular convolution window and therefore a mechanism for background suppression is required. We propose a non-stationary DCF that utilizes depth to modulate the DCF content with respect to the filter position. Specifically, we define the non-stationary deep DCF as

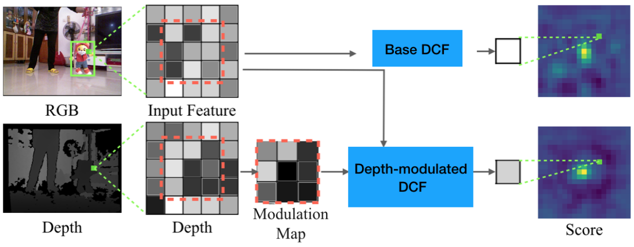

Fig. 2. Depth modulates the DCF by re-weighting the DCF kernels according to the depth similarity with the tested target position. Top: the results from the base DCF. Bottom : the results from the depth modulated DCF.

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

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 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

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

The loss is optimized using the steepest decent algorithm from [7] 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 deep DCF described in Sec. 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 [25] bounding box regression introduced in [15].

The IoUNet is trained offline on image pairs of the same target using a large number of video sequences. The 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

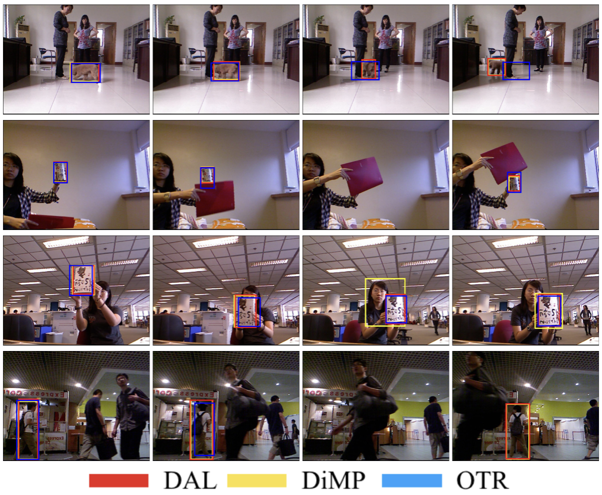

Fig. 3. Qualitative comparison of DAL, OTR [5] and DiMP [7] on the PTB. All trackers localizes the target and give precise bounding boxes (the first two rows). DAL shows better discriminative ability when strong distractor appears (human face in the third row). DAL reports target disappearance more accurately. Song would like to add the OTR results

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 (Sec. 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.

C. Long-term tracker architecture

A long-term tracker is required to address the situation in which the target disappears for long duration and re-appears later on. Target loss prediction and re-detection play a crucial role in these scenarios. In our case, the short-term tracker is composed of a robust localizer i.e., a non-stationary deep DCF (Sec. III-A) and an accurate bounding box refinement module (Sec. III-B), and is used for continuous, short-term, target localization. Periods of unreliable target localization are detected by a non-stationary target presence classifier (Sec. III-D). Once the target is lost, the target search range progressively increases over the consecutive frames. Target is re-detected by applying the non-stationary DCF from Sec. 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 (Sec. III-D) to switch between target presence/absence states and identify periods during which it is safe to update the target model.

D. 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 values indicate low target presence likelihood. Thus the correlation-based target presence indicator is defined simply as $\beta_{\text{DCF}}(\tau) = \{1 : \text{ if } \rho_{\text{DCF}} > \tau;\ 0 : \text{ otherwise}\}$.

The temporal depth consistency is used as another indicator. The target depth distribution is encoded by depth histograms $\mathcal{G}_i \in G, i = 1, ..., N_G$, extracted from the depth images inside the 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 the 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 depth consistency indicator is 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.

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

TABLE I
SUMMARY OF THE TARGET PRESENCE INDICATORS (SECTION III-D).

State	**Conditions**
Target lost	$\neg\beta_{\text{DCF}}(\tau)\ \wedge\ \neg\beta_{\text{dep}}(\tau_D)$
Target re-detected	$\beta_{\text{DCF}}(\tau_h)$
Update model	$\beta_{\text{DCF}}(\tau_u)\ \wedge\ \beta_{\text{dep}}(\tau_D)$

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 [7]. The depth modulation hyperparameter is set to $\alpha = 0.1$. The bins of 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 empirically set to $\tau_l = 0.2$, $\tau = 0.25$, $\tau_h = 0.3$ and $\tau_D = 0.8$. 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 the same values were used in all experiments.

B. State-of-the-art Comparison

The proposed depth-aware long-term (DAL) tracker was evaluated on the three available RGB-D benchmarks: Princeton tracking benchmark [1] (PTB), STC tracking benchmark [2] and Color-and-depth tracking benchmark [3]

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 [5]	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 [7]	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 [20]	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++ [4]	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 [17]	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 [1]	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 [16]	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 [19]	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 [26]	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 [18]	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 [27]	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 [18]	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 [28]	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 [2]	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 [2]. 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 [7]	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 [5]	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++ [4]	0.45	0.35	0.43	0.30	0.14	0.39	0.40	0.43	0.38	0.40	0.46
ca3dms+toh [20]	0.43	0.25	0.39	0.29	0.17	0.33	0.41	0.48(3)	0.35	0.39	0.44
STC [2]	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 [18]	0.39	0.29	0.38	0.21	0.04	0.25	0.38	0.47	0.27	0.31	0.37
PT [1]	0.35	0.20	0.32	0.13	0.02	0.17	0.32	0.39	0.27	0.27	0.30
DS-KCF [27]	0.34	0.26	0.34	0.16	0.07	0.20	0.38	0.39	0.23	0.25	0.29
OAPF [16]	0.26	0.15	0.21	0.15	0.15	0.18	0.24	0.29	0.18	0.23	0.28

(CDTB). The top performing trackers on all benchmarks were included to our experiments.

Princeton Tracking Benchmark [1] (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.

DAL achieves the average success rate higher than 0.8, outperforming all RGBD trackers and outperform the SoTA RGBD tracker OTR [5] and the SoTA RGB tracker DiMP [7] 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 "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.

The STC benchmark [2] is complementary to 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 [29]. Tracking performance is evaluated by the success and precision plots (Fig 4). 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.

Results are reported in Table III. DAL outperforms top-performing RGBD trackers by a large margin. The top RGBD tracker OTR is outperformed significally by 30.6%, while DiMP is outperformed by 4.9%. The improved performance is

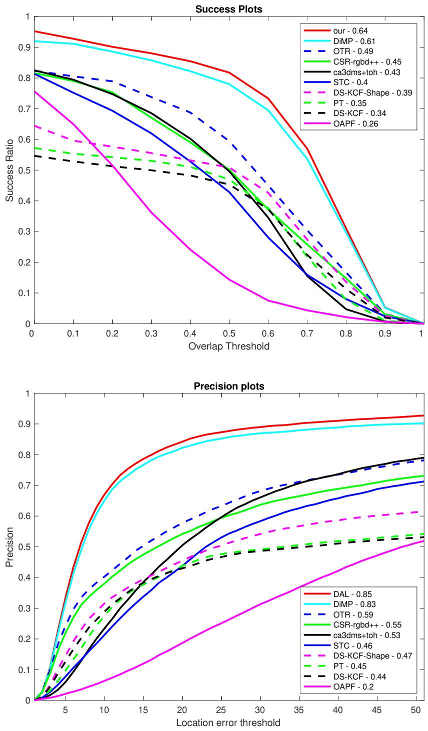

Fig. 4. Success and precision plots on STC benchmark [2].

consistent across all the attributes, except SCC (Surrounding Color Clutter).

The CDTB dataset [3] 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.

Tracking results are presented in Fig 6 & Fig 5. The proposed tracker outperforms the top-performer in CDTB [3], 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% mostly due to the better target representation including deep features and the deep non-stationary DCF. The proposed tracker outperforms SoTA RGB short-term DiMP *wrt.* 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 [6] 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 [30]. This complicated architecture 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 in our tests) and has large memory footprint due to several network branches. On the other hand, DAL has simpler architecture and runs 10× faster thanks to efficient use of depth, while attaining comparable tracking accuracy.

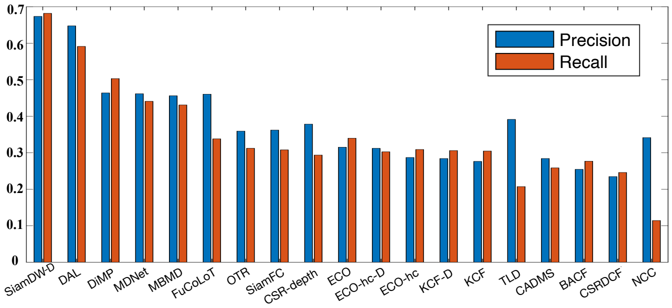

Fig. 5. Tracking precision and recall calculated at the optimal point (maximum F-measure). Evaluated on the CDTB dataset.

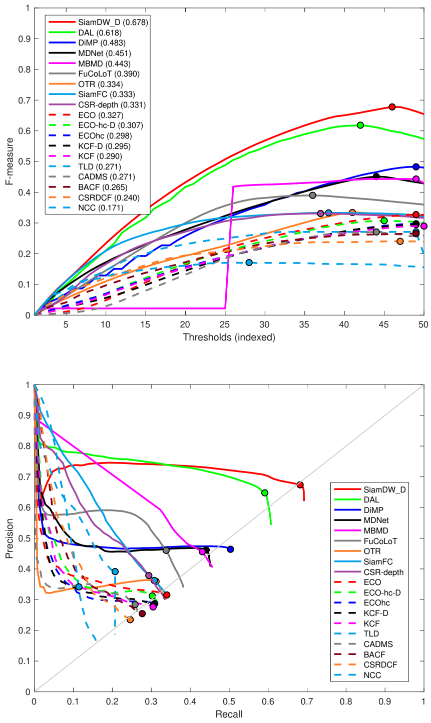

Fig. 6. The overall tracking performance on the CDTB dataset is presented as tracking F-measure (top) and tracking Precision-Recall (bottom). Trackers are ranked by their optimal tracking performance (maximum F-measure).

C. Ablation Study

The following variants of DAL were evaluated: (i) the proposed tracker (DAL), (ii) depth trigger is omitted (DAL$^{-LT(\tau_D)}$), (iii) depth trigger is omitted and the base DCF is used (DAL$^{-\alpha,-LT(\tau_D)}$), (iv) no long-term tracking

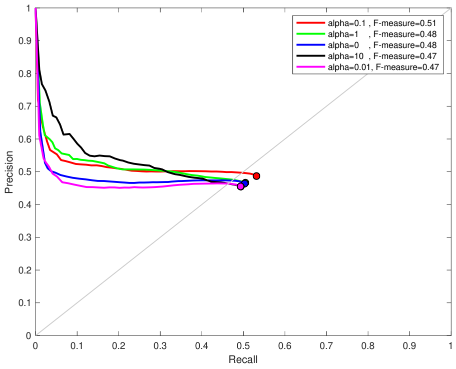

Fig. 7. Precision-Recall curves and F-measure as function of varying α. for depth-modulated DCF. Evaluated on CDTB dataset.

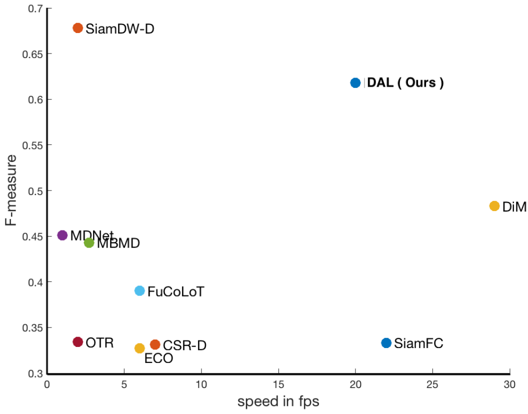

Fig. 8. Tracker practicality evaluation with respect to F-measure and Speed (fps) on the CDTB dataset.

(DAL^{-LT}), (v) no long-term tracking and using the base DCF (DAL$^{-\alpha,-LT}$).

The CDTB F-measure for the variants is (i) **0.62**, (ii) 0.58, (iii) 0.55, (iv) 0.51 and (v) 0.48 with proportional improvements of (iv) +6.3%, (iii) +14.6%, (ii) +20.8% and (i) +29.1% to the baseline (v). These results clearly indicate the contributions of each proposed element in the DAL tracker.

The α parameters in (3) 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.

Speed vs. accuracy was evaluated for the ten top-performing trackers on the CDTB dataset. 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.

V. Conclusions

We proposed a novel deep DCF formulation for RGBD tracking. The formulation is based on the concept of non-stationary DCF where tracker features are modulated by the depth content of the target regions during tracking. The simple formulation provides a strong short-term RGBD tracker, improving the performance from 5% to 6% on all RGBD tracking benchmarks. We also proposed a long-term tracking model where the same non-stationary DCF is used in target re-detection. Simple DCF and depth similarity scores effectively trigger between the modes (tracking vs. re-detection) and when the target model is updated. The DAL tracker consistently achieves superior performance over the state-of-the-art RGB and RGBD trackers (DiMP and OTR) on the three available RGBD tracking benchmarks (PTB, STC and CDTB) and runs significantly faster than the winner of the last VOT-RGBD challenge (CDTB) - 2 fps vs. 20 fps.

References

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

[2] 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.

[3] 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.

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

[5] 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.

[6] 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.

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

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

[9] 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.

[10] 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.

[11] 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.

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

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

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

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

[20] 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.

[21] 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.

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

[23] 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.

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

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

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

[27] 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.

[28] P. Springstübe, S. Heinrich, and S. Wermter, “Continuous convolutional object tracking,” in *ESANN*, 2018.

[29] 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.

[30] 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.