ICPR 2020
25th INTERNATIONAL CONFERENCE ON PATTERN RECOGNITION
Milan, Italy 10 | 15 January 2021

DAL: A Deep Depth-Aware Long-term Tracker

Yanlin Qian[1]*, Song Yan[1]*, Alan Lukežič†, Matej Kristan†, Joni-Kristian Kämäräinen*, Jiri 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

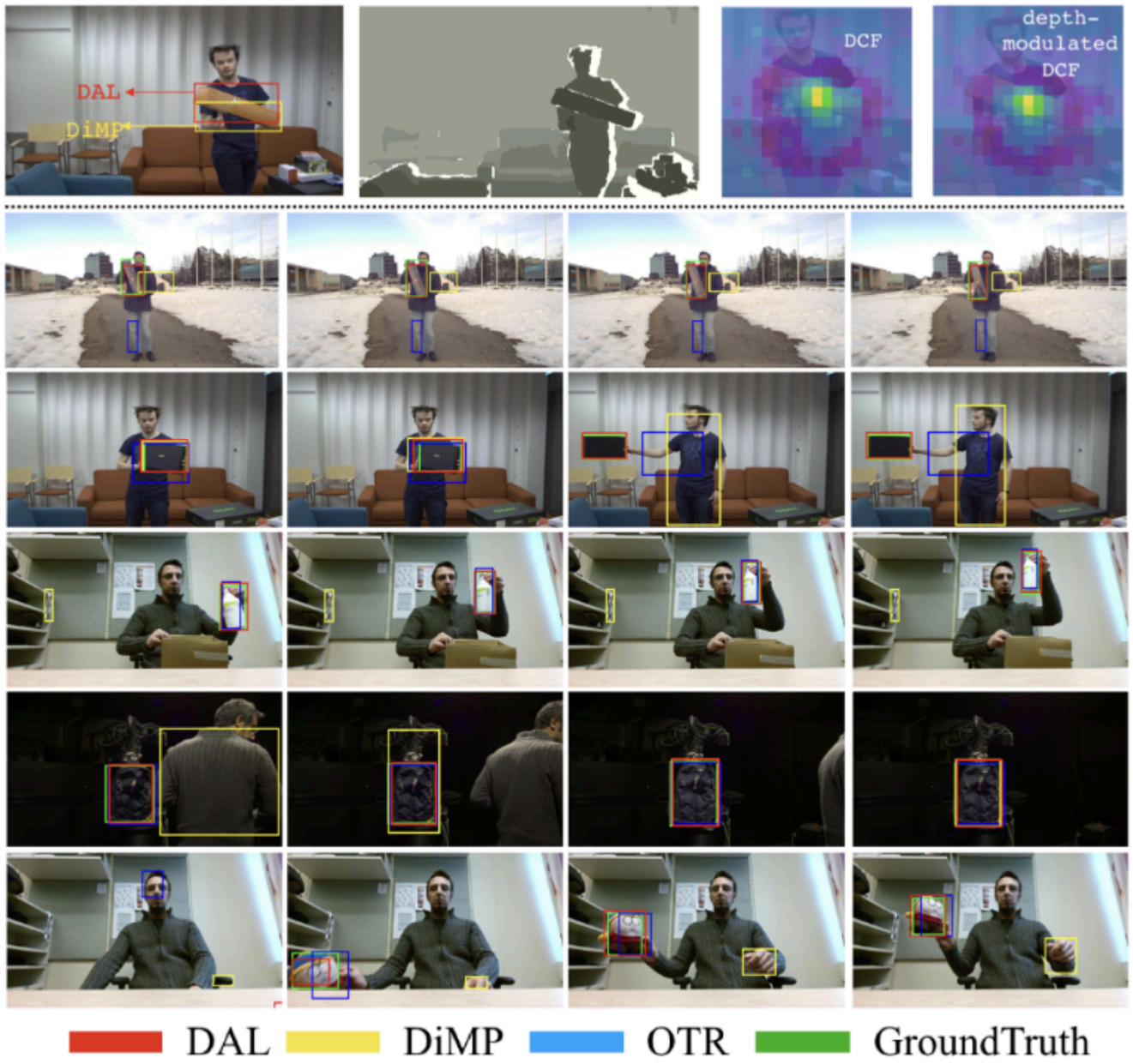
1 equal contribution

Tampere University

CZECH TECHNICAL UNIVERSITY IN PRAGUE

University *of Ljubljana*
Faculty *of Computer and Information Science*

Introduction



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.

Related work

- PTB[1] 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*. [2] addresses RGBD tracker with occlusion awareness and Bibi *et.al.* [3] further models a target using sparse 3D cuboids.

- Based on KCF, Hannuna *et.al*. [4] uses depth for occlusion detection and An *et.al*. [5] extends KCF with depth channel.

- Liu *et.al*. [6] presents a 3D mean-shift-based tracker.

- Kart *et.al*. [7] applies graph cut segmentation on color and depth information, generating better foreground mask for training CSR-DCF [8]. They then extend the idea with building an object-based 3D model [9], relying on a SLAM system Co-Fusion [10].

[1] Song, Shuran, and Jianxiong Xiao. "Tracking revisited using RGBD camera: Unified benchmark and baselines." ICCV2013
[2] Meshgi, Kourosh, et al. "An occlusion-aware particle filter tracker to handle complex and persistent occlusions." *CVIU* 150 (2016): 81-94.
[3] Bibi, Adel, Tianzhu Zhang, and Bernard Ghanem. "3d part-based sparse tracker with automatic synchronization and registration." CVPR2016
[4] Hannuna, Sion, et al. "Ds-kcf: a real-time tracker for rgb-d data." *Journal of Real-Time Image Processing* (2019): 1-20.
[5] An, Ning, Xiao-Guang Zhao, and Zeng-Guang Hou. "Online RGB-D tracking via detection-learning-segmentation." *ICPR*2016.
[6] Liu, Ye, et al. "Context-aware three-dimensional mean-shift with occlusion handling for robust object tracking in RGB-D videos." *TMM* 21.3 (2018): 664-677.
[7] Kart, Ugur, Joni-Kristian Kamarainen, and Jiri Matas. "How to make an RGBD tracker?." *ECCV-W*2018.
[8] Lukezic, Alan, et al. "Discriminative correlation filter with channel and spatial reliability." *CVPR*2017.
[9] Kart, Ugur, et al. "Object tracking by reconstruction with view-specific discriminative correlation filters." *CVPR*2019.
[10] Rünz, Martin, and Lourdes Agapito. "Co-fusion: Real-time segmentation, tracking and fusion of multiple objects." *ICRA*2017.

Proposed method

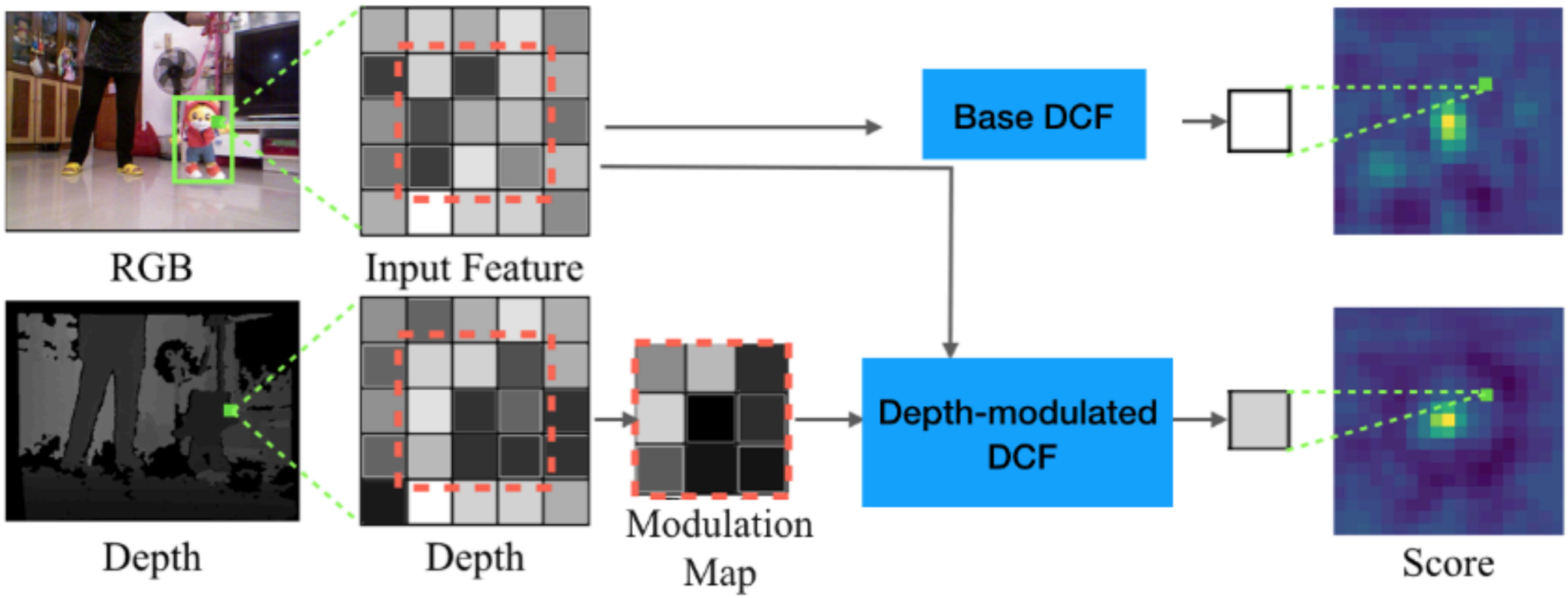


Fig. 2. Visualization of depth-modulated DCF. Depth modulates the DCF by re-weighting the DCF kernels according to the depth similarity with the tested target position. Top: the confidence score map of the target object resulting from base DCF; Bottom: the corresponding score map obtained by our depth-modulated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

Proposed method

Base DCF is optimized by steepest descent by steepest descent on the following loss :

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

We utilize depth to modulate the DCF content with respect to the filter position :

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

The modulation map is then defined as :

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

The loss for training the non-stationary DCF becomes :

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

Long-term tracker architecture

The maximum of the depth-modulated DCF correlation response, ρ_{DCF} **indicates the target presence likelihood.** Thus the correlation-based target presence indicator is defined as :

$$\beta_{\mathrm{DCF}}(\tau) = \{1 : \rho_{\mathrm{DCF}} > \tau; 0 : \text{ otherwise}\}.$$

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

$$\rho^i_{\mathrm{dep}} = \sum_j^{n_B} \sqrt{\mathscr{H}_j \mathscr{G}^i_j},$$

The depth consistency indicator is therefore defined as :

$$\beta_{\mathrm{dep}}(\tau) = \{1 : \rho^i_{\mathrm{dep}} > \tau \ \forall \ i; 0 : \text{ otherwise}\}$$

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

Experiments

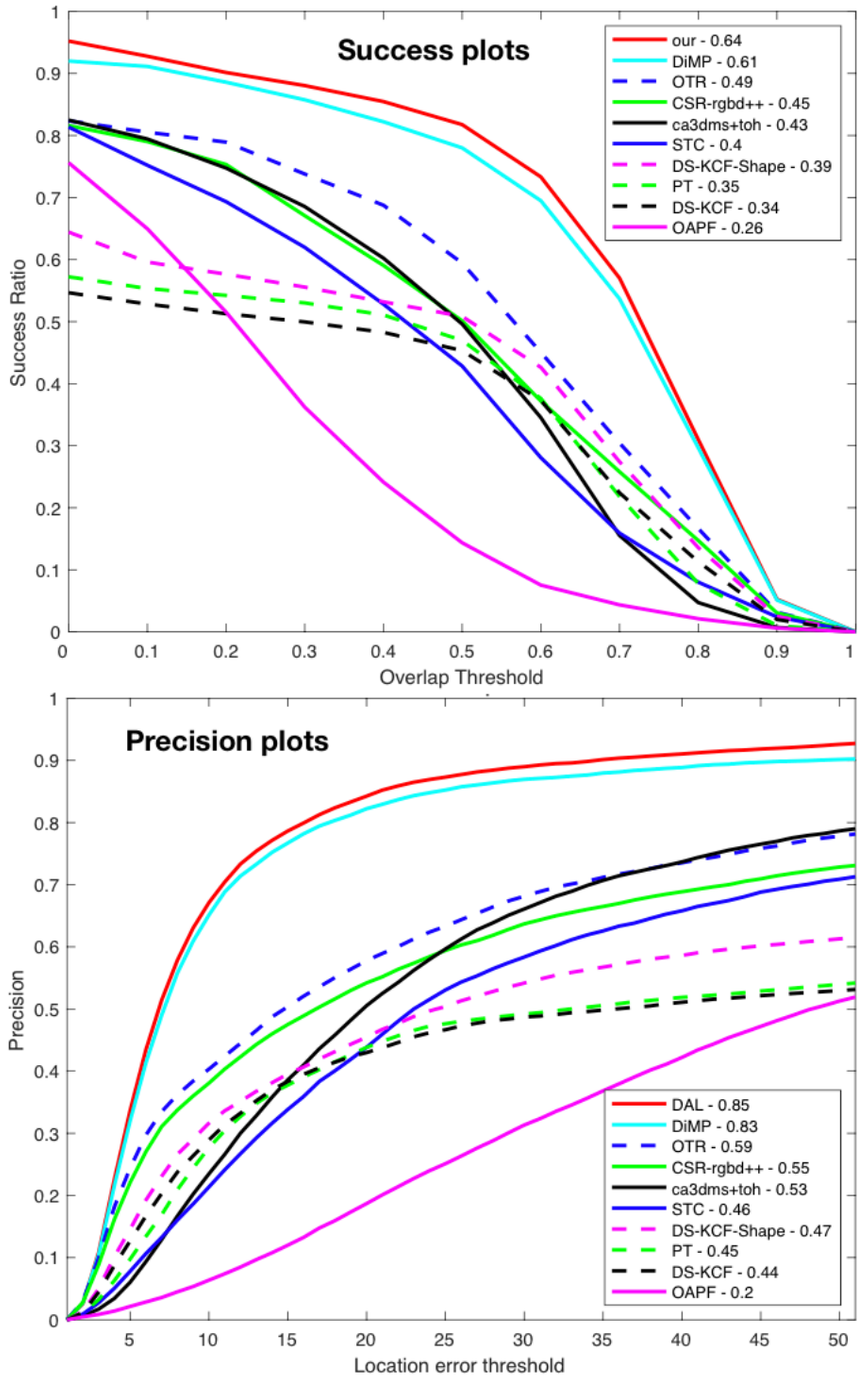


Fig. 4. Success and precision plots on STC benchmark [9].

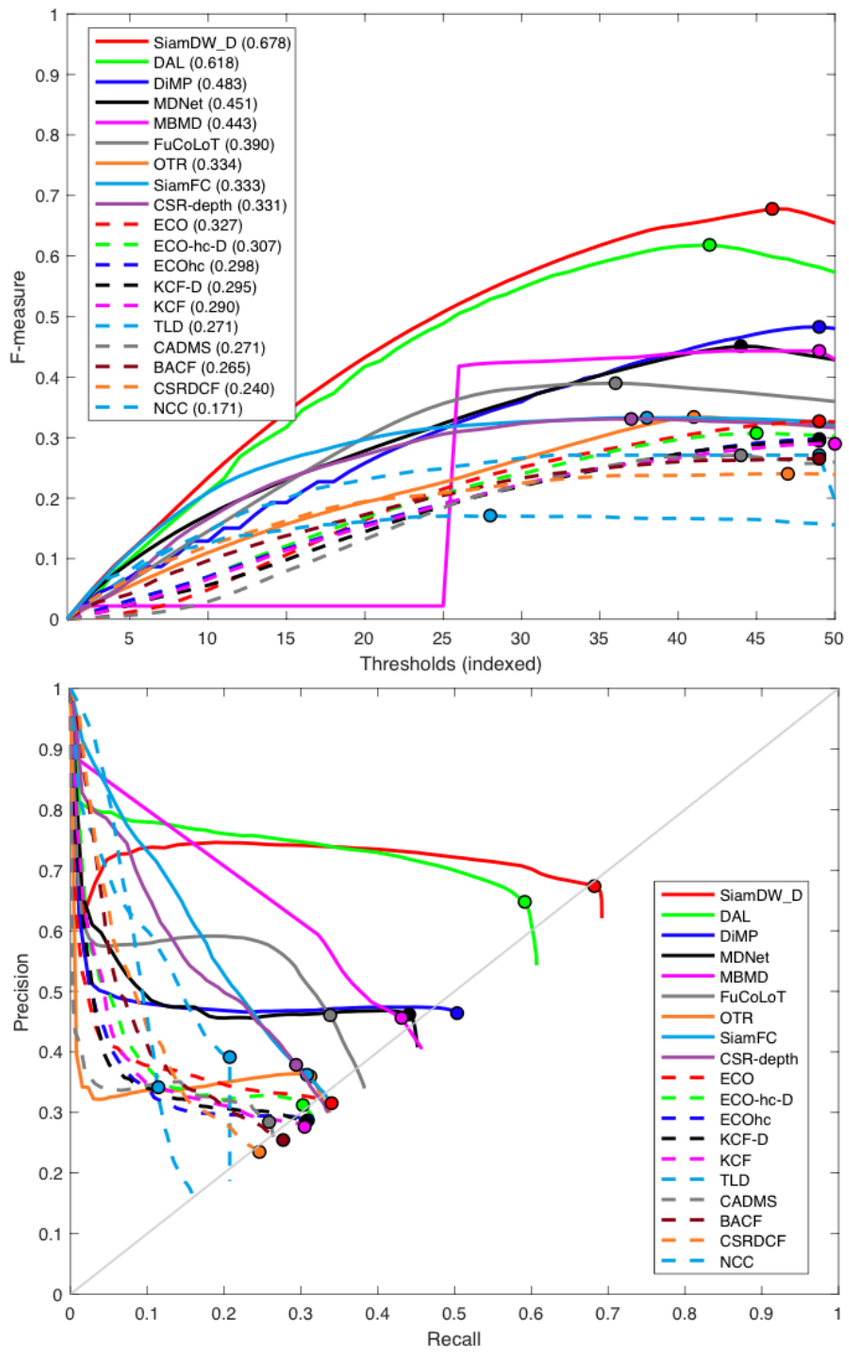


Fig. 5. The overall tracking performance is presented as tracking F-measure (top) and tracking Precision-Recall (bottom) on the CTDB dataset. Trackers are ranked by their optimal tracking performance (maximum F-measure).

Experiments

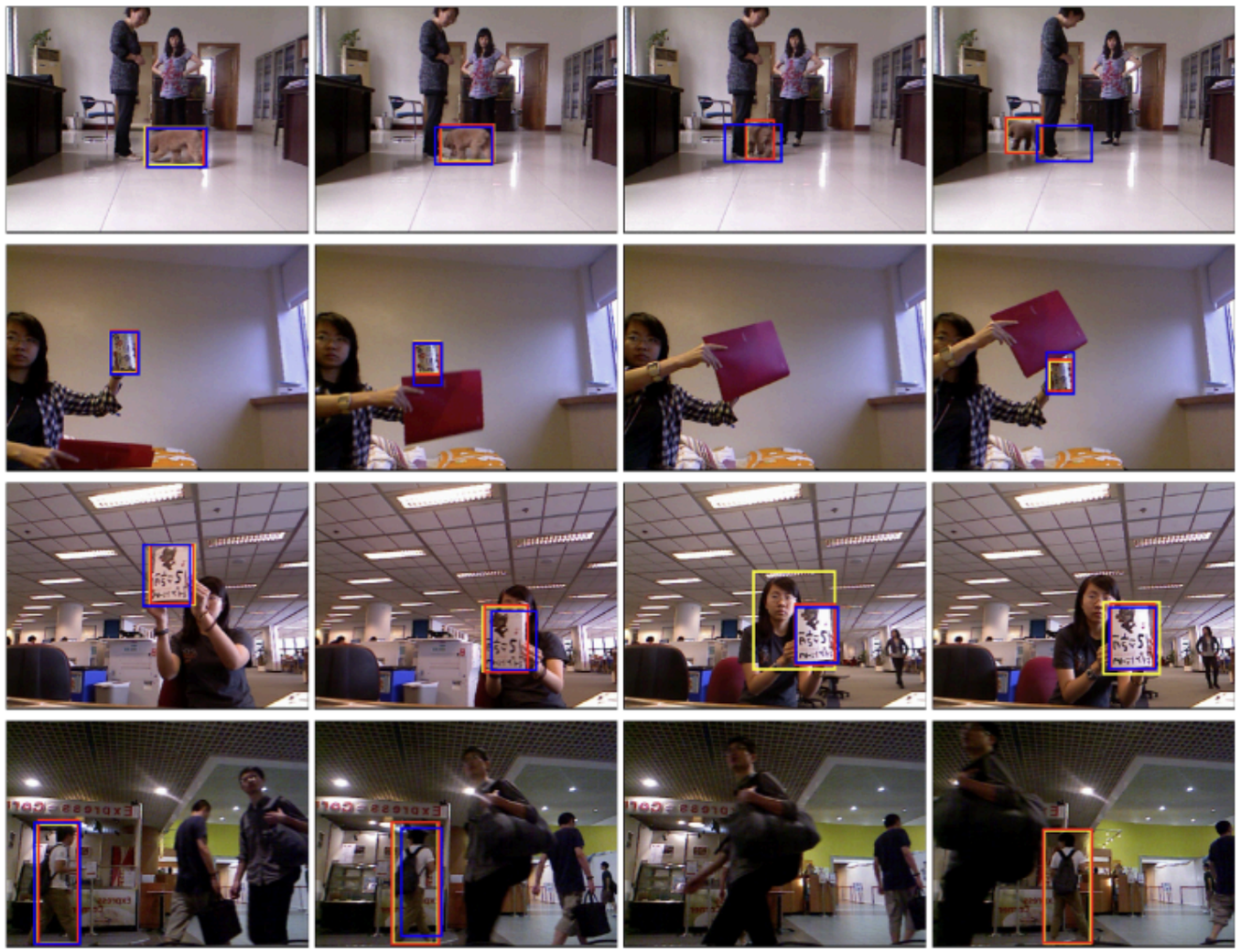


Fig. 3. Qualitative comparison of DAL, DiMP and OTR on the PTB. All trackers localize the target and give precise bounding boxes (the first two columns). With depth-modulated DCF, our tracker shows better discriminative ability when strong distractor appears (human face in the third row and human legs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

Experiments

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

Experiments

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

Experiments

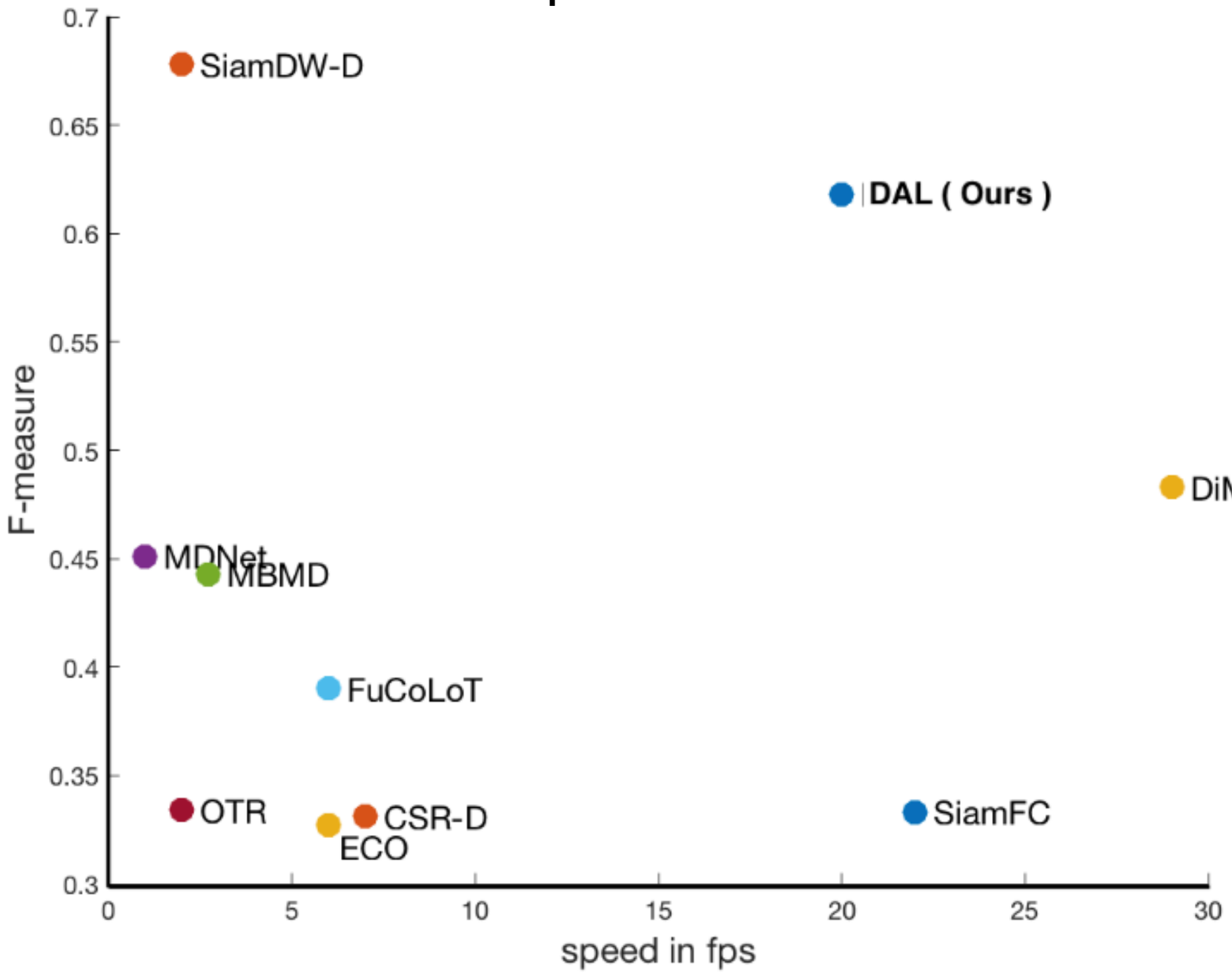


Fig. 8. Tracker practicality evaluation with respect to F-measure and Speed (in frames-per-second) on the CDTB dataset

Conclusion

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).