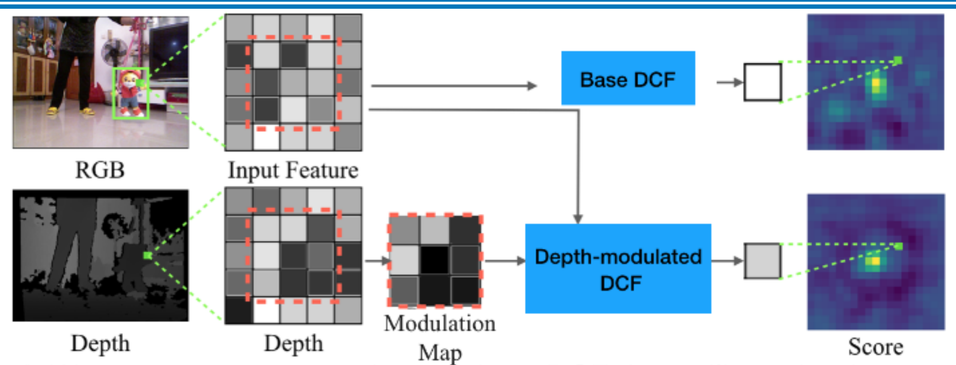
DAL: A Deep Depth-Aware Long-Term Tracker

Yanlin Qian*, Song Yan*, 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

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

Abstract: 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.



Method: We propose a non-stationary deep DCF that utilizes depth to modulate the DCF content with respect to the filter position:

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

where $\mathbf{f}$ is a stationary base filter, $\boldsymbol{\Theta}(x,y)$ 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:

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

where α is a hyper parameter that controls the modulation strength.

The loss for training the non-stationary DCF becomes :

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

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 $z \le T$, where T is a threshold on the error.

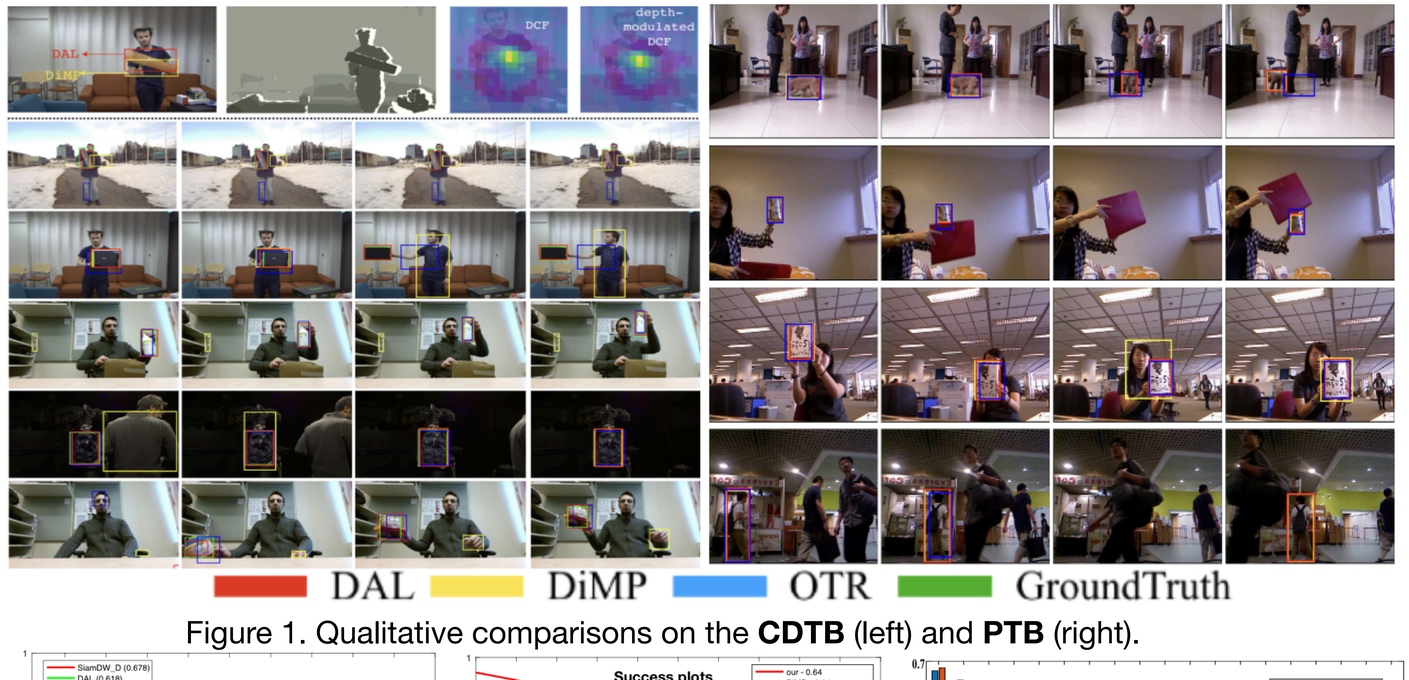


Figure 1. Qualitative comparisons on the **CDTB** (left) and **PTB** (right).

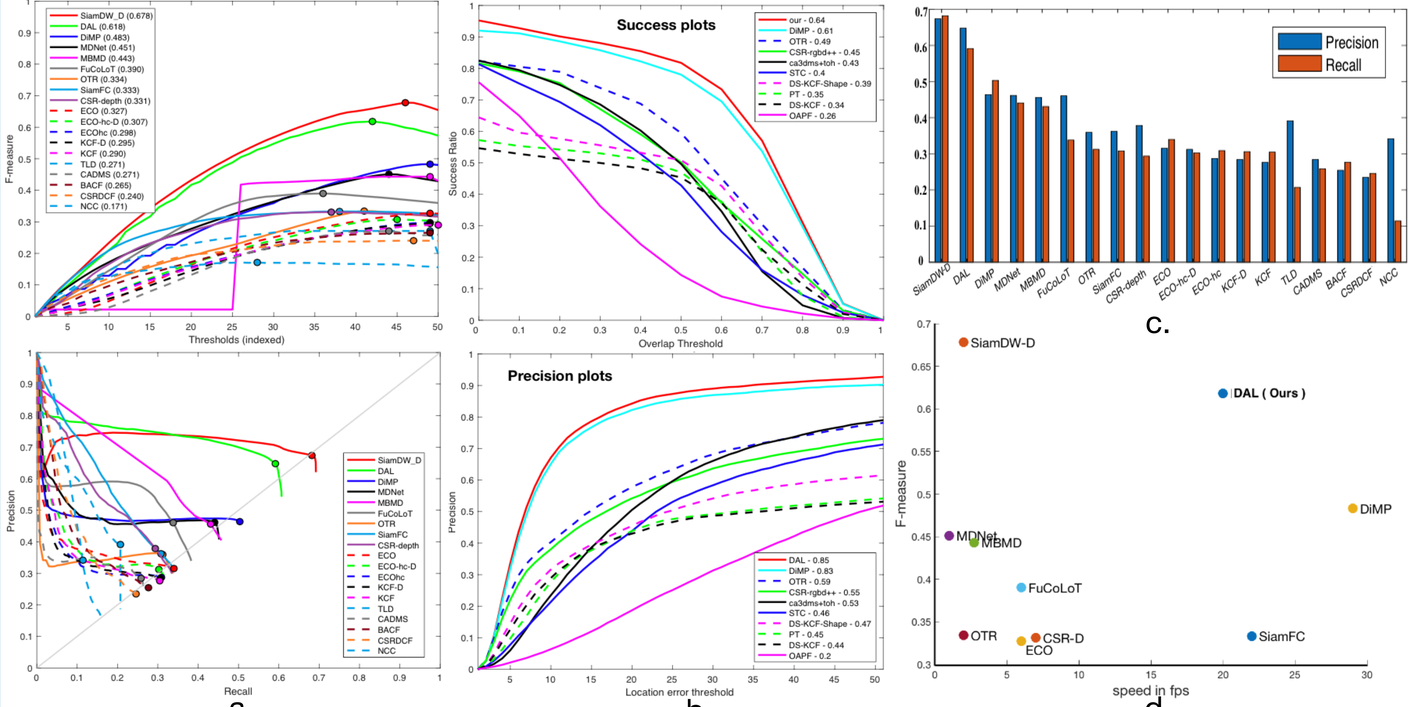


Figure 2. **a**) The overall tracking performance is presented as tracking F-measure (top) and tracking Precision-Recall (bottom) on the CTDB dataset. **b**) Success and precision plots on STC benchmark. **c**) The overall tracking performance is presented as tracking F-measure (top) and tracking Precision-Recall (bottom) on the CTDB dataset. **d**) Precision-Recall curves and F-measure as function of varying α for depth-modulated DCF. Evaluated on CDTB dataset