- Supplementary Material -
DepthTrack: Unveiling the Power of RGBD Tracking

Song Yan[1,†], Jinyu Yang[2,3,†], Jani Käpylä[1,†], Feng Zheng[2], Aleš Leonardis[3], Joni-Kristian Kämäräinen[1]

[1]Tampere University [2]Southern University of Science and Technology [3]University of Birmingham

{song.yan,jani.kapyla,joni.kamarainen}@tuni.fi, jinyu.yang96@outlook.com,
zhengf@sustech.edu.cn, a.leonardis@cs.bham.ac.uk

1. Performance Metrics

Following the settings of the VOT challenges [3, 4], we firstly employ the average performance over all sequences as follows:

$$\begin{aligned} Pr(\tau_\theta) &= \frac{1}{N}\sum_{N}^{i} Pr^i(\tau_\theta), \\ Re(\tau_\theta) &= \frac{1}{N}\sum_{N}^{i} Re^i(\tau_\theta), \end{aligned} \tag{1}$$

where $Pr^i(\tau_\theta)$ and $Re^i(\tau_\theta)$ denote the precision (*Pr*) and recall (*Re*) over frames in the i^{th} sequence of all N test videos. The confidence threshold is denoted as τ_θ. We refer the above evaluation as *sequence-based evaluation*.

To better handle the imbalance problem of the video lengths, we also propose a *frame-based evaluation* metric as follows :

$$\begin{aligned} Pr(\tau_\theta) &= \frac{1}{N_p}\sum_{t\in\{t:A_t(\tau_\theta)\neq\varnothing\}} \Omega(A_t(\tau_\theta), G_t), \\ Re(\tau_\theta) &= \frac{1}{N_g}\sum_{t\in\{t:G_t\neq\varnothing\}} \Omega(A_t(\tau_\theta), G_t), \\ F(\tau_\theta) &= \frac{2Re(\tau_\theta)Pr(\tau_\theta)}{Re(\tau_\theta)+Pr(\tau_\theta)}, \end{aligned} \tag{2}$$

where $\Omega(A_t(\tau_\theta), G_t)$ indicates the intersection-over-union (IoU) between prediction result and groundtruth, and $F(\tau_\theta)$ is the F-score metric. G_t denotes the groundtruth of the target and $A_t(\tau_\theta)$ denotes the corresponding prediction at frame t. If the predicted confidence score θ_t at frame t is below τ_θ, then the output is an empty set $A_t(\tau_\theta) = \varnothing$. N_p denotes the number of frames in which the target is predicted visible, and N_g denotes the number of frames in which the target is indeed visible.

[†]Equal contribution.

2. Visual Attributes

The tracking performance on a particular attribute is to verify how trackers behave in a specific scenario. The optimal F-score over all frames for a specific attribute is adopted to measure the performance for the scenarios when the object is visible. For the cases of target disappearance, including *out-of-frame* and *full occlusion*, we apply the method of binary classification according to OxUvA [6]. The definitions of all visual attributes are listed in Table 1.

The object presence and absence are separately treated as positive and negative class. We declare that the situation, where the target is invisible and predicted as absent, is true negative (TN). True-negative rate (TNR) is defined with the ratio of present objects predicted as absence. It is used to evaluate the attributes of *out-of-frame* and *full occlusion*. Please refer to [6] for the details of TNR.

$$TNR(\tau_\theta) = \frac{1}{N}\sum_{N}^{t}(\theta_t < \tau_\theta), \tag{3}$$

where θ_t denotes the predicted confidence score at frame t, and τ_θ denotes the confidence threshold. N denotes the number of frames belonging to the attribute *out-of-frame* or *full occlusion*. The optimal TNR is the averaged value over all possible τ_θ.

3. Quantitative Results

Sequence-based *vs.* frame-based metrics. We proposed the frame-based evaluation metric to alleviate the imbalance problems of the video lengths. For most trackers, the frame-based results are slightly lower than the sequence-based results. Since the former one focuses more on the longer sequences, it indicates that long sequences are more challenging compared to the short ones. Thus, we can see that the frame-based protocols evaluated on each frame rather than the entire sequence obtains a fairer accuracy

Table 1: Tracking visual attributes including 10 manually annotated attributes and 5 ones calculated from the groundtruth.

Attribute	Tag	Description	Annotation
Aspect-ratio Change	AC	When the ratio between maximum and minimum target size in 21 consecutive frames was larger than 1.5.	Calculated
Background Clutter	BC	Background near the target has the similar appearance as the target.	Manually
Camera Motion	CM	The camera view is not fixed.	Manually
Dark Scene	DS	No visible light is in the scenario.	Manually
Depth Change	DC	The ratio between maximum and minimum of depth median in target region in 21 frames was larger than 1.5.	Calculated
Fast Motion	FM	The target center moves by at least 30% of its size in consecutive frames.	Calculated
Full Occlusion	FO	The target is fully occluded.	Manually
Non-rigid Deformation	ND	The non-rigid object deforms.	Manually
Out-of-plane Rotation	OP	Target rotates out of the plane.	Manually
Out-of-frame	OF	Partial or the whole target leaves the view.	Manually
Partial Occlusion	PO	The target is partially occluded.	Manually
Reflective Targets	RT	Interface of the target is reflective.	Manually
Size Change	SC	When the ratio between maximum and minimum target size in 21 consecutive frames is larger than 1.5.	Calculated
Similar Objects	SO	There is adjacent objects whose appearance is similar to the target.	Manually
Unassigned	NaN	There is no aforementioned cases appeared in the frame.	Calculated

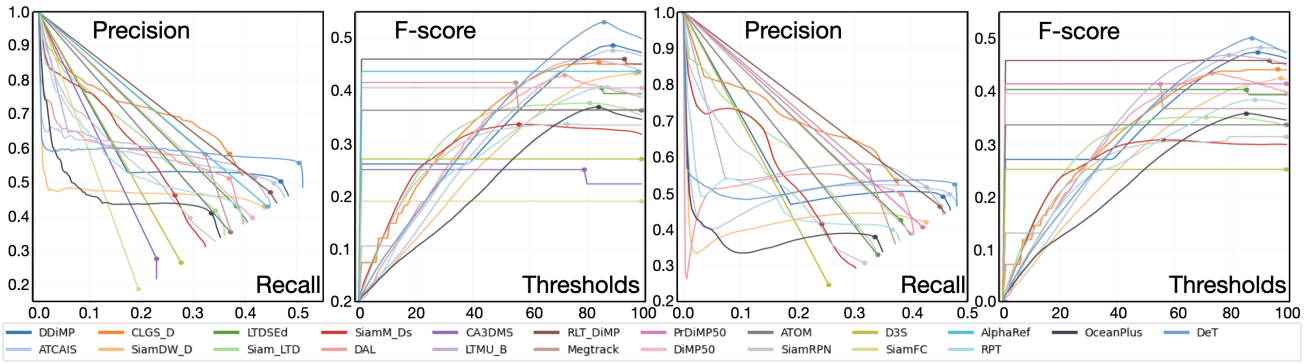


Figure 1: The Precision-Recall and F-score curves of the evaluated trackers (the best F-score point marked in each graph). Left: sequence-based; Right: frame-based evaluation.

evaluation. The Precision-Recall and F-Score plots of all evaluated trackers are shown in Fig. 1

4. The Variants of DeT

We choose ATOM [2] and DiMP50 [1] as "masters" for DeT, which use the ResNet-18 and ResNet-50 RGB backbones, respectively. The architectures of the variants, DeT-ATOM and DeT-DiMP50 are shown in Fig. 4a and Fig. 4b.

5. Cross-Dataset Evaluation

Attribute-based performance analysis. F-scores of each attribute for the proposed DeT and its variants on the CDTB benchmark [5] are shown in Fig. 2. The variant DeT-DiMP50-Max wins 8 of 13 attributes while DeT-DiMP50-Max wins 3 (FM, ND and PO) and DeT-ATOM-Mean wins 2 (DS and OF). **Qualitative results.** Fig. 3 shows the qualitative results of the baseline trackers and our proposed two DeT variants including DeT-DiMP50-Max and DeT-ATOM-Max, on the CDTB benchmark. Both RGB and RGBD trackers perform well in most sequences, *e.g.* the

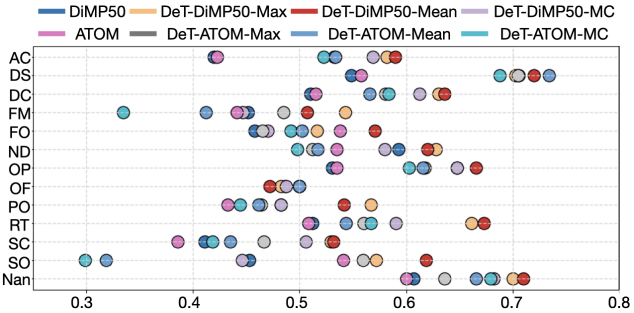


Figure 2: Optimal F-scores for the visual attributes on CDTB dataset [5]. ATOM and its variants perform very similar on the *out-of-frame* (OF) attribute.

sequences containing single target (2nd row) and the sequences of simple background (3rd and 6th rows). CDTB contains few sequences in which depth values are missing, *e.g.* 7th row of Fig. 3. We can see that the proposed variants can perform well in the case of depth data missing.

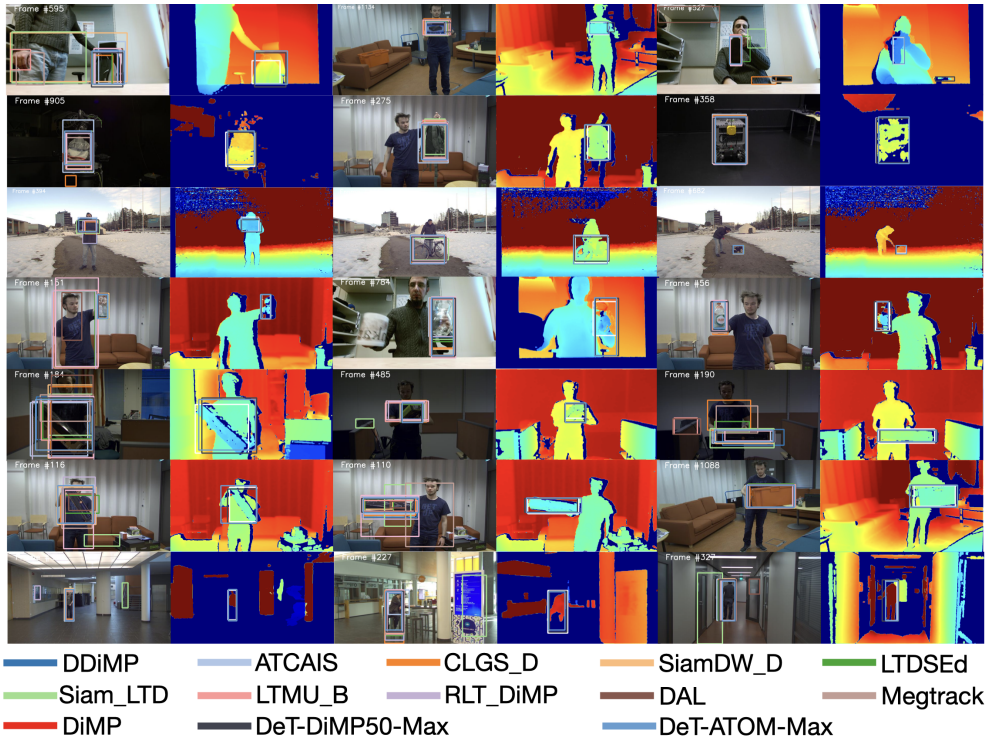
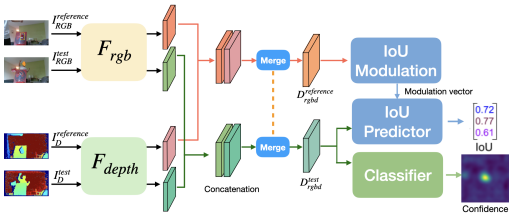
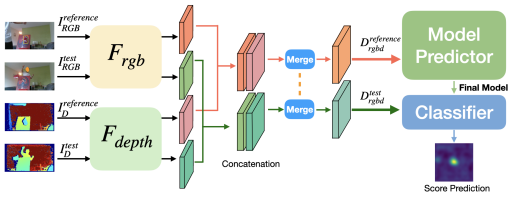


Figure 3: Examples from the CDTB benchmark (targets marked with white boxes in depth images): similar background color (1st row), single targets (2nd row), wild scenes (3rd row), glasses (4th row), dark scenes (5th row), long boxes (6th row), depth missing scenes (7th row). In addition to the existing baselines, the DeT-DiMP50-Max and DeT-ATOM-Max outputs are marked in the depth images.



(a) DeT-ATOM



(b) DeT-DiMP50

Figure 4: The architectures of the DeT variants.

References

[1] Goutam Bhat, Martin Danelljan, Luc Van Gool, and Radu Timofte. Learning discriminative model prediction for tracking. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6182–6191, 2019.

[2] Martin Danelljan, Goutam Bhat, Fahad Shahbaz Khan, and Michael Felsberg. Atom: Accurate tracking by overlap maximization. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[3] Matej Kristan, Aleš Leonardis, Jiří Matas, Michael Felsberg, Roman Pflugfelder, Joni-Kristian Kämäräinen, Martin Danelljan, Luka Čehovin Zajc, Alan Lukežič, Ondrej Drbohlav, et al. The eighth visual object tracking vot2020 challenge results. In *European Conference on Computer Vision*, pages 547–601. Springer, 2020.

[4] Matej Kristan, Jiri Matas, Ales Leonardis, Michael Felsberg, Roman Pflugfelder, Joni-Kristian Kamarainen, Luka Cehovin Zajc, Ondrej Drbohlav, Alan Lukezic, Amanda Berg, et al. The seventh visual object tracking vot2019 challenge results. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019.

[5] Alan Lukezic, Ugur Kart, Jani Kapyla, Ahmed Durmush, Joni-Kristian Kamarainen, Jiri Matas, and Matej Kristan. Cdtb: A color and depth visual object tracking dataset and benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10013–10022, 2019.

[6] Jack Valmadre, Luca Bertinetto, Joao F Henriques, Ran Tao, Andrea Vedaldi, Arnold WM Smeulders, Philip HS Torr, and Efstratios Gavves. Long-term tracking in the wild: A benchmark. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 670–685, 2018.