Depth-only Object Tracking

Song Yan¹
song.yan@tuni.fi
Jinyu Yang²³
jinyu.yang96@outlook.com
Aleš Leonardis³
a.leonardis@cs.bham.ac.uk
Joni-Kristian Kämäräinen¹
joni.kamarainen@tuni.fi

- ¹ Tampere University Finland
- ² Southern University of Science and Technology China
- ³ University of Birmingham United Kingdom

Abstract

Depth (D) indicates occlusion and is less sensitive to illumination changes, which make depth attractive modality for Visual Object Tracking (VOT). Depth is used in RGBD object tracking where the best trackers are deep RGB trackers with additional heuristic using depth maps. There are two potential reasons for the heuristics: 1) the lack of large RGBD tracking datasets to train deep RGBD trackers and 2) the long-term evaluation protocol of VOT RGBD that benefits from heuristics such as depth-based occlusion detection. In this work, we study how far D-only tracking can go if trained with large amounts of depth data. To compensate the lack of depth data, we generate depth maps for tracking. We train a "Depth-DiMP" from the scratch with the generated data and finetune it with the available small RGBD tracking datasets. The depth-only DiMP achieves good accuracy in depth-only tracking and combined with the original RGB DiMP the end-to-end trained RGBD-DiMP outperforms the recent VOT 2020 RGBD winners.

1 Introduction

The Visual Object Tracking (VOT) Challenge has been running since 2013. In 2019 a new challenge, VOT-RGBD, and dataset for RGBD tracking were introduced. Interestingly, the winners of VOT-RGBD 2019 [and 2020 [are actually RGB-only trackers (MD-Net [], ATOM [] and DiMP []) with additional heuristics using the D channel. On one hand, VOT-RGBD is a long-term tracking challenge that benefits from heuristic full occlusion detection, and on the other hand, the lack of large RGBD tracking datasets makes it impossible to train RGBD trackers from scratch. It remains unclear to what degree depth is useful in object tracking.

In this work, we aim to answer the above research question by training a deep depthonly tracker from scratch using depth tracking data and evaluating it with the previous VOT-RGBD datasets (CDTB [] in 2019-2021 and DepthTrack [] in 2021). This work provides the following novel contributions: 1) we introduce synthetic depth-only tracking datasets that are generated from the large RGB datasets using monocular depth estimation; 2) we introduce a depth-only DiMP (Depth-DiMP) and its extension RGBD-DiMP which are trained using both generated and real data. With the help of Depth-DiMP, 3) we report a careful analysis of depth-only tracking and compare the performance of RGB and D-only trackers on CDTB and DepthTrack datasets. Interestingly, our Depth-DiMP is almost on par with the best RGB trackers and RGBD-DiMP wins the VOT-RGBD winners on the Depth-Track test set. Our analysis shows that there are scenes for which depth is an important complementary cue for object tracking.

2 Related work

Visual object tracking. In recent years, there has been steady progress from conventional "engineered" trackers toward learning-based deep trackers. Discriminative Correlation Filter (DCF) [1] is a representative example of the conventional trackers. DCF formulation leads to optimized detection filters that are used to find the target in the next frame. This principled idea has inspired the main architecture of the two well-performing deep trackers: ATOM [1] and DiMP [1]. Their architectures have three processing modules: i) feature extraction, ii) target classifier and iii) target estimation. For the depth-only D and RGBD trackers in this work we adopt the DiMP architecture. A DiMP extension, Probabilistic DiMP (PrDiMP) [1], has been proposed, but we adopted the original DiMP due to its simplicity and stable public implementation.

RGBD and D Object Tracking. RGBD tracking has been a separate track in the VOT Challenge since 2019 [III]. The three best trackers in the VOT2020-RGBD challenge [III], ATCAIS, DDiMP and CLGSD, are RGBD extensions of the deep RGB trackers, ATOM, DiMP and MDNet [III]. They use depth channel to support the RGB trackers with heuristic processing. Our work differs substantially from all previous RGBD trackers in the sense that our tracker is data-driven similar to the original ATOM and DiMP. To the best of authors' knowledge there are no other works addressing the generic object tracking with depth only.

RGBD tracking datasets. Deep RGBD tracker research suffers from the fact that there are no high quality and large RGBD tracking datasets. The only available datasets are PTB, STC and CDTB. PTB [was perhaps the first published RGBD tracking benchmark and contains 100 video sequences, but the groundtruth is not publicly available. STC [is more recent and of better quality, but contains only 36 sequences. The largest publicly available dataset is CDTB [is that contains 80 indoor and outdoor sequences captured with multiple different sensors. CDTB is also used in the VOT-RGBD 2019-2021 evaluations [id , id , id]. VOT2021-RGBD provides an additional dataset, DepthTrack [id], from which we use 120 sequences divided to training (70) and test (50) sets. Cross-dataset comparisons are made between CDTB and DepthTrack.

3 Methods

In this section, we briefly describe the original DiMP and then introduce its depth-only, Depth-DiMP, and RGBD variants that differ only in their feature extraction parts.

3.1 DiMP

The ATOM [2] and DiMP [2] trackers share the same overall architecture: 1) a pre-trained feature extraction module, 2) an offline trained target estimation module and 3) an online

trained classifier module.

Feature extraction. DiMP employs the ResNet [1] network as its feature extraction module. The deeper ResNet-50 provides better performance, but is slower than the lighter ResNet-18 (43 vs. 57 fps). We use ResNet-50 features extracted from the same layers as in the original DiMP. ResNet is pre-trained with the ImageNet image categorization dataset [1] to learn strong semantic features. Ideally, ResNet should be trained with D and RGBD depth data for the D and RGBD trackers, but since there are no large semantic tasks depth datasets, we utilize RGB-trained ResNet features as the initial features for both RGB and depth.

Target classifier. The first step in the DiMP processing pipeline is to estimate the target location in the current frame given its location in the previous frame (*reference frame*). The target location is estimated by an online trained *target classifier* module. The ATOM classifier module is a simple two-layer fully connected regressor network,

$$f(x; w) = \phi_2(w_2 * \phi_1(w_1 * x)) \tag{1}$$

where $w = \{w_1, w_2\}$ are the network weights, * is the multi-channel convolution, ϕ_s are activation functions and x is the ResNet feature map. During tracking negative and positive examples are generated from the reference frame and the weights are re-optimized over a small number of iterations using the Gauss-Newton method [IX]. It is noteworthy that this step is similar to the DCF formulation where w_2 resembles the discriminative correlation filters for tracking. DiMP takes this step further by introducing a more complex and powerful classifier network. DiMP optimization is based on the regularized loss

$$L(f) = \frac{1}{|S_{train}|} \sum_{(x,c) \in S_{train}} || r(x \circledast f, c) ||^2 + || \lambda f ||^2$$
 (2)

where f are the filter weights to be optimized and c denotes the desired output response for the randomly selected training samples S_{train} . \circledast denotes convolution and λ is a regularization factor. Suitable candidates for c are a Gaussian peaked at the groundtruth location or a triangular basis function of DiMP. DiMP defines an error residual $r(\cdot,\cdot)$ that combines least-squares and the hinge loss and is optimized using the Gauss-Newton method.

Target estimation. The target estimation networks of ATOM and DiMP share the same structure. Given the position estimated from the classifier module at the maximum score of the response $x \circledast f$ and the bounding box width and height from the previous frame 10 bounding box proposals are generated by adding uniform random noise. The target estimation module is applied to each proposal. The final estimate is the average of three best proposals after the target estimation stage. The target estimation provides Intersection-over-Union (IoU) score for each proposal B using

$$IoU(B) = g(c(x_0, B_0), z(x, B))$$
 (3)

where g is the predictor network. The predictor network is trained offline with the largest available RGB tracking datasets (e.g. LaSOT [\square] and TrackingNet[\square]) and with augmented examples from the diverse set of classes in COCO [\square]. $c(\cdot, \cdot)$ is the feature representation of the initial region B_0 from the feature map x_0 of reference image while $z(\cdot, \cdot)$ is the proposal representation from the feature map x of the test image.

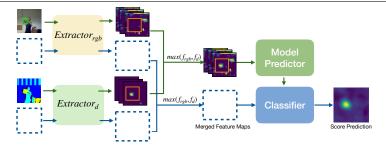


Figure 1: The RGBD-DiMP framework. $max(f_{rgb}, f_d)$ denotes element-wise maximum pooling that combines the two modalities, RGB and D, before the DiMP predictor and classifier modules. To illustrate pooling, high response depth features merged with RGB features are pinpointed with orange boxes. The dashed blue boxes denote the reference branch.

3.2 Depth DiMP

In all experiments we wanted to keep the number of parameters (weights) of the original DiMP fixed so that we could compare the effect of training data only. For different inputs, RGB, D and RGBD, this was achieved by altering only the ResNet-50 based feature extraction module of DiMP. These changes for D and RGBD are briefly explained next.

Depth-DiMP. The input affects only the feature extraction module of DiMP, i.e. ResNet-50 features extracted from the input. The depth values can be encoded to a single channel, but to keep the DiMP structures unchanged and the same total number of parameters (weights) we kept all three channels and experimented with different encoding of the depth. The two encoding experimented where *channel copying* and *colormap encoding*. The channel copying is simply repeating the same normalized depth map three times and is denoted by $3 \times D$. The colormap encoding is performed by mapping pixels of the normalized depth map to RGB vectors of a predefined color matrix (JET used in the experiments). This encoding is denoted by "ColMap". Depth normalization is performed by normalizing all values to [0,1] and before normalization metric depth values are clipped to the range [0.0, 10.0] meters.

RGBD DiMP. To verify that RGBD tracking benefits from learning more realistic deep depth features, we combined RGB DiMP with Depth-DiMP by adding an additional feature extraction module for the depth input. Before the other DiMP modules the RGB and D streams are merged. For merging, we adopt the element-wise maximum pooling. The RGBD-DiMP framework and merging of the two channels are illustrated in Figure 1.

4 Experiments

4.1 Performance metrics and settings

The major difference between long- and short-term trackers is the re-detection mechanism and the most common way is just to brute-force search the previous location with an increasing region until the target is found. To focus on how generic object tracking benefits from depth, we converted the long-term (LT) sequences into short-term (ST) ones by using two

rules to cut the sequences: i) all frames from the beginning (frame #1) are kept until the target disappears (frame #D) and re-appears again (frame #D+R-1); and ii) a new ST sequence is started (frame #D+R+S) when the target size is at least 75% of its size in the frame #1. This procedure is iterated until the end of each LT clip. However, to report numbers consistently with the previous RGBD works we adopted the LT evaluation protocol of the VOT challenge. For each frame, the tracker predicts a target *confidence score* and *bounding box*. The bounding box coordinates are used to calculate the tracking precision (Pr) by measuring the overlap ratio with the groundtruth box. Tracking recall (Re) measures the fraction of the frames labelled as "target present" and for which the confidence score is above a threshold. The overall evaluation F-score is the harmonic mean of Pr and Re. For training the default settings in $[\Box]$ were used.

4.2 Datasets

RGBD datasets for evaluation. For the depth-only (D) and RGBD experiments we used the two largest and most recent RGBD datasets: CDTB [] and DepthTrack []. CDTB was used in the VOT-RGBD challenge from 2019 to 2021 and DepthTrack was introduced as an additional dataset to the 2021 challenge. Our LT-to-ST procedure produced 152 ST sequences from the original 80 CDTB LT sequences. Similarly, the DepthTrack 50 LT test sequences resulted to 124 ST sequences.

The 70 LT training sequences of DepthTrack were used in training as is as the training part does not differ for LT and ST. For the cross-dataset evaluations we trained with the DepthTrack training set and tested with CDTB-ST.

Generated depth data for training. The original DiMP was trained with the training sets of TrackingNet [14], LaSOT [14], GOT10k [144] and COCO [144]. TrackingNet is the largest of these datasets including 30,000 sequences. The largest existing RGBD datasets are PTB, STC, CDTB and DepthTrack of which DepthTrack is the largest containing 120 sequences. Collecting and annotating more RGBD sequences is time consuming and tedious, and therefore we decided to convert the RGB datasets to RGBD using monocular depth estimation. Depth data has much less variation than RGB and therefore we converted the two smaller datasets, LaSOT [15] (1,500 seqs.) and Got10k [15] (10,000 seqs.). We also used the COCO images [15] to train the tracker to have a richer set of objects.

To quantitatively evaluate the generated depth sequences we adopted the *Ordinal Error* (ORD) from monocular depth literature $[\mathbf{G}, \mathbf{M}]$. Monocular depth methods do not produce metric depth but values in [0,1], and therefore evaluation is based on whether depth order of pixels match with groundtruth order. Ordinal error is defined as

$$ORD = \frac{\sum_{i} \omega_{i} \mathbb{1}(\ell_{i,\tau} \neq \ell_{i,\tau}^{*})}{\sum_{i} \omega_{i}}, \text{ where } \ell = \begin{cases} +1, & p_{0}/p_{1} \geq 1+\tau \\ -1, & p_{0}/p_{1} \leq \frac{1}{1+\tau} \\ 0, & \text{otherwise} \end{cases}$$
(4)

Computation is based on 50,000 randomly selected depth pixel pairs $[p_0,p_1]$ from groundtruth and monocular estimated depth maps. If their ratio is greater or smaller than the thresholds defined by the parameter τ ($\tau=0.03$ adopted from literature), then positive or negative indicator values ℓ (groundtruth) and ℓ^* (monocular estimated) are produced. ORD computes the proportion of pixel pairs that do not match. ORD 0.0 is the ideal value and 1.0 means that all

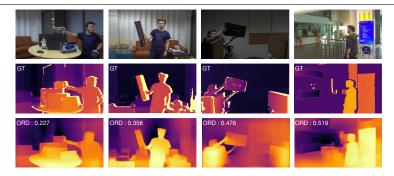


Figure 2: CDTB RGB frames (top), groundtruth depth maps (middle) and DenseDepth estimated depth maps (bottom) with their ordinal error (ORD) values.

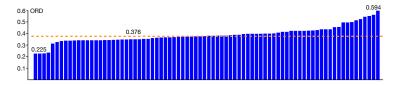


Figure 3: ORD histogram of 80 CDTB sequences. Lower ORD is better. Minimum, maximum and median values are marked. Mean ORD is 0.386.

pixels are in opposite order (for $\omega_i = 1$). We tested a number of monocular depth estimation methods and found the DenseDepth [\blacksquare] to produce the best results. Generated depth maps are shown in Figure 2 for various ORD values and the ORD value histogram of all CDTB sequences are shown in Figure 3. It should be noted that depth values are missing from large parts of the groundtruth images, but otherwise the median/average ORD values 0.376/0.386 indicate sufficient quality. Moreover, the generated RGBD sequences were manually verified and about 800 sequences were removed.

4.3 Cross-dataset evaluation

In the first experiment we investigated the contributions of depth-only training and depth-only fine-tuning of DiMP. To further emphasize the limitation that ResNet-50 is pretrained with RGB from ImageNet we denote it as "ResNet50-RGB". The RGB DiMP is the original DiMP from its authors. Depth variants of DiMP were trained with the generated depth data of LaSOT, GOT10k and COCO. For sanity check we report results also for DiMP trained with different color channels of the same datasets.

CDTB-ST. The results are summarized in Table 1 and provide the following findings: 1) off-the-shelf RGB DiMP is superior to single color channels (R/G/B) and depth as expected; 2) fine-tuning with the DepthTrack training set sequences systematically improves the results of all trackers and, in particular, depth tracking results are substantially improved (F-score jumps +20% from 0.421 to 0.508); 3) training DiMP from the scratch using channel specific training data further improves the results of all trackers - G channel obtains better accuracy than the original RGB (F-score increases from 0.665 to 0.701) and the depth-only F-score further improves +9% from 0.508 to 0.556.

Pr	Re	F	Backbone	Pre-trained	Fine-tuned	Input	Coding		
0.659	0.671	0.665	ResNet50-RGB	RGB		RGB	RGB		
0.681	0.689	0.685	ResNet50-RGB	RGB	DepthTrack-RGB	RGB	RGB		
Off-the-shelf RGB DiMP									
0.625	0.624	0.625	ResNet50-RGB	RGB		R	$3\times R$		
0.640	0.637	0.638	ResNet50-RGB	RGB		G	$3\times G$		
0.624	0.622	0.623	ResNet50-RGB	RGB		В	$3\times B$		
0.443	0.402	0.421	ResNet50-RGB	RGB		D	$3\times D$		
0.476	0.431	0.452	ResNet50-RGB	RGB		D	ColMap		
+	fine-tun	ing with	real depth data						
0.663	0.680	0.672	ResNet50-RGB	RGB	DepthTrack-G	G	$3\times G$		
0.510	0.506	0.508	ResNet50-RGB	RGB	DepthTrack-D	D	$3\times D$		
0.502	0.513	0.507	ResNet50-RGB	RGB	DepthTrack-D	D	ColMap		
0.444	0.410	0.426	ResNet50-RGB	RGB	DepthTrack-D [†]	D	$3\times D$		
0.455	0.412	0.433	ResNet50-RGB	RGB	DepthTrack-D [†]	D	ColMap		
Depth-	Depth-only DiMP trained w/ generated depth								
0.701	0.701	0.701	ResNet50-RGB	G	DepthTrack-G	G	$3\times G$		
0.698	0.703	0.700	ResNet50-RGB	G	DepthTrack-G	G	ColMap		
0.487	0.375	0.424	ResNet50-RGB	G	DepthTrack-D	D	$3\times D$		
0.549	0.563	0.556	ResNet50-RGB	\mathbf{D}^{\dagger}	D [†] DepthTrack-D		$3\times D$		
0.520	0.529	0.525	ResNet50-RGB	D^{\dagger}	DepthTrack-D	D	ColMap		

Table 1: Depth-only tracking results for the CDTB-ST (ST: short-term) dataset. The results are cross-dataset where CDTB was only used in testing. D^{\dagger} denotes DenseDepth generated depth data. The best depth-only (D) F-score is bolded and the second best underlined.

Pr	Re	F	Backbone	Pre-trained	Fine-tuned	Input	Coding			
0.534	0.483	0.507	ResNet50-RGB	RGB		RGB	RGB			
0.542	0.488	0.513	ResNet50-RGB	RGB	CDTB-RGB	RGB	RGB			
	Off-the-shelf RGB DiMP									
0.461	0.387	0.420	ResNet50-RGB	RGB		G	$3\times G$			
0.350	0.246	0.288	ResNet50-RGB	RGB		D	$3\times D$			
0.362	0.267	0.307	ResNet50-RGB	RGB		D	ColMap			
+	+ fine-tuning with real depth data									
0.463	0.413	0.437	ResNet50-RGB	RGB	CDTB-G	G	$3\times G$			
0.412	0.335	0.370	ResNet50-RGB	RGB	CDTB-D	D	$3\times D$			
0.381	0.350	0.364	ResNet50-RGB	RGB	CDTB-D	D	ColMap			
0.386	0.306	0.341	ResNet50-RGB	RGB	$CDTB-D^{\dagger}$	D	$3\times D$			
0.390	0.328	0.356	ResNet50-RGB	RGB	CDTB-D [†]	D	ColMap			
Depth-only DiMP trained w/ generated depth										
0.419	0.307	0.355	ResNet50-RGB	D^{\dagger}	CDTB-D	D	$3\times D$			
0.418	0.325	<u>0.366</u>	ResNet50-RGB	\mathbf{D}^{\dagger}	CDTB-D	D	ColMap			

Table 2: Cross-dataset results for the DepthTrack-ST test set.

These results verify that by training and fine-tuning DiMP with depth data - even with synthetic - improves the results substantially. The best color channel (Green) F-score improved from 0.638 to 0.701 (+10%) and the depth (D) from 0.421 to 0.556 (+32%). Overall, after these optimizations the Depth-DiMP is only 16% behind the original RGB DiMP.

DepthTrack-ST. To validate the above findings the datasets were swapped and DepthTrack-ST test set was used in testing and CDTB in fine-tuning. The results are summarized in Table 2. For DepthTrack the sequences are selected to particularly challenge RGB tracking and this is evident from the results. Fine-tuning again improves the results 28% (F-score from 0.288 to 0.370), but in this case generated data does not provide additional benefit (0.366) as

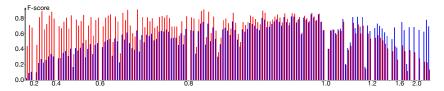


Figure 4: The plot of F-scores (Y-axis) and the ratios (X-axis) between the RGB DiMP tracker (red) and the Depth-DiMP tracker (D^{\dagger} -DepthTrack-3×D) (blue) for each sequence in CDTB-ST. Ratios < 1.0 represent RGB dominance (RGB performs better) and > 1.0 D dominance. 0.5 means that the RGB F-score is twice better than the corresponding D F-score and 2.0 means that the depth is twice better.

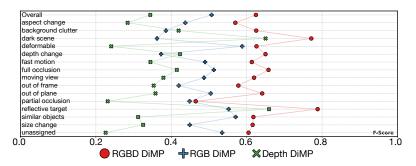


Figure 5: Visual attribute specific F-scores for RGB, D and RGBD (D^{\dagger} -CDTB-ColMap) DiMP variants on the DepthTrack-ST test set.

the sequences are very different from test data. However, it should be noted that if there is no fine-tuning data available the generated depth still improves the results.

4.4 Sequence level analysis

The results in Table 1 show that the best depth-only DiMP obtains F-score only 0.1 points behind the state-of-the-art RGB DiMP (0.556 vs. 0.665). Since the results of RGB DiMP do not substantially improve after fine-tuning (0.701 vs. 0.665) it is intriguing to find whether there are sequences particularly suitable or unsuitable for depth-only tracking.

To investigate the complementary properties of RGB and D we computed the graph in Figure 4 where all CDTB-ST sequences are sorted from left to right so that for the leftmost images RGB is more dominant and for the rightmost D is more dominant. In the middle are sequences for which color and depth perform comparably. In Figure 4 for most of the sequences RGB is better than D (ratio < 1.0), but, on the other hand, in 20% of the sequences (36 out of 152) the depth is better than RGB (> 1.0) and in half of them depth is clearly better.

Example sequences. The 10 best and worst sequences for the depth-only tracking are listed in Table 3 and Figure 6 shows two example sequences in which target shares similar texture and color with distractors (a person in these examples) and 3D rotates to change its 2D shape quickly. These changes finally lock the RGB tracker to wrong object. On the other hand, the distinct depth of the targets make their depth-only tracking easy.

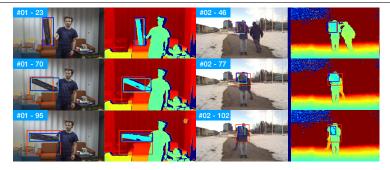


Figure 6: CDTB-ST example sequences where the Depth-DiMP tracker (D^{\dagger} -DepthTrack- $3\times D$) (blue boxes) is clearly superior to the RGB DiMP tracker (red boxes). Green boxes denote the groundtruth bounding boxes. #S-N denotes the N-th frame of S-th Seq. in Table 3.

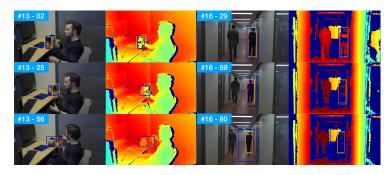


Figure 7: CDTB-ST example sequences where the RGB DiMP tracker (red boxes) is clearly superior to the Depth-DiMP tracker (D^{\dagger} -DepthTrack-3×D) (blue boxes).

Scenes for which depth-only fails are illustrated in Figure 7. Two reasons can be observed: the target and distractor share the same depth, but have different texture, or the target is too far and the depth sensor fails to get measurements.

Visual attributes. To provide more insights about scenes in which depth is an important cue, we computed attribute specific F-scores for the DepthTrack-ST (Figure 5). The scores verify that there are scene types where depth outperforms RGB: *background clutter*, *dark*, *depth change* and *reflective* scenes. RGB is better than D for the other scene types (attributes), but together RGB and D are clearly superior in all of them.

4.5 RGBD tracking

We combined the optimized Depth-DiMP and RGB DiMP to RGBD-DiMP and compared its performance to the VOT2020-RGBD winners with the novel DepthTrack-ST test set. The results in Table 4 show that the 2020 second, DDiMP, now wins the last year winner, ATCAIS, but our end-to-end trained RGBD-DiMP obtains the best results with 16% higher F-score than DDiMP. It should be noted that the short-term setting in these experiments is different from the long-term setting used in the VOT-RGBD challenge.

		D DiMP			RGB DiMP			
id	Seq. name	Pr	Re	F	Ratio	Pr	Re	F
01	box_room_noocc_3_1	0.689	0.689	0.689	← 5.1	0.135	0.135	0.135
02	XMG_outside_2	0.771	0.771	0.771	$\leftarrow 3.4$	0.229	0.229	0.229
03	boxes_office_occ_1_2	0.621	0.621	0.621	$\leftarrow 2.6$	0.235	0.235	0.235
04	box_darkroom_noocc_9_1	0.643	0.643	0.643	$\leftarrow 2.3$	0.278	0.278	0.278
05	boxes_room_occ_1_1	0.669	0.693	0.681	$\leftarrow 1.9$	0.354	0.360	0.357
06	box_darkroom_noocc_5_1	0.677	0.679	0.678	$\leftarrow 1.8$	0.385	0.387	0.386
07	box_room_occ_1_1	0.595	0.688	0.638	← 1.7	0.346	0.403	0.372
08	bag_outside_3	0.741	0.741	0.741	← 1.7	0.444	0.444	0.444
09	box_room_noocc_2_1	0.417	0.417	0.417	$\leftarrow 1.6$	0.262	0.262	0.262
10	toy_office_occ_1_2	0.642	0.708	0.673	$\leftarrow 1.5$	0.541	0.367	0.437
11	trophy_room_occ_1_2	0.083	0.103	0.092	7.7→	0.653	0.770	0.706
12	two_mugs_5	0.108	0.106	0.107	$6.3 \rightarrow$	0.707	0.640	0.672
13	thermos_office_occ_1_2	0.091	0.112	0.100	$4.5 \rightarrow$	0.416	0.494	0.452
14	box_room_noocc_8_1	0.220	0.220	0.220	$3.7 \rightarrow$	0.804	0.804	0.804
15	humans_corridor_occ_2_B_4	0.246	0.298	0.270	$3.3 \rightarrow$	0.882	0.882	0.882
16	humans_corridor_occ_2_B_1	0.160	0.372	0.224	$2.9 \rightarrow$	0.538	0.825	0.651
17	jug_3	0.259	0.259	0.259	$2.8 \rightarrow$	0.718	0.718	0.718
18	bottle_box_2	0.232	0.434	0.303	$2.7 \rightarrow$	0.846	0.808	0.827
19	trashcan_room_occ_1_6	0.333	0.342	0.337	$2.6 \rightarrow$	0.882	0.868	0.875
20	bottle_room_occ_1_2	0.558	0.209	0.304	$2.6 \rightarrow$	0.787	0.787	0.787

Table 3: 10 best CDTB-ST sequences for D and RGB DiMP trackers respectively. Ratios are converted to indicate how many times better the other modality is as denoted by the arrow.

Tracker	Pr	Re	F	Input
ATCAIS	0.566	0.444	0.498	RGB+D
DDiMP	0.607	0.478	0.535	RGB+D
CLGS_D	0.710	0.367	0.484	RGB+D
Siam_LTD	0.433	0.389	0.410	RGB+D
DiMP-RGBD	0.639	0.611	0.625	RGB+ColMap
DiMP-RGBD	0.610	0.567	0.588	RGB+3×D

Table 4: Comparison of the four best VOT2020-RGBD trackers [122] and the RGBD data trained RGBD-DiMP with the DepthTrack-ST test set.

5 Conclusion

We experimentally verified that there are tracking scenes for which depth provides dominant cue and where depth-only trackers are superior to state-of-the-art RGB trackers. Moreover, we showed that the end-to-end trained RGBD-DiMP using generated and real depth and RGB data is clearly superior to state-of-the-art RGBD trackers and does not need additional heuristics. Our data and code will be published to facilitate fair comparisons and more work on depth-only object tracking.

Acknowledgements

We thank VOT guru Matej Kristan for the original idea of using generated depth in training.

References

- [1] Ibraheem Alhashim and Peter Wonka. High quality monocular depth estimation via transfer learning. *arXiv preprint arXiv:1812.11941*, 2018.
- [2] Goutam Bhat, Martin Danelljan, Luc Van Gool, and Radu Timofte. Learning discriminative model prediction for tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6182–6191, 2019.
- [3] Weifeng Chen, Zhao Fu, Dawei Yang, and Jia Deng. Single-image depth perception in the wild. *Advances in neural information processing systems*, 29:730–738, 2016.
- [4] Martin Danelljan, Goutam Bhat, Fahad Shahbaz Khan, and Michael Felsberg. Atom: Accurate tracking by overlap maximization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4660–4669, 2019.
- [5] Martin Danelljan, Luc Van Gool, and Radu Timofte. Probabilistic regression for visual tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7183–7192, 2020.
- [6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.
- [7] Heng Fan, Liting Lin, Fan Yang, Peng Chu, Ge Deng, Sijia Yu, Hexin Bai, Yong Xu, Chunyuan Liao, and Haibin Ling. Lasot: A high-quality benchmark for large-scale single object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5374–5383, 2019.
- [8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [9] João F Henriques, Rui Caseiro, Pedro Martins, and Jorge Batista. High-speed tracking with kernelized correlation filters. *IEEE transactions on pattern analysis and machine* intelligence, 37(3):583–596, 2014.
- [10] Lianghua Huang, Xin Zhao, and Kaiqi Huang. Got-10k: A large high-diversity benchmark for generic object tracking in the wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.
- [11] 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 2206–2241, 2019.
- [12] 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, 2020.

- [13] Matej Kristan, Jiri Matas, Ales Leonardis, Michael Felsberg, Roman Pflugfelder, Joni-Kristian Kamarainen, Hyung Jin Chang, Martin Danelljan, Luka Cehovin, Alan Lukezic, Ondrej Drbohlav, Jani Kapyla, Gustav Hager, Song Yan, Jinyu Yang, Zhongqun Zhang, and Gustavo Fernandez. The ninth visual object tracking vot2021 challenge results. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 2711–2738, October 2021.
- [14] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.
- [15] 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.
- [16] Matthias Muller, Adel Bibi, Silvio Giancola, Salman Alsubaihi, and Bernard Ghanem. Trackingnet: A large-scale dataset and benchmark for object tracking in the wild. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 300–317, 2018.
- [17] Hyeonseob Nam and Bohyung Han. Learning multi-domain convolutional neural networks for visual tracking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4293–4302, 2016.
- [18] Jorge Nocedal and Stephen Wright. *Numerical optimization*. Springer Science & Business Media, 2006.
- [19] Shuran Song and Jianxiong Xiao. Tracking revisited using rgbd camera: Unified benchmark and baselines. In *Proceedings of the IEEE international conference on computer vision*, pages 233–240, 2013.
- [20] Ke Xian, Jianming Zhang, Oliver Wang, Long Mai, Zhe Lin, and Zhiguo Cao. Structure-guided ranking loss for single image depth prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 611–620, 2020.
- [21] Jingjing Xiao, Rustam Stolkin, Yuqing Gao, and Aleš 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*, 48(8): 2485–2499, 2017.
- [22] Song Yan, Jinyu Yang, Jani Käpylä, Feng Zheng, Aleš Leonardis, and Joni-Kristian Kämäräinen. Depthtrack: Unveiling the power of rgbd tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10725–10733, 2021.