RGBD1K: A Large-Scale Dataset and Benchmark for RGB-D Object Tracking

Xue-Feng Zhu[1], **Tianyang Xu**[1], **Zhangyong Tang**[1], **Zucheng Wu**[1], **Haodong Liu**[1], **Xiao Yang**[1], **Xiao-Jun Wu**[1*], **Josef Kittler**[2]

[1]School of Artificial Intelligence and Computer Science, Jiangnan University, Wuxi, Jiangsu, P.R. China
[2]Centre for Vision, Speech and Signal Processing, University of Surrey, Guildford, GU2 7XH, UK
xuefeng_zhu95@163.com; {tianyang.xu; wu_xiaojun}@jiangnan.edu.cn; j.kittler@surrey.ac.uk

Abstract

RGB-D object tracking has attracted considerable attention recently, achieving promising performance thanks to the symbiosis between visual and depth channels. However, given a limited amount of annotated RGB-D tracking data, most state-of-the-art RGB-D trackers are simple extensions of high-performance RGB-only trackers, without fully exploiting the underlying potential of the depth channel in the offline training stage. To address the dataset deficiency issue, a new RGB-D dataset named RGBD1K is released in this paper. The RGBD1K contains 1,050 sequences with about 2.5M frames in total. To demonstrate the benefits of training on a larger RGB-D data set in general, and RGBD1K in particular, we develop a transformer-based RGB-D tracker, named SPT, as a baseline for future visual object tracking studies using the new dataset. The results, of extensive experiments using the SPT tracker demonstrate the potential of the RGBD1K dataset to improve the performance of RGB-D tracking, inspiring future developments of effective tracker designs. The dataset and codes will be available on the project homepage: https://github.com/xuefeng-zhu5/RGBD1K.

Introduction

Visual Object Tracking (VOT) aims at detecting the position and scale of an object of interest in every frame of a video. The tracking capability plays a significant role in the gamut of perceptual functionalities in computer vision and pattern recognition (Xue et al. 2020; Griffiths et al. 2017; Smeulders et al. 2013). The development of visual object tracking techniques has been ongoing for decades. In recent years in particular, with the access to large-scale annotated datasets, such as GOT10K (Huang, Zhao, and Huang 2019), TrackingNet (Muller et al. 2018), LaSOT (Fan et al. 2019), etc., the development of advanced visual object trackers has been accelerated by deep learning. Trained offline, using millions of labelled video frames, tracking networks are capable to learn robust feature representations, resulting in remarkable performance improvements, compared with conventional online learning methods (Kristan et al. 2019, 2020, 2021).

Recently, with the widespread availability of low-cost RGB-D sensors, the task of visual object tracking has broadened to include RGB-D videos. An RGB-D data is comprised of a three-channel RGB image and a single-channel depth map. Compared to conventional RGB-only tracking, the additional depth maps of RGB-D videos provide supplementary spatial information that facilitates object tracking in complicated scenarios (Bagautdinov, Fleuret, and Fua 2015; Meshgi et al. 2016). Nevertheless, the existing RGB-D tracking methods generally build upon high-performance RGB-only trackers, adopting the depth information in the online tracking stage to support reasoning about partially occluded targets, and re-detection of disappearing targets (Camplani et al. 2015; Kart, Kamarainen, and Matas 2018; Hannuna et al. 2019).

However, RGB-D trackers are not evolving as swiftly as RGB-only trackers (Kristan et al. 2019, 2020, 2021). The main reason is the lack of training data for RGB-D tracking. The publicly available annotated RGB-D videos cannot support offline training of an RGB-D tracking network. More specifically, while the existing datasets for RGB-only trackers contain thousands of video sequences with millions of annotated frames, the existing RGB-D datasets contain only 416 video sequences in total.

Recently, a new RGB-D tracking dataset named DepthTrack as well as an offline trained RGB-D tracker DeT have been made public (Yan et al. 2021b). However, the training set of DepthTrack contains only 150 videos captured in realistic scenarios. The vast majority of the RGB-D training data used for the development of tracker DeT was generated from RGB-only tracking datasets using monocular depth estimation techniques. The real training data collected by the depth camera occupies only a very small proportion. The performance of a deep RGB-D tracker trained in this way depends largely on the quality of the monocular depth estimation used for reconstructing the depth information. In summary, the existing RGB-D data is far from sufficient to promote the rapid development of RGB-D tracking.

In order to further motivate the investigation of RGB-D tracking, and its use, we collect a new RGB-D dataset named RGBD1K. RGBD1K contains 1,050 sequences with about 2.5M frames in total. Of these, 1,000 videos are reserved for training and 50 videos for testing. For the training videos, considering the annotation cost as well as the fact that the top one-fifth of frames of a long-term video contains representative visual and depth appearance variations

*Xiao-Jun Wu is the corresponding author.

Dataset	Videos number	Total frames	Average length	Annotated frames	Scene attributes	ST/ LT
PTB	100	21,542	215	21,542	5	ST
STC	36	9,009	250	9,009	12	ST
CDTB	80	101,956	1,274	101,956	13	LT
DepthTrack	200	294,591	1,473	294,591	15	LT
RGBD1K	1,050	2,503,400	2,384	717,900	15	LT

Table 1: An overview of the existing RGB-D datasets. The ST/ LT means Short-Term/ Long-Term.

for the learning of a tracking model, only the first 600 frames of each video are annotated. Therefore, 600,000 annotated frames of RGBD1K can be utilized for supervised learning of deep RGB-D tracking methods. Regarding the test set, all the frames are annotated, containing 117,900 frames in total. Additionally, we annotate each frame with 15 challenging attributes. The per-frame attributes facilitate the analysis of the trackers. To the best of our knowledge, RGBD1K is the largest dataset for RGB-D tracking presently in existence. The Table 1 summarises the existing RGB-D datasets, including PTB (Song and Xiao 2013), STC (Xiao et al. 2017), CDTB (Lukezic et al. 2019), DepthTrack (Yan et al. 2021b) and our RGBD1K. As evident from the table, the proposed RGBD1K has the largest number of videos, frames, annotations, and the average length of sequences.

In order to demonstrate the impact of the new dataset on the accuracy of RGB-D tracking, we propose a new baseline tracker based on spatial transformer learning, named SPT. Specifically, we extend the RGB-only tracking network (Yan et al. 2021a) to an RGB-D version and introduce a novel fusion module designed to fuse the features from the two modalities. The SPT is trained offline using the 1,000 training videos of the RGBD1K dataset. Extensive experiments, including the ablation experiments for self-analysis, and the comparative evaluation, are conducted on the RGBD1K, DepthTrack and CDTB datasets. The corresponding results demonstrate the effectiveness of our RGBD1K dataset and the competitiveness of our new baseline tracker SPT.

Related Work

Recently, the advancement of RGB-D tracking has been stimulated by the emergence of RGB-D tracking datasets. In this section, we briefly introduce the techniques that are closely related to this tracking task.

RGB-D Tracking Datasets

There are four RGB-D object tracking datasets publicly available, including Princeton Tracking Benchmark (PTB) (Song and Xiao 2013), Spatio-Temporal Consistency dataset (STC) (Xiao et al. 2017), Color and Depth Tracking Benchmark (CDTB) (Lukezic et al. 2019) and DepthTrack (Yan et al. 2021b). The specific properties of these four datasets are provided in Table 1. PTB (Song and Xiao 2013) is the seminal publicly available RGB-D tracking dataset. PTB contains 100 challenging videos recorded indoors for RGB-D tracking evaluation. According to the target category, target size, movement, occlusion and motion type, these video sequences are labelled according to 11 attribute categories. STC (Xiao et al. 2017) is an RGB-D tracking dataset comprising 36 video sequences with 12 per-frame annotated attributes. STC contains both indoor and outdoor scenarios. CDTB (Lukezic et al. 2019) is the existing largest test dataset for RGB-D tracking, comprising 80 video sequences captured in long-term tracking scenarios. All these sequences are annotated with 13 per-frame attributes. CDTB dataset has been recently adopted in the VOT-RGBD 2019 (Kristan et al. 2019), 2020 (Kristan et al. 2020) and 2021 (Kristan et al. 2021) challenges. DepthTrack (Yan et al. 2021b) is the most recent RGB-D tracking dataset and is also the first dataset for offline training the RGB-D trackers. It contains 200 RGB-D video sequences captured both indoors and outdoors, in which 150 videos can be adopted for offline learning. The remaining 50 videos are used for the tracking performance evaluation.

RGB-D Tracking Methods

Since the advent of the seminal publicly available RGB-D tracking dataset PTB, the research on RGB-D tracking has received widespread attention. Given the remarkable performance of the existing RGB-only tracking framework, most of the RGB-D tracking algorithms in the literature are simple RGB-only tracker extensions, utilizing the supplementary depth information effectively to improve the performance.

Based on the early colour-only Struck tracker and classical mean-shift tracker respectively, a local depth pattern feature (Awwad, Hussein, and Piccardi 2015) and a 3-D mean-shift (Liu et al. 2018) are proposed for RGB-D tracking. Besides, a 3D part-based sparse tracker with occlusion handling is developed from the particle filter framework (Bibi, Zhang, and Ghanem 2016). Considering the promising performance of Discriminative Correlation Filter (DCF) based trackers on RGB videos (Danelljan et al. 2017; Xu et al. 2019; Zhu et al. 2021), the RGB-D trackers DS-KCF (Camplani et al. 2015), DS-KCF-shape (Hannuna et al. 2019), DM-DCF (Kart, Kamarainen, and Matas 2018) and OTR (Kart et al. 2019) are developed. Recently, deep learning has been shown to exhibit promising performance in RGB-only tracking (Bhat et al. 2019; Lukezic, Matas, and Kristan 2020; Zhao et al. 2021b, 2022). The existing best performing RGB-D trackers are extensions of offline trained RGB-only trackers. For example, in VOT-RGBD challenges (Kristan et al. 2019, 2020, 2021), the trackers STARK_RGBD, TALGD, ATCAIS are developed from the deep RGB trackers STARK (Yan et al. 2021a), ATOM (Danelljan et al. 2019) and DiMP (Bhat et al. 2019). More recently, Yan *et al.* propose an end-to-end offline trained RGB-D tracker DeT (Yan et al. 2021b), which is

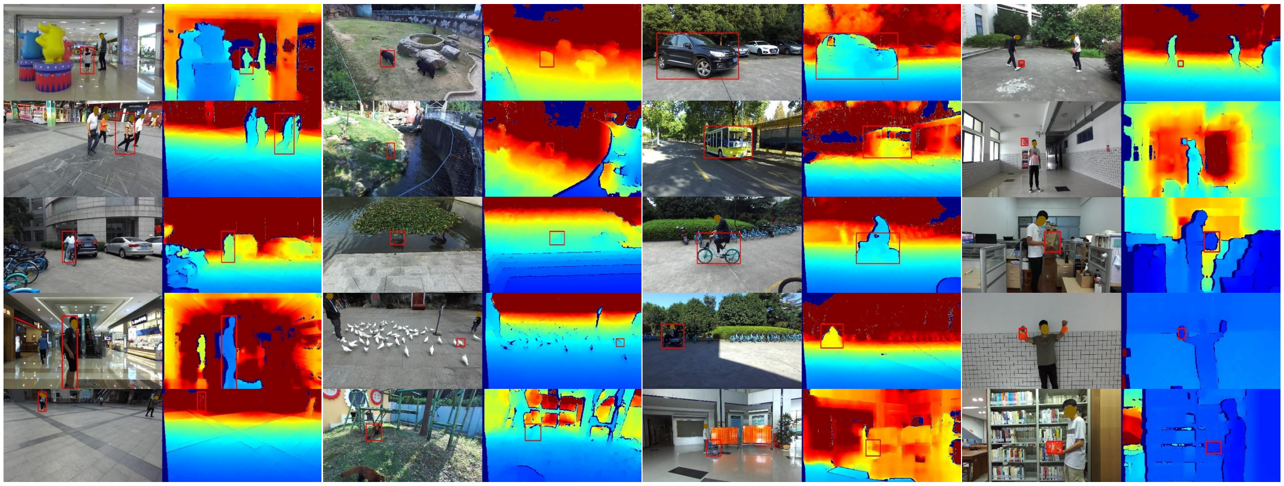

Figure 1: RGB-D image samples from the RGBD1K. The targets are marked with red boxes, and the depth maps are converted to colour maps for more clear visualization. The first column is five samples of human, including *child*, *elder*, *woman*, *man* and *couple*. The second column is five samples of animal, such as *bear*, *tiger*, *swan*, *pigeon* and *red panda*. The third column is samples of vehicle, like *car*, *bus*, *bicycle*, *motorbike* and *cart*. The final column is five samples of articles for daily use, including *basketball*, *balloon*, *box*, *doll* and *book*.

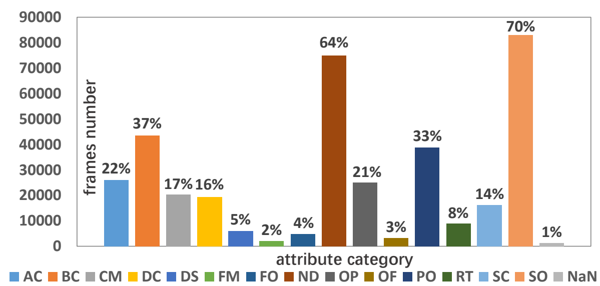

Figure 2: The distribution of frames of different attributes in the RGBD1K test set.

based on the framework of the RGB-only trackers, ATOM and DiMP.

The RGBD1K Dataset

Video Sequences

RGBD1K contains 1,000 training sequences and 50 test sequences. In total, the training set contains 2,385,500 frames and the test set contains 117,900 frames. All the 1,050 sequences of our RGBD1K dataset are captured indoors and outdoors using the stereo camera ZED. The ZED camera provides time-synchronized and pixel-aligned RGB and depth frames. The video sequences share the same frame rate of 25 frames per second (fps). The RGB images are stored using 24-bit (8-bit each channel) JEPG format, meanwhile, the depth maps are stored using 16-bit PNG format.

The RGBD1K covers a considerable number of object categories, including more than 100 different types concerned with humans, animals, vehicles and articles for daily use. Fig. 1 provides some cases of different object classes. We also select dozens of different scenes to record these sequences, such as office buildings, shopping malls, zoos, sports fields, etc. Besides, some video sequences are captured from a first-person perspective and an overlooking perspective to simulate the perspectives of moving robots, UAVs and surveillance cameras. For more statistical analysis and examples of different scenarios and object classes in the proposed RGBD1K dataset, please refer to the supplementary material.

Data Annotation

As to each video, we annotate the frames with the target bounding box. It is universally acknowledged that data annotation is critical to research but time-consuming. Considering that a short clip of a video sequence can contain sufficient visual and depth appearance variations, as well as to reduce the time cost, for the training set, we only annotate the frames of one segment of each video. Specifically, we only annotate the first 600 frames of each sequence for the training set. Although on average each video is only annotated with 1/4 of its length, we argue that the appearance variations in the annotated clips are sufficient for the learning of the spatio-temporal changing targets and scenarios (Valmadre et al. 2018; Kristan et al. 2018). Besides, the unlabeled part is tightly related to the labelled part. Such partially annotated videos can be directly adopted for supervised learning, with the potential also to be utilized effectively for semi-supervised learning. Meanwhile, it saves labour costs. For the test set, all the frames of each sequence are annotated.

For further performance analysis of tracking methods, we annotate each frame of the test set with 15 attributes as proposed by CDTB (Lukezic et al. 2019) and DepthTrack (Yan et al. 2021b), including Aspect-ratio Change (AC), Background Clutter (BC), Camera Motion (CM), Depth Change (DC), Dark Scene (DS), Fast Motion (FM), Full Occlusion

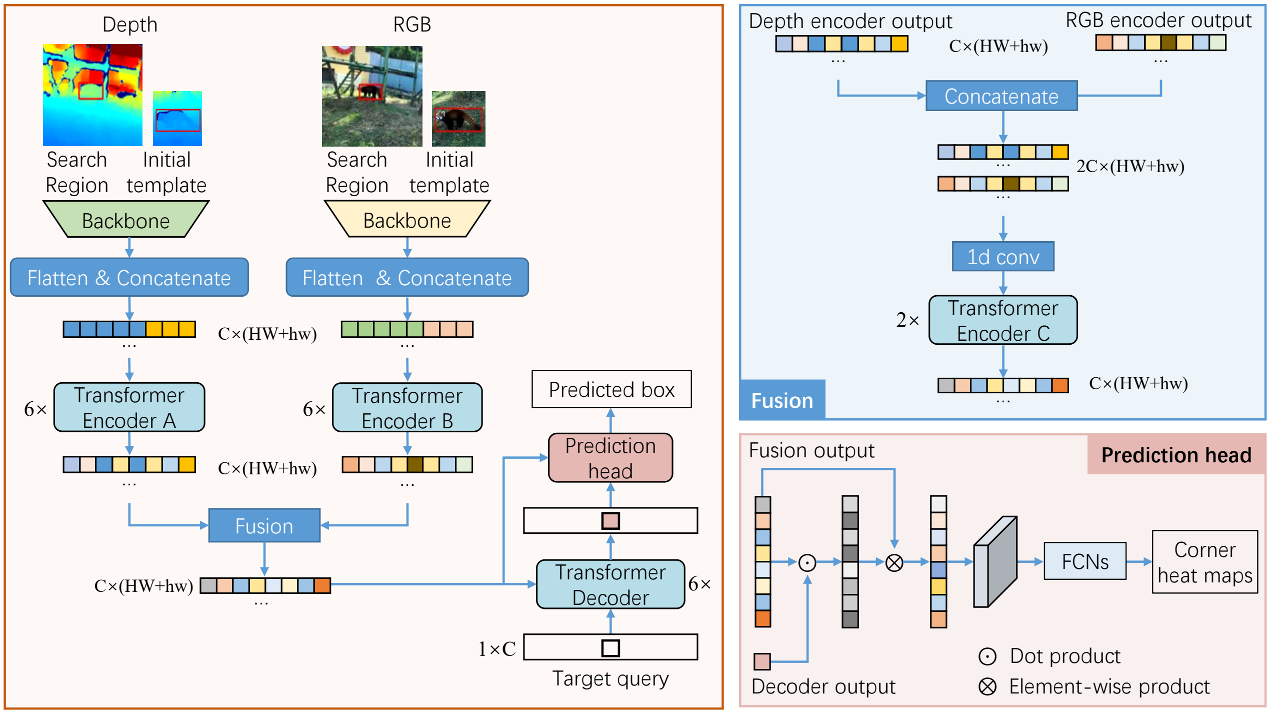

Figure 3: Illustration of the framework of the proposed SPT tracker. The transformer encoder A and the transformer encoder B have the same structure, which stacks 6 encoder layers. The transformer encoder C stacks 2 encoder layers.

(FO), Non-rigid Deformation (ND), Out-of-plane Rotation (OP), Out of Frame (OF), Partial Occlusion (PO), Reflective Target (RT), Size Change (SC), Similar Objects (SO) and Unassigned (NaN). The attributes AC, DC, FM, SC and NaN are calculated from the RGB-D images and the bounding box annotations. The remaining 10 attributes are annotated manually. These scene attributes are beneficial for the trackers to analyse their merits and demerits in specific challenges. For a detailed definition of each attribute, please refer to the supplementary material.

The distribution of frames in each attribute category of the RGBD1K test set is reported in Fig. 2. From the figure, we can observe that only 1% of the frames are marked without any scene attributes, which indicates that the RGBD1K test set is challenging. Among the sequences, approximate 64% of the frames are of non-rigid deformable targets. Typically, a deformable object implies a high probability of drastic appearance variations, which means it is more difficult for stable tracking. In addition, 70% frames are marked with the challenge attribute of similar objects. The interference of similar objects in the background is an important issue worth studying for robust tracking. Besides, background clutter and partial occlusion are also essential challenging factors in the RGBD1K test set. Although some attributes contain a small number of frames, such as FO and OF only occupy 4% and 3% respectively, they are still very valuable for practical applications. An RGB-D frame with the attribute of FO or OF means the target is invisible in the current frame. Despite that the targets in only 7% of the frames are invisible in total, this means that on average each video of the test set has about 165 frames of the target disappearance. The frequent long period of target disappearance and reappearance complicates the tracking analysis, requiring perceptual capability for the RGBD tracker.

Performance Measures

While there are no explicit restrictions on the use of RGBD1K, when evaluating trackers on the test set we advocate the use of the long-term tracking evaluation protocol from (Lukežič et al. 2018), which is applied in the VOT-RGBD challenges (Kristan et al. 2019, 2020, 2021). The reason is that there are a certain proportion of frames in which the targets are invisible in the RGBD1K dataset, *i.e.* the target may disappear and reappear several times in one video. For a tracker to be evaluated on RGBD1K, the ability to localise the target as well as to predict the target absence, and recapture the re-emerged target, is of significance for a robust tracking system. Therefore, the long-term VOT evaluation protocol is precisely suited for evaluating trackers on our dataset.

The tracking Precision and Recall from (Lukežič et al. 2018) are applied as the performance measures. Specifically, Precision is defined as the average overlap ratio of the predicted and ground truth targets on the frames where the target is detected. The Recall represents the average overlap ratio of the predicted target bounding box and the ground truth annotation, measured on the frames where the target is visible. The primary performance measure is F-score obtained by calculating the tracking F-measure that combines tracking Precision and Recall. Besides, trackers can

Dataset	RGBD1K			DepthTrack			CDTB		
Method	Pr	Re	F-score	Pr	Re	F-score	Pr	Re	F-score
STARK-S	0.480	0.510	0.495	0.490	0.511	0.500	0.630	0.701	0.664
STARK-S-FT	0.509	0.537	0.522	0.497	0.517	0.507	0.638	0.706	0.670
SPT	0.545	0.578	0.561	0.527	0.549	0.538	0.654	0.726	0.688

Table 2: A comparison of the STARK-S, STARK-S-FT and SPT on the RGBD1K, DepthTrack and CDTB datasets.

Method	DDiMP	ATCAIS	DRefine	SLMD	DAL	DeT	TSDM	TALGD	Siam_LTD	SPT
Pr	**0.557**	0.511	0.532	0.554	0.562	0.438	0.455	0.485	0.543	0.545
Re	0.534	0.451	0.462	0.526	0.407	0.419	0.361	0.415	0.318	**0.578**
F-score	0.545	0.479	0.494	0.540	0.472	0.428	0.403	0.447	0.398	**0.561**
ST/LT	ST	LT	LT	LT	LT	ST	LT	LT	LT	ST

Table 3: The tracking results on the RGBD1K test set.

be conveniently evaluated on the RGBD1K by using the VOT challenge toolkit (Kristan et al. 2021). For more details on the evaluation metrics, readers can refer to the literature (Lukežič et al. 2018) or our supplementary material.

A New Baseline RGB-D Tracker

To demonstrate the significance of the RGBD1K dataset as well as to inspire new designs for RGB-D tracking, we propose a new RGB-D tracking baseline coined as SPT. The SPT is developed from the recent state-of-the-art transformer-based tracker STARK (Yan et al. 2021a). STARK is a distinguished RGB-only tracker, achieving remarkable performance on RGB-only tracking datasets.

The SPT is formed by extending the STARK-S (STARK without the temporal structure) to an RGB-D version with a dedicated feature fusion module. The architecture of SPT is presented in Fig. 3. Firstly, the search regions and the initial templates of the two modalities are input to the backbone to extract deep CNN features respectively. The backbone used here is the ResNet-50 network (He et al. 2016). The features of search regions and templates are of $H \times W \times C$ and $h \times w \times C$, respectively. Then, the features of each modality are flattened and concatenated, following a 6-layer stacked transformer encoder to fuse the template-search appearance for the specific modality. Finally, the outputs of two modality-specific encoders are fused by our feature fusion module.

Here we introduce the proposed feature fusion module in detail. Firstly, the depth encoder output and the RGB encoder output are concatenated across channels. Then a $1d$ convolutional layer is adopted to reduce the channel number of the concatenated features from $2C$ to C. Finally, we introduce a transformer encoder stacking 2 encoder layers to further fuse and enhance the features of the two modalities. Each encoder layer is composed of a multi-head self-attention module and a feed-forward network.

The rest parts of the framework include the target query, the transformer decoder and the target bounding box prediction head (Yan et al. 2021a). The transformer decoder, stacking 6 decoder layers, takes a learnable target query and the fused features as input. Each decoder layer contains a self-attention, encoder-decoder attention, and a feed-forward network. Later, the output of the transformer decoder and the fused features are fed into the bounding box prediction head to predict the target box coordinates.

In the bounding box prediction module, firstly, the decoder output is used to calculate the similarities with fused features, and the similarities are used to enhance the fused features. Then the enhanced features are reshaped and passed through fully-convolutional networks to generate a top-left corner heat map and a bottom-right corner heat map. With the top-left and bottom-right corner points, the object bounding box can be determined. The loss function of SPT is the combination between l_1 loss and the IoU loss. For more details on each component of SPT, please refer to the supplementary material.

Evaluation

We perform extensive experiments on RGBD1K, DepthTrack and CDTB datasets. In this section, we describe the implementation details of our tracker SPT, including the parameters setup and the experimental platform. Then, the results of ablation studies are presented, to demonstrate the effectiveness of our dataset as well as the proposed feature fusion module of the SPT tracker. Finally, we provide the results and corresponding analysis of comparative experiments.

Implementation Details

The proposed SPT tracker is trained and evaluated with an Intel i9-CPU and one NVIDIA GeForce RTX 3090 GPU. The training and test parameters are set the same as Stark, except for the learning rate and training epoch number. The learning rate is set as 10^{-5} and the total epoch number is 250. As to the backbone, transformer encoder A, B, transformer decoder and box prediction head of SPT, we initialize their weights by using the weights of corresponding components of the officially published STARK-S model. Then the SPT is trained on the training set of RGBD1K. The tracking speed of SPT is about 25 fps.

Ablation Study

In order to demonstrate the effectiveness of the proposed RGBD1K dataset for RGB-D tracking, firstly, we

Method	DDiMP	ATCAIS	CLGS_D	SiamDW_D	LTDSEd	Siam_LTD	SiamM_Ds	DAL	DeT	SPT
Pr	0.503	0.500	0.584	0.429	0.430	0.418	0.463	0.512	**0.560**	0.527
Re	0.469	0.455	0.369	0.436	0.382	0.342	0.264	0.369	0.506	**0.549**
F-score	0.485	0.476	0.453	0.432	0.405	0.376	0.336	0.429	0.532	**0.538**
ST/LT	ST	LT	LT	LT	LT	LT	LT	LT	ST	ST

Table 4: The tracking results on the DepthTrack dataset.

Method	DDiMP	ATCAIS	CLGS_D	SiamDW_D	LTDSEd	Siam_LTD	SiamM_Ds	OTR	DeT	SPT
Pr	0.703	**0.709**	0.725	0.677	0.674	0.626	0.685	0.364	0.674	0.654
Re	0.689	0.696	0.664	0.685	0.643	0.489	0.677	0.312	0.642	**0.726**
F-score	0.696	**0.702**	0.693	0.681	0.658	0.549	0.681	0.336	0.657	0.688
Speed (fps)	4.7	1.3	7.3	3.8	5.7	13.0	19.4	1.8	36.8	25.3
ST/LT	ST	LT	LT	LT	LT	LT	LT	LT	ST	ST

Table 5: The tracking results on the CDTB dataset.

construct three trackers, including STARK-S (Yan et al. 2021a), STARK-S-FT, and our SPT. STARK-S, the STARK tracker with ResNet-50 as the backbone (without a temporal branch), is the baseline tracker of SPT. We use the officially released trained model for STARK-S. The STARK-S-FT is the STARK-S tracker fine-tuned on the RGBD1K using only all the RGB images of the training set. The SPT is trained with the RGB-D images of RGBD1K. The results of the test set of RGBD1K are provided in Table 2. Fine-tuned with the RGB images of the training set of RGBD1K, the STARK-S-FT improves the performance from 0.480, 0.510, and 0.495 to 0.509, 0.537 and 0.522 in terms of Precision, Recall and F-score, respectively. Trained with RGB-D images of the training set of RGBD1K, SPT further improves the results to 0.545 for Precision, 0.578 for Recall and 0.561 for F-score. This improvement enables us to draw the conclusion that the challenging RGB images, as well as the depth images of RGBD1K, are beneficial for improving RGB-D tracking performance.

To further confirm the merit of RGBD1K, we conduct the same experiments on two other datasets DepthTrack and CDTB, to explore the performance among STARK-S, STARK-S-FT and SPT. It is worth noting that the trackers are trained only using the RGBD1K without sequences from DepthTrack or CDTB to fine-tune the tracking networks or corresponding hyper-parameters. The results on DepthTrack and CDTB are exhibited in Table 2. After training on the RGBD1K dataset, the trackers STARK-S-FT and SPT achieve significant performance improvement on the DepthTrack and CDTB datasets. Especially, the SPT trained with RGB-D data from the RGBD1K, improves the results of STARK-S from 0.490, 0.511 and 0.500 to 0.527, 0.549 and 0.538 and from 0.630, 0.701 and 0.664 to 0.654, 0.726 and 0.688 in terms of Precision, Recall and F-score on the DepthTrack and CDTB datasets, respectively. Concerning the F-score measure, the SPT improves the STARK-S by *7.6%* and *3.6%* on DepthTrack and CDTB, respectively. Apparently, the results shown in Table 2 demonstrate the generalised advantages of the proposed RGBD1K dataset in training end-to-end RGB-D trackers. Furthermore, extensive ablation experiments and analyses about the fusion module are provided in the supplementary material.

Comparison with SOTA Methods

To demonstrate the superiority of the tracker SPT, we conduct quantitative, qualitative and attribute-based experiments with state-of-the-art trackers. The attribute-based experiments can be found in the supplementary material.

Quantitative Comparison: We compare the proposed SPT with a considerable number of recent state-of-the-art RGB-D trackers on the RGBD1K test set. In Table. 3, we report the results of RGB-D trackers, including DDiMP, ATCAIS and Siam_LTD submitted to the VOT-RGBD 2020 challenge (Kristan et al. 2020), TALGD, DRefine and SLMD from the VOT-RGBD 2021 challenge (Kristan et al. 2021), DAL (Qian et al. 2021), DeT (Yan et al. 2021b), TSDM (Zhao et al. 2021a) and SPT. Detailed results of some other RGB-only trackers on the RGBD1K test set are provided in the supplementary material. Generally, the F-score is the most important performance measure in the VOT protocol and the trackers are ranked according to F-score values. As can be seen, on the RGBD1K test set, the SPT achieves the best F-score, and the short-term RGB-D tracker DDiMP is the second-best tracker. Compared to the tracker DDiMP, our SPT tracker obtains *2.9%* improvement in terms of F-score. Besides, compared with the long-term RGB-D trackers, such as ATCAIS, DRefine, SLMD and DAL, the proposed SPT tracker is also predominant, with gains of *17.1%*, *13.5%*, *3.9%*, and *18.8%* on F-score, respectively. The tracking performance gain of the SPT indicates that training with the proposed RGBD1K facilitates more robust RGB-D tracking.

To further reflect the transferability and domain advantage of our RGBD1K, we compare the SPT tracker with state-of-the-art trackers on the DepthTrack and CDTB datasets. It is still worth noting that our SPT tracker trained using RGBD1K is directly used to test on the DepthTrack and CDTB datasets without fine-tuning any parameters. The results on DepthTrack and CDTB are provided in Table. 4 and 5, respectively.

In Table. 4, the SPT tracker is compared with DDiMP, ATCAIS, CLGS_D and Siam_LTD from the VOT-RGBD 2020 challenge (Kristan et al. 2020), SiamDW_D, LTDSEd and SiamM_Ds from the VOT-RGBD 2019 challenge (Kristan et al. 2019), DAL (Qian et al. 2021) and DeT (Yan

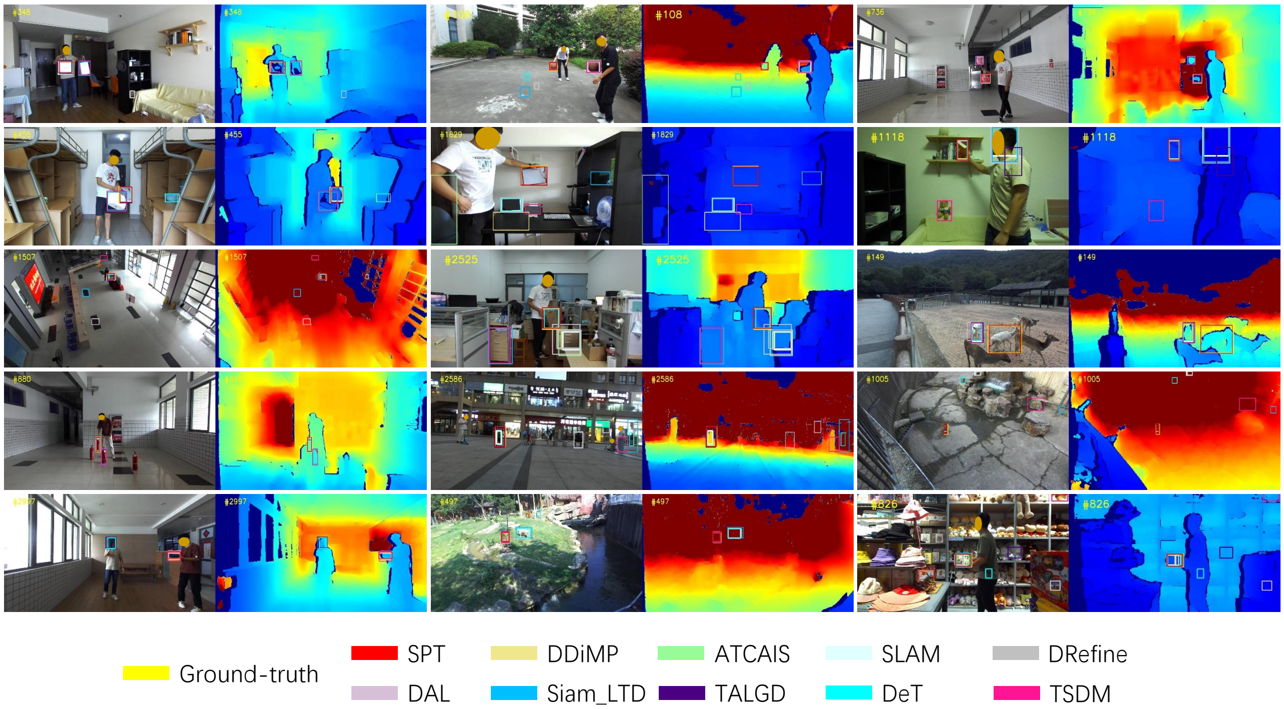

Figure 4: An illustration of the qualitative experimental results on several challenging sequences of the RGBD1K test set. Each RGB-D image is a sample from one particular sequence. The colour bounding boxes distinguish the ground-truth annotation and the results obtained by SPT, DDiMP, ATCAIS, SLAM, DRefine, DAL, Siam_LTD, TALGD, DeT, and TSDM, respectively.

et al. 2021b). From the results, the SPT achieves the best F-score of 0.538 and Recall of 0.549 on the DepthTrack dataset. In Table. 5, the proposed SPT tracker is compared with DDiMP, ATCAIS, CLGS_D SiamDW_D, LTDSEd, Siam_LTD, SiamM_Ds, OTR (Kart et al. 2019) and DeT. As can be seen, the SPT achieves significant superiority against the state-of-the-art trackers in terms of Recall. Although the SPT achieves inferior Precision and F-score compared to DDiMP, ATCAIS and CLGS_D, our SPT tracker has an obvious advantage in tracking speed. The results provided above authenticate that the method offline trained with a large amount of real RGB-D data, such as the proposed baseline SPT tracker, can provide superior performance for RGB-D tracking. On the other hand, the results also confirm the significance of the proposed RGBD1K dataset for advanced RGB-D object tracking, although each video of RGBD1K is only annotated with the first 600 frames.

Qualitative Comparison: To intuitively display the advantages of our method, in Fig. 4, we provide a qualitative comparison of the tested RGB-D trackers, including SPT, DDiMP, ATCAIS, SLAM, DRefine, DAL, Siam_LTD, TALGD, DeT, and TSDM, on several challenging videos from the RGBD1K dataset. As can be seen in the figure, although suffering from different challenging factors, such as similar objects, partial occlusion, reappearing from full occlusion, camera motion, etc, our SPT can perform precise and steady tracking on these challenging videos. Trained with additional depth images offline, the SPT can effectively alleviate the problem of similar objects, since two objects may vary in depth appearances when they are similar in visual appearance. Besides, the proposed fusion module in SPT enables to effectively fuse and enhance the features of RGB and depth modalities, making the depth information and the visual information complement each other, which helps the SPT to mitigate various complicated issues in RGB-D videos. Therefore, undoubtedly, our SPT tracker can achieve promising RGB-D tracking performance.

Conclusion

In this work, we proposed a large-scale dataset for RGB-D tracking as well as a baseline tracker based on an end-to-end deep network. This work is motivated by the scarcity of available annotated RGB-D videos that has hindered the development of RGB-D tracking. The proposed RGBD1K dataset contains more than twice as many videos as all the existing publicly available RGB-D videos for RGB-D object tracking. To demonstrate the utility of the RGBD1K dataset, we designed a new baseline method named SPT for RGB-D tracking. The SPT is trained offline using all the RGB-D videos of the training set of RGBD1K. The extensive experimental results obtained using the RGBD1K test set, DepthTrack test set and CDTB dataset, have demonstrated the benefits of training on RGBD1K, and its capacity to promote the development of RGB-D trackers in the future. The supplementary material is available here: https://drive.google.com/drive/folders/1xpHP-R2VTbp8n2sfipHlNAgdHKXrvsnC?usp=share_link.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (62020106012, U1836218, 61672265), the 111 Project of Ministry of Education of China (B12018), and the Engineering and Physical Sciences Research Council (EPSRC) (EP/N007743/1, MURI/EPSRC/DSTL, EP/R018456/1).

References

Awwad, S.; Hussein, F.; and Piccardi, M. 2015. Local depth patterns for tracking in depth videos. In *Proceedings of the 23rd ACM international conference on Multimedia*, 1115–1118.

Bagautdinov, T.; Fleuret, F.; and Fua, P. 2015. Probability occupancy maps for occluded depth images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2829–2837.

Bhat, G.; Danelljan, M.; Gool, L. V.; and Timofte, R. 2019. Learning discriminative model prediction for tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6182–6191.

Bibi, A.; Zhang, T.; and Ghanem, B. 2016. 3d part-based sparse tracker with automatic synchronization and registration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1439–1448.

Camplani, M.; Hannuna, S. L.; Mirmehdi, M.; Damen, D.; Paiement, A.; Tao, L.; and Burghardt, T. 2015. Real-time RGB-D Tracking with Depth Scaling Kernelised Correlation Filters and Occlusion Handling. In *BMVC*, volume 3, 01–12.

Danelljan, M.; Bhat, G.; Khan, F. S.; and Felsberg, M. 2019. Atom: Accurate tracking by overlap maximization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4660–4669.

Danelljan, M.; Bhat, G.; Shahbaz Khan, F.; and Felsberg, M. 2017. Eco: Efficient convolution operators for tracking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 6638–6646.

Fan, H.; Lin, L.; Yang, F.; Chu, P.; Deng, G.; Yu, S.; Bai, H.; Xu, Y.; Liao, C.; and Ling, H. 2019. Lasot: A high-quality benchmark for large-scale single object tracking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5374–5383.

Griffiths, E.; Kalyanaraman, A.; Ranjan, J.; and Whitehouse, K. 2017. An empirical design space analysis of doorway tracking systems for real-world environments. *ACM Transactions on Sensor Networks (TOSN)*, 13(4): 1–34.

Hannuna, S.; Camplani, M.; Hall, J.; Mirmehdi, M.; Damen, D.; Burghardt, T.; Paiement, A.; and Tao, L. 2019. Ds-kcf: a real-time tracker for rgb-d data. *Journal of Real-Time Image Processing*, 16(5): 1439–1458.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Huang, L.; Zhao, X.; and Huang, K. 2019. Got-10k: A large high-diversity benchmark for generic object tracking in the wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(5): 1562–1577.

Kart, U.; Kamarainen, J.-K.; and Matas, J. 2018. How to make an rgbd tracker? In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, 01–15.

Kart, U.; Lukezic, A.; Kristan, M.; Kamarainen, J.-K.; and Matas, J. 2019. Object tracking by reconstruction with view-specific discriminative correlation filters. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1339–1348.

Kristan, M.; Leonardis, A.; Matas, J.; Felsberg, M.; Pflugfelder, R.; Kämäräinen, J.-K.; Danelljan, M.; Zajc, L. Č.; Lukežič, A.; Drbohlav, O.; et al. 2020. The eighth visual object tracking VOT2020 challenge results. In *European Conference on Computer Vision*, 547–601. Springer.

Kristan, M.; Leonardis, A.; Matas, J.; Felsberg, M.; Pflugfelder, R.; ˇCehovin Zajc, L.; Vojir, T.; Bhat, G.; Lukezic, A.; Eldesokey, A.; et al. 2018. The sixth visual object tracking vot2018 challenge results. In *Proceedings of the European Conference on Computer Vision Workshops*, 01–52. Springer.

Kristan, M.; Matas, J.; Leonardis, A.; Felsberg, M.; Pflugfelder, R.; Kamarainen, J.-K.; Cehovin Zajc, L.; Drbohlav, O.; Lukezic, A.; Berg, A.; et al. 2019. The seventh visual object tracking vot2019 challenge results. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 01–36.

Kristan, M.; Matas, J.; Leonardis, A.; Felsberg, M.; Pflugfelder, R.; Kämäräinen, J.-K.; Chang, H. J.; Danelljan, M.; Cehovin, L.; Lukežič, A.; et al. 2021. The ninth visual object tracking vot2021 challenge results. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2711–2738.

Liu, Y.; Jing, X.-Y.; Nie, J.; Gao, H.; Liu, J.; and Jiang, G.-P. 2018. Context-aware three-dimensional mean-shift with occlusion handling for robust object tracking in RGB-D videos. *IEEE Transactions on Multimedia*, 21(3): 664–677.

Lukezic, A.; Kart, U.; Kapyla, J.; Durmush, A.; Kamarainen, J.-K.; Matas, J.; and Kristan, M. 2019. Cdtb: A color and depth visual object tracking dataset and benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10013–10022.

Lukezic, A.; Matas, J.; and Kristan, M. 2020. D3S-A discriminative single shot segmentation tracker. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7133–7142.

Lukežič, A.; Zajc, L. Č.; Vojíř, T.; Matas, J.; and Kristan, M. 2018. Now you see me: evaluating performance in long-term visual tracking. *arXiv preprint arXiv:1804.07056*, 01–10.

Meshgi, K.; Maeda, S.-i.; Oba, S.; Skibbe, H.; Li, Y.-z.; and Ishii, S. 2016. An occlusion-aware particle filter tracker to handle complex and persistent occlusions. *Computer Vision and Image Understanding*, 150: 81–94.

Muller, M.; Bibi, A.; Giancola, S.; Alsubaihi, S.; and Ghanem, B. 2018. Trackingnet: A large-scale dataset and benchmark for object tracking in the wild. In *Proceedings* *of the European Conference on Computer Vision (ECCV)*, 300–317.

Qian, Y.; Yan, S.; Lukežič, A.; Kristan, M.; Kämäräinen, J.-K.; and Matas, J. 2021. DAL: A Deep Depth-Aware Long-term Tracker. In *2020 25th International Conference on Pattern Recognition (ICPR)*, 7825–7832. IEEE.

Smeulders, A. W.; Chu, D. M.; Cucchiara, R.; Calderara, S.; Dehghan, A.; and Shah, M. 2013. Visual tracking: An experimental survey. *IEEE transactions on pattern analysis and machine intelligence*, 36(7): 1442–1468.

Song, S.; and Xiao, J. 2013. Tracking revisited using RGBD camera: Unified benchmark and baselines. In *Proceedings of the IEEE international conference on computer vision*, 233–240.

Valmadre, J.; Bertinetto, L.; Henriques, J. F.; Tao, R.; Vedaldi, A.; Smeulders, A. W.; Torr, P. H.; and Gavves, E. 2018. Long-term tracking in the wild: A benchmark. In *Proceedings of the European conference on computer vision (ECCV)*, 670–685.

Xiao, J.; Stolkin, R.; Gao, Y.; and Leonardis, A. 2017. 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.

Xu, T.; Feng, Z.-H.; Wu, X.-J.; and Kittler, J. 2019. Joint group feature selection and discriminative filter learning for robust visual object tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7950–7960.

Xue, X.; Li, Y.; Yin, X.; Shang, C.; Peng, T.; and Shen, Q. 2020. Semantic-Aware Real-Time Correlation Tracking Framework for UAV Videos. *IEEE Transactions on Cybernetics*, 01–12.

Yan, B.; Peng, H.; Fu, J.; Wang, D.; and Lu, H. 2021a. Learning spatio-temporal transformer for visual tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10448–10457.

Yan, S.; Yang, J.; Käpylä, J.; Zheng, F.; Leonardis, A.; and Kämäräinen, J.-K. 2021b. Depthtrack: Unveiling the power of rgbd tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10725–10733.

Zhao, P.; Liu, Q.; Wang, W.; and Guo, Q. 2021a. TSDM: Tracking by SiamRPN++ with a Depth-refiner and a Mask-generator. In *2020 25th International Conference on Pattern Recognition (ICPR)*, 670–676. IEEE.

Zhao, S.; Xu, T.; Wu, X.-J.; and Kittler, J. 2022. Distillation, Ensemble and Selection for building a Better and Faster Siamese based Tracker. *IEEE Transactions on Circuits and Systems for Video Technology*, 01–13.

Zhao, S.; Xu, T.; Wu, X.-J.; and Zhu, X.-F. 2021b. Adaptive feature fusion for visual object tracking. *Pattern Recognition*, 111: 107679.

Zhu, X.-F.; Wu, X.-J.; Xu, T.; Feng, Z.; and Kittler, J. 2021. Robust Visual Object Tracking via Adaptive Attribute-Aware Discriminative Correlation Filters. *IEEE Transactions on Multimedia*, 24: 301–312.