- Supplementary Materials -
ARKitTrack: A New Diverse Dataset for Tracking Using Mobile RGB-D Data

A. Comparison on ARKitTrack-VOS-Test

We compare 3[1] state-of-art recent popular methods [1, 4, 11] using their provided models which are finetuned on Davis [6] and YoutubeVOS2019 [8]. As shown in Table 1, the overall performance on our dataset is always lower than that on DAVIS2017 [6] and YouTubeVOS2019 [8]. It confirms that the proposed ARKitTrack is more challenging than the existing RGB VOS datasets.

B. Quantitative Results

RGB-D VOT. We compare the proposed method with 7 recent RGB-D trackers. The 7 RGB-D tracker includes 3 recent methods (DeT [10], DAL [7], and TSDM [12]) and the top four trackers from the VOT-RGBD 2021 challenge [3] (STARK_RGBD, TALGD, ATCAIS, DDiMP). Among them, ATCAIS and DDiMP are also the top trackers of VOT-RGBD 2020 challenge [2]. The quantitative results are shown in Figure 1.

RGB-D VOS. We select 4 state-of-the-art RGB-VOS methods for comparison on ARKitTrack-VOS-Test, including STCN [1], RPCM [9], AOT (SwinB-L) [11] and QDMN [4]. Besides, We design a variant named STCN_RGBD for RGB-D VOS by adding an additional depth branch to STCN and fusing RGBD features through concatenation. The quantitative results are shown in Figure 2 and 3.

C. Frame-level Attributes

We summarize the attribute description of ARKitTrack-VOT-Test, which is shown in Table 2. We state that per-frame attribute annotations can be used to fully exploit the effectiveness of attribute-aware trackers. We follow the pre-

[1]As RPCM [9] does not provide a pre-trained model, the results are not presented

Table 1. Comparison results of 3 state-of-the-art RGB methods on the existing RGB VOS datasets.

Tracker	ARKitTrack			DAVIS2017			YoutubeVOS2018			Description	
	$\mathcal{J}\&\mathcal{F}\uparrow$	$\mathcal{J}\uparrow$	$\mathcal{F}\uparrow$	$\mathcal{J}\&\mathcal{F}\uparrow$	$\mathcal{J}\uparrow$	$\mathcal{F}\uparrow$	$\mathcal{J}\&\mathcal{F}\uparrow$	$\mathcal{J}\uparrow$	$\mathcal{F}\uparrow$	Type	Year
STCN	0.526	0.491	0.560	0.854	0.822	0.886	0.830	0.799	0.861	RGB	2021
AOT-SwinB-L	0.735	0.704	0.766	0.854	0.824	0.884	0.845	0.811	0.879	RGB	2022
QDMN	0.701	0.670	0.732	0.856	0.825	0.886	0.830	0.862	0.560	RGB	2022

Table 2. Per-frame attributes include 11 manually annotated attributes and 5 ones calculated from the groundtruth.

Tag	Attribute	Description	Annotation
AC	Aspect-ratio Change	When the ratio between the maximum and minimum aspect in 21 consecutive frames was larger than 1.5.	Calculated
BC	Background Clutter	The background near the target has the similar color or texture as the target.	Manually
DC	Depth Clutter	The depth map near the target has complex depth distribution or the similar depth as the target.	Manually
EI	Extreme Illumination	The target is in low or high light condition.	Manually
FM	Fast Moving	The target center moves by at least 30% of its size in consecutive frames.	Calculated
FO	Full Occlusion	The target is fully occluded.	Manually
LD	Low Depth Quality	When the number of low confidence depth values in the bounding box was more than 50%.	Calculated
ND	Non-rigid Deformation	The non-rigid object deformation.	Manually
OP	Out-of-plane Rotation	Target rotates out of the plane.	Manually
OV	Out-of-View	The target is partially or completely missing in the current view.	Manually
PO	Partial Occlusion	The target is partially occluded.	Manually
SO	Similar Objects	There are adjacent objects whose appearance is similar to the target.	Manually
SC	Size Change	When the ratio between the maximum and minimum target size in 21 consecutive frames is larger than 1.5.	Calculated
RT	Reflective Target	Interface of the target is reflective.	Manually
TB	Target Blur	Target is blurry caused by illumination or motion.	Manually
NaN	Unassigned	There are no aforementioned cases appearing in the frame.	Calculated

Figure 1. Quantitative results of several RGB-D visual tracking methods on ARKitTrack-VOT-Test.

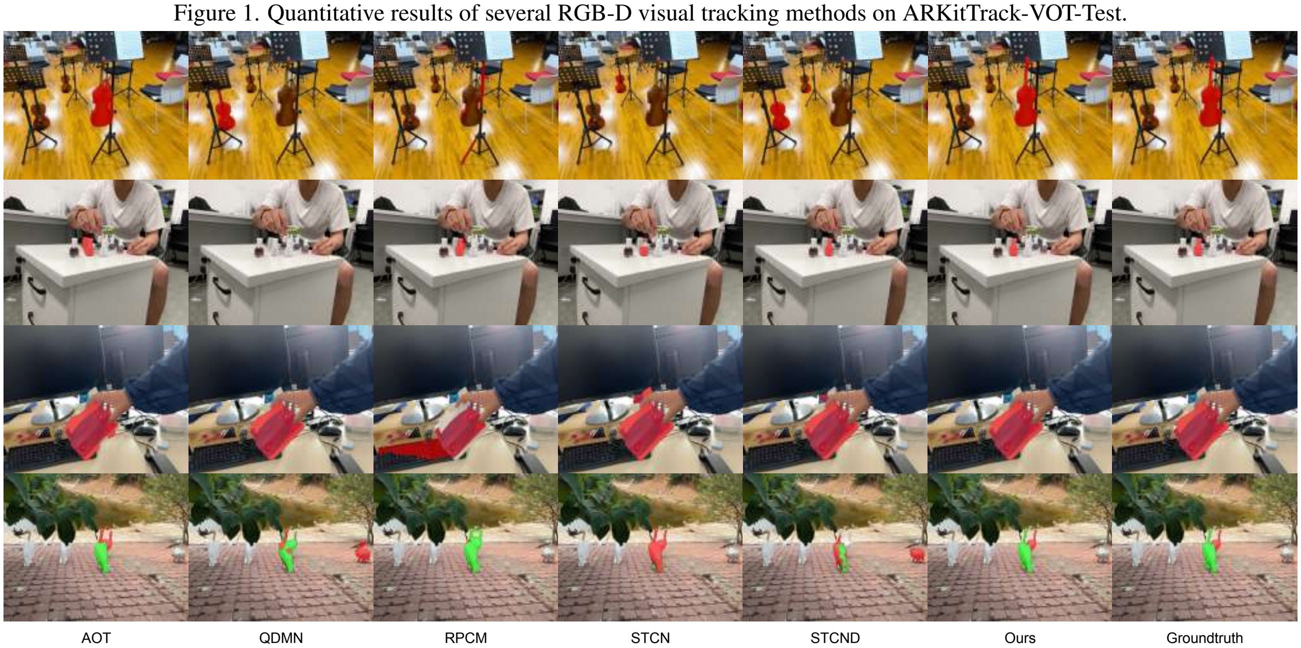

Figure 2. Quantitative results of several RGB(-D) video object segmentation methods on ARKitTrack-VOS-Test.

Figure 3. Quantitative results of several RGB(-D) video object segmentation methods over time. GT denotes the groundtruth.

vious works [5, 10] to perform the per-attribute evaluation.

D. Failure Cases

As shown in Figure 4 and 5, we provide some failure cases to show the limitations of our method. For both VOT and VOS, our method often suffers target missing when there are many similar objects or the target moves fast. Complicated environments, such as occlusion and depth clutter, can also cause tracking failure.

E. Back-Projected BEV Feature

In our work, we fuse color and depth information in the BEV space and back-project the fused feature to the image plane for 2D tasks. We visualize some back-projected features for better understanding, shown in Figure 6. By exploring the space geometry cues in the BEV space, the target information can be enhanced.

References

[1] Ho Kei Cheng, Yu-Wing Tai, and Chi-Keung Tang. Rethinking space-time networks with improved memory coverage for efficient video object segmentation. In *Advances in Neural Information Processing Systems*, pages 11781–11794, 2021. 1

[2] Matej Kristan, Ales Leonardis, Jiri Matas, Michael Felsberg, Roman P. Pflugfelder, Joni-Kristian Kämäräinen, Martin Danelljan, et al. The eighth visual object tracking VOT2020 challenge results. In Adrien Bartoli and Andrea Fusiello, editors, *Proceedings of the European Conference on Computer Vision Workshops*, pages 547–601, 2020. 1

[3] Matej Kristan, Jirí Matas, Ales Leonardis, Michael Felsberg, Roman P. Pflugfelder, Joni-Kristian Kämäräinen, Hyung Jin Chang, Martin Danelljan, et al. The ninth visual object tracking VOT2021 challenge results. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 2711–2738, 2021. 1

[4] Yong Liu, Ran Yu, Fei Yin, Xinyuan Zhao, Wei Zhao, Weihao Xia, and Yujiu Yang. Learning quality-aware dynamic memory for video object segmentation. In *Proceedings of the European Conference on Computer Vision*, 2022. 1

[5] Alan Lukezic, Ugur Kart, Jani Käpylä, 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 International Conference on Computer Vision*, pages 10012–10021, 2019. 3

[6] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, and A. Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 1

[7] Yanlin Qian, Song Yan, Alan Lukezic, Matej Kristan, Joni-Kristian Kämäräinen, and Jiri Matas. DAL: A deep depth-aware long-term tracker. In *Proceedings of IEEE Conference on Pattern Recognition*, pages 7825–7832, 2020. 1

[8] Ning Xu, Linjie Yang, Yuchen Fan, Jianchao Yang, Dingcheng Yue, Yuchen Liang, Brian L. Price, Scott Cohen, and Thomas S. Huang. Youtube-vos: Sequence-to-sequence video object segmentation. In *Proceedings of the European Conference on Computer Vision*, pages 603–619, 2018. 1

[9] Xiaohao Xu, Jinglu Wang, Xiao Li, and Yan Lu. Reliable propagation-correction modulation for video object segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022. 1

[10] Song Yan, Jinyu Yang, Jani Käpylä, Feng Zheng, Ales Leonardis, and Joni-Kristian Kämäräinen. Depthtrack: Unveiling the power of RGBD tracking. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 10705–10713, 2021. 1, 3

[11] Zongxin Yang, Yunchao Wei, and Yi Yang. Associating objects with transformers for video object segmentation. In *Advances in Neural Information Processing Systems*, pages 2491–2502, 2021. 1

[12] Pengyao Zhao, Quanli Liu, Wei Wang, and Qiang Guo. TSDM: tracking by siamrpn++ with a depth-refiner and a mask-generator. In *Proceedings of IEEE Conference on Pattern Recognition*, pages 670–676, 2020. 1

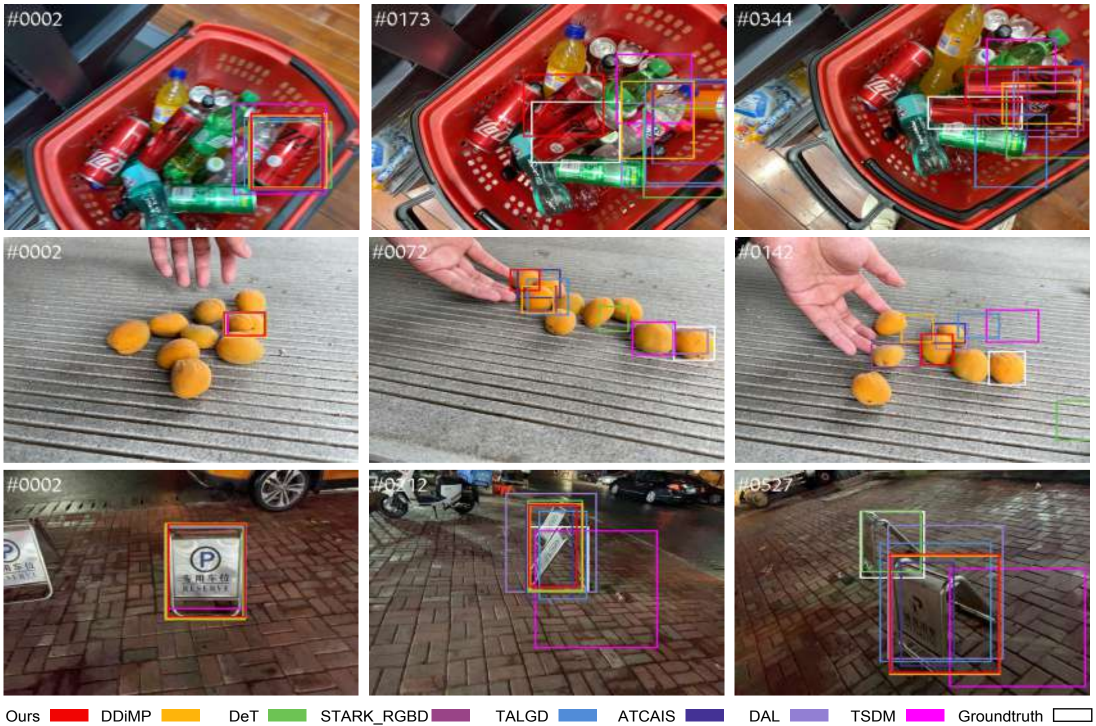

Figure 4. Failure cases of RGB-D VOT methods. Our tracker and many other methods fail to track in depth clutter (1st row), fast motion (2nd row), and similar objects (3rd row).

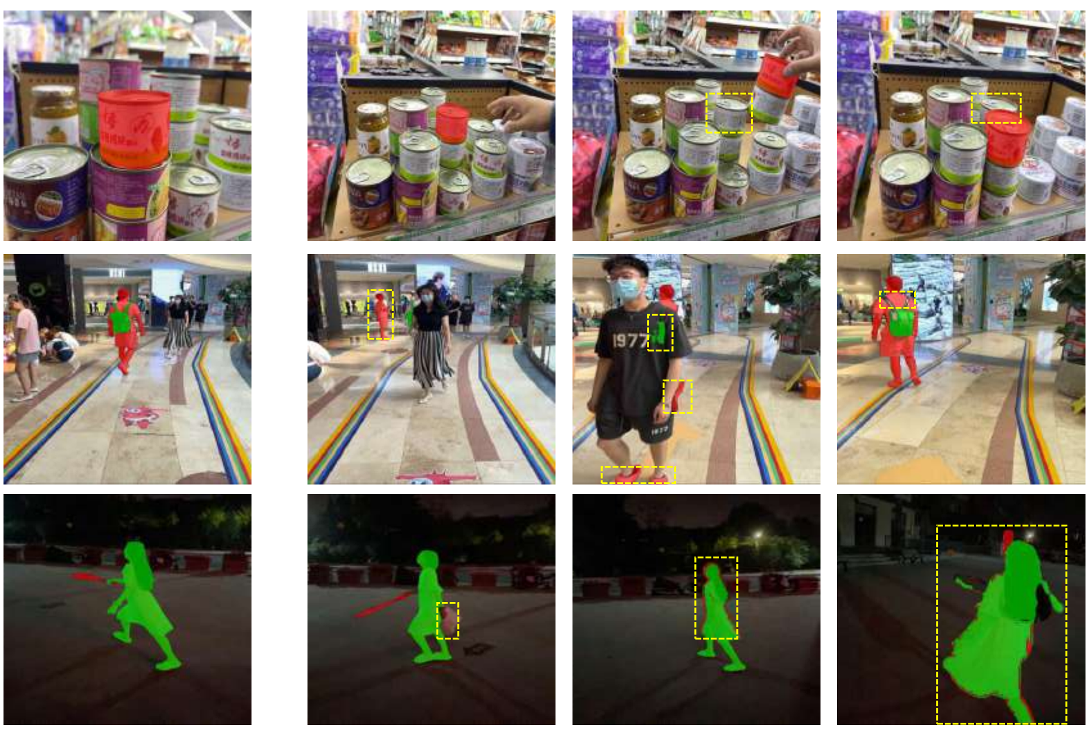

Figure 5. Failure cases. We box the failed segmentation regions out in the Yellow dashed rectangular. First row: multiple cans with the same appearance are being selected. We fail to discriminate the target one that is occluded by others. Second row: A person is walking in a mall. We cannot catch up as the man is covered by the background. Third row: a girl is playing table tennis. We fail to segment as the target is moving quickly with a large motion blur and depth clutter.

Figure 6. Some examples of the back-projected BEV features.