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

Xue-Feng Zhu · Tianyang Xu · Zhangyong Tang · Zucheng Wu · Haodong Liu · Xiao Yang · Xiao-Jun Wu · Josef Kittler
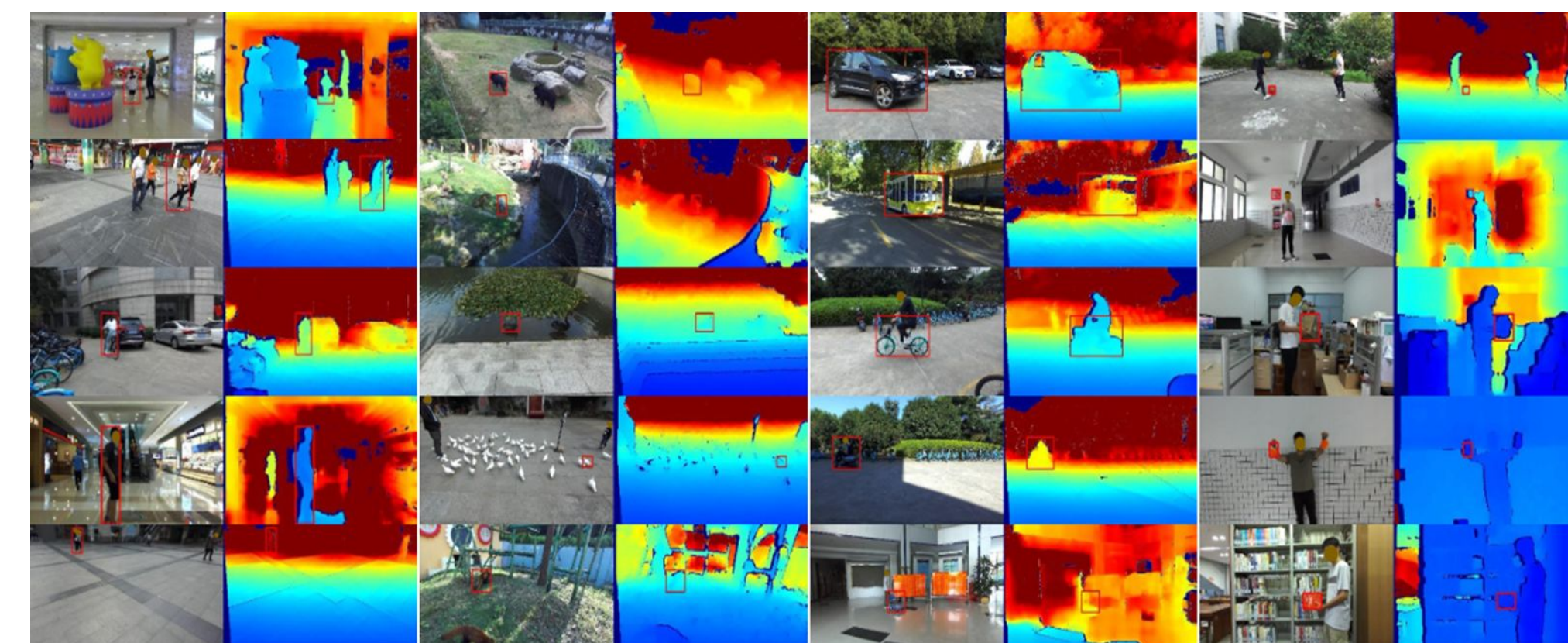
JIANGNAN UNIVERSITY   UNIVERSITY OF SURREY

Introduction

RGB-D object tracking has attracted considerable attention recently. 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. 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 are available on the project homepage: https://github.com/xuefeng-zhu5/RGBD1K.
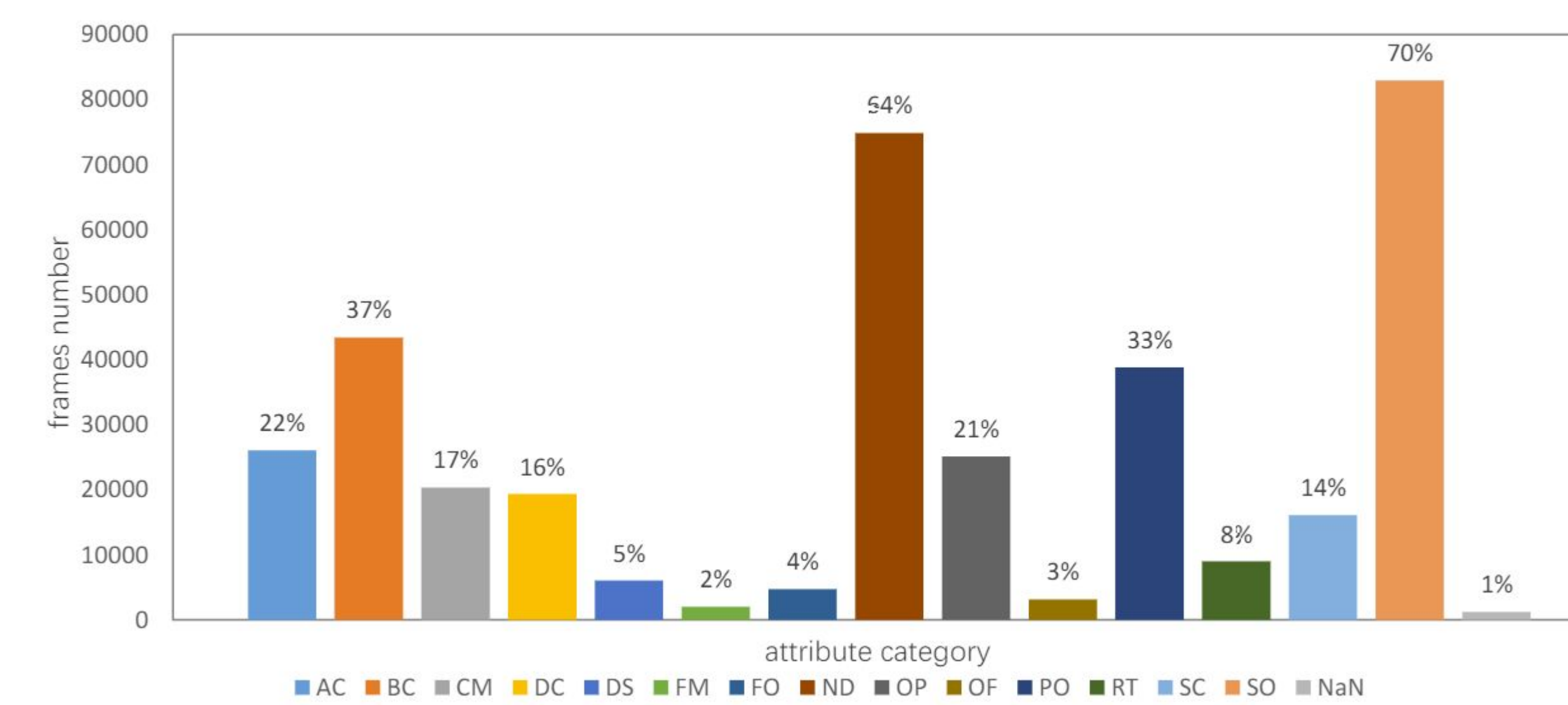
RGBD1K dataset

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

↑ Comparation of the existing RGB-D datasets.

- 1,050 videos, 2.5 million frames
- 717,900 annotated frames
- more than 110 object categories
- dozens of various scenes
- 15 challenging attributes

↑ The distribution of frames of different attributes in the RGBD1K test set.

Baseline tracker SPT

$$L = \lambda_{iou} L_{iou}(b_i, \hat{b}_i) + \lambda_{L_1} L_1(b_i, \hat{b}_i)$$

↑ 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
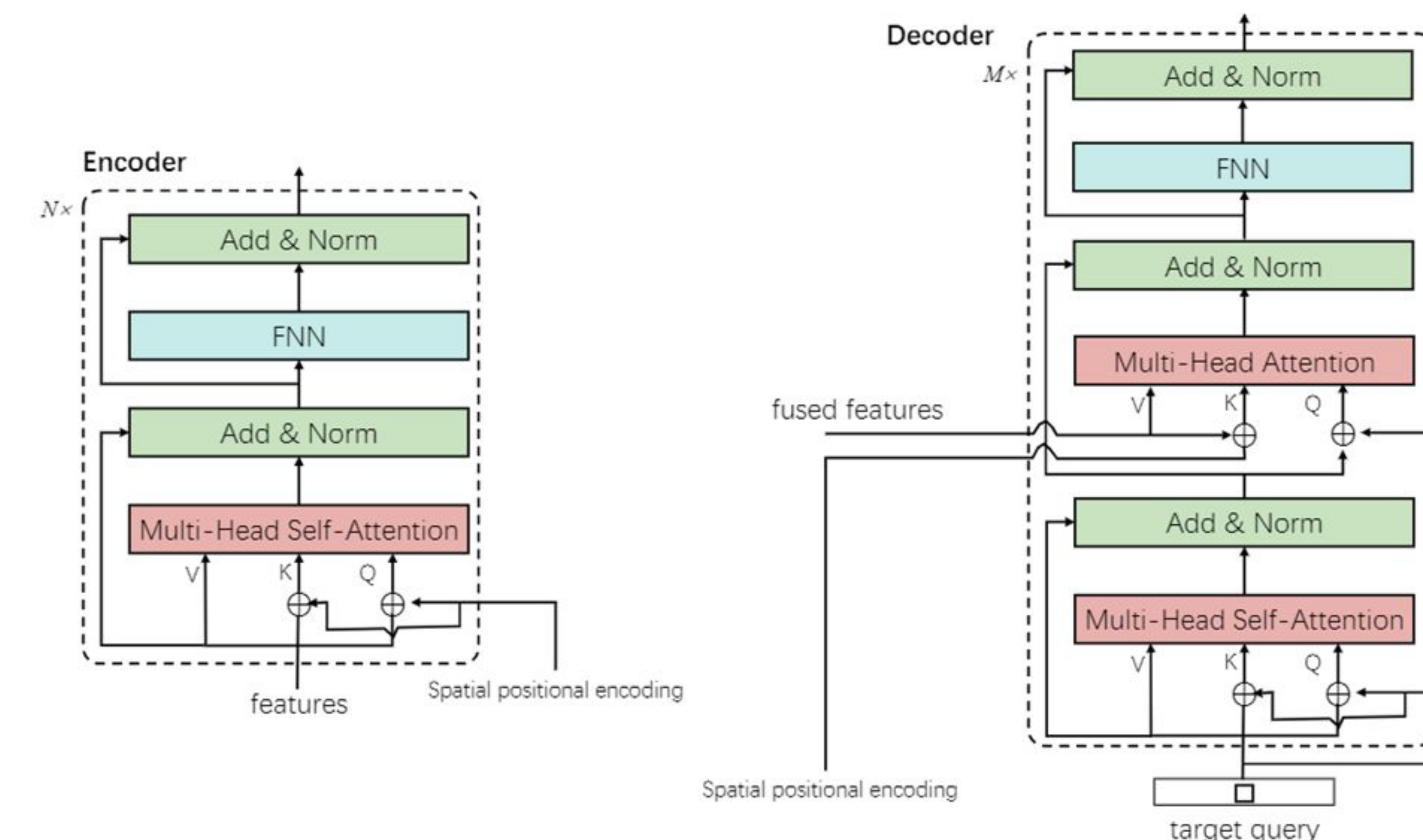
↑ Illustration of the architectures of transformer encoder and decoder.

Ablation study

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

↑ Ablation performance on the RGBD1K, DepthTrack and CDTB datasets. STARK-S is the baseline of SPT. 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.

Comparason with SOTA trackers

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

↑ The tracking results on the RGBD1K test set.

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

↑ The tracking results on the DepthTrack test set.

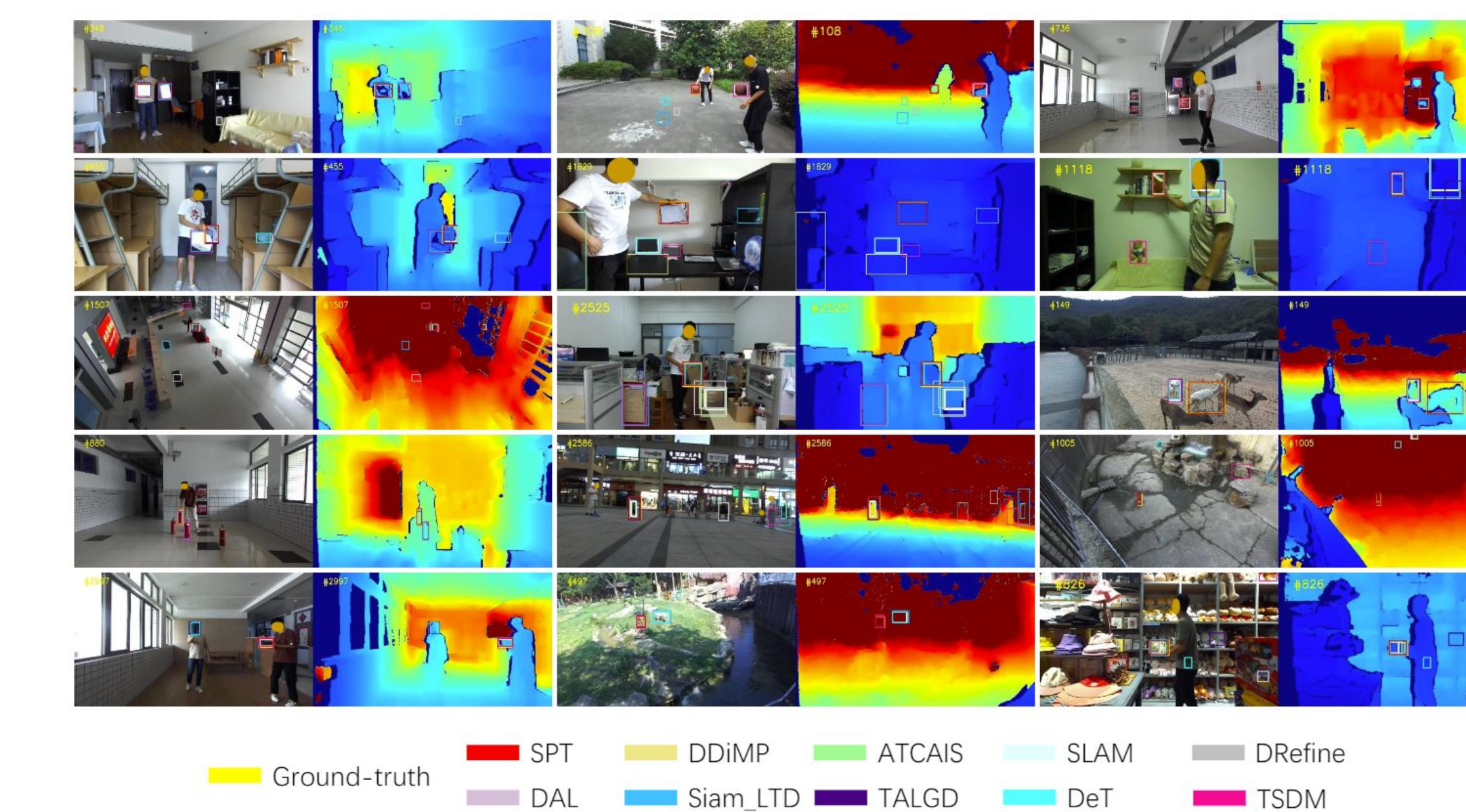| Method	DDiMP	ATCAIS	CLGS_D	SiamDW_D	LTDSEd	Siam_LTD	SiamM_Ds	OTR	DeT	SPT
Pr	0.703	0.709	0.725	0.677	0.674	0.626	0.685	0.364	0.674	0.654
Re	0.689	0.696	0.664	0.685	0.643	0.489	0.677	0.312	0.642	0.726
F-score	0.696	0.702	0.693	0.681	0.658	0.549	0.681	0.336	0.657	0.688
Speed (fps)	4.7	1.3	7.3	3.8	5.7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

↑ The tracking results on the CDTB dataset.

Qualitative experimental results

Acknowledgements