Efficient RGB-T Tracking via Cross-Modality Distillation

1. Attribute-based performance

We further analyze the attribute-based performance on RGBT234 [4] and LasHeR [6] dataset.

1.1. RGBT234 dataset

RGBT234 [4] is a large-scale RGB-T tracking dataset. It contains 12 challenge attribute labels, including no occlusion (NO), partial occlusion (PO), heavy occlusion (HO), low illumination (LI), low resolution (LR), thermal crossover (TC), deformation (DEF), fast motion (FM), scale variation (SV), motion blur (MB), camera moving (CM) and background clutter (BC).

As shown in Table 1, we analyze the attribute-based performance on RGBT234. For clarity, we only illustrate the four re-trained trackers and another four advanced trackers, i.e., JMMAC [11], M5L [7], CAT [5] and MANet++ [12]. From the results, we can see that our proposed method still performs well in most annotated attributions. Compared with such trackers based on MDNet (i.e., M5L [7], CAT [5], MANet++ [12], DAFNet [2] and FANet [14]), our method has remarkable improvements in case of PO, LI, DEF and SV. Compared with mfDiMP [10], which is based on DiMP [1] and employs two ResNet50 [3] for feature extraction, our algorithm achieves competitive performance but significantly reduce parameters.

1.2. LasHeR dataset

LasHeR [6] is currently the largest RGB-T tracking dataset. In addition to such challenges in RGBT234, LasHeR contains more challenges, including total occlusion (TO), hyaline occlusion (HO), high illumination (HI), abrupt illumination variation (AIV), similar appearance (SA), aspect ratio change(ARC), out-of-view (OV) and frame lost (FL).

As shown in Table 2, we further analyze the attribute-based performance on LasHeR. The results of our proposed method and some other state-of-the-art trackers, including MANet [8], DAPNet [13], DAFNet [2], MACNet [9], CAT [5], mfDiMP [10], FANet [14] and MANet++ [12], demonstrate that the our method performs the best under the most challenging conditions. First, in adverse lighting conditions, thermal crossover and low resolution, our method outperforms all other trackers. This demonstrates that the proposed method can enable such a compact model to fully explore the complementary information within multi-modal images. Second, our framework is robust to significant appearance changes, such as deformation, scale variation, camera moving and similar appearance. Finally, our model struggles to handle the out of view challenge and hyaline occlusion challange. It may be due to the fact that student model can not learn effective information from the teacher model when targets are invisible.

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. 1

[2] Yuan Gao, Chenglong Li, Yabin Zhu, Jin Tang, Tao He, and Futian Wang. Deep adaptive fusion network for high performance rgbt tracking. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2019. 1, 2

[3] 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. 1

[4] Chenglong Li, Xinyan Liang, Yijuan Lu, Nan Zhao, and Jin Tang. RGB-T object tracking: Benchmark and baseline. *Pattern Recognition*, 96:106977, 2019. 1

[5] Chenglong Li, Lei Liu, Andong Lu, Qing Ji, and Jin Tang. Challenge-aware rgbt tracking. In *European Conference on Computer Vision*, pages 222–237. Springer, 2020. 1, 2

[6] Chenglong Li, Wanlin Xue, Yaqing Jia, Zhichen Qu, Bin Luo, and Jin Tang. Lasher: A large-scale high-diversity benchmark for rgbt tracking. *arXiv preprint arXiv:2104.13202*, 2021. 1

[7] Yifan Liu, Ke Chen, Chris Liu, Zengchang Qin, Zhenbo Luo, and Jingdong Wang. Structured knowledge distillation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2604–2613, 2019. 1, 2

[8] Cheng Long Li, Andong Lu, Ai Hua Zheng, Zhengzheng Tu, and Jin Tang. Multi-adapter RGBT tracking. In *Proceedings of the IEEE Conference on Computer Vision Workshops*, 2019. 1, 2

[9] Hui Zhang, Lei Zhang, Li Zhuo, and Jing Zhang. Object tracking in RGB-T videos using modal-aware attention net-

Table 1. Attribute-based precision rate and success rate (PR/SR) scores obtained by using different trackers on RGBT234 dataset. The numbers with red colors indicate the best results.

Trackers	JMMAC [11]	M5L [7]	CAT [5]	MANet++ [12]	MANet [8]	DAFNet [2]	FANet [14]	mfDiMP [10]	Student-Distill
Pub. Info.	TIP2021	TIP2022	ECCV2020	TIP2021	ICCVW2019	ICCVW2019	TIV2021	ICCVW2019	2023
NO	93.2/69.4	93.1/64.6	89.8/65.4	93.2/66.8	91.4/66.6	91.4/63.8	93.1/63.4	92.6/68.7	90.4/65.7
PO	84.1/61.1	86.3/58.9	85.2/59.3	85.1/59.3	83.9/59.1	86.1/58.7	83.9/56.6	88.2/62.7	89.6/64.1
HO	67.7/48.3	66.5/45.0	70.4/47.1	70.0/48.0	67.6/46.9	68.7/47.0	68.9/47.0	72.1/49.4	71.7/49.5
LI	84.0/58.8	82.1/54.7	81.1/55.1	81.0/54.7	75.9/52.7	80.5/54.1	81.5/54.6	85.5/60.7	88.3/61.7
LR	84.0/58.8	82.3/53.5	82.3/54.5	82.0/53.9	78.7/51.8	79.3/51.1	81.0/50.7	76.2/50.1	73.7/48.1
TC	74.9/52.6	82.1/56.4	80.3/57.6	80.3/57.6	75.0/54.3	77.9/54.6	76.9/52.8	79.8/55.5	75.8/51.2
DEF	70.6/51 9	73.6/51.1	75.3/53.5	76.2/54.1	73.6/53.5	75.1/53.2	75.4/53.5	81.6/59.6	82.7/60.9
FM	61.0/41.7	72.8/49.5	70.0/45.3	73.1/47.0	71.3/46.4	69.1/45.6	68.6/44.4	76.5/54.3	72.5/52.2
SV	83.7/61.6	79.6/54.2	78.9/55.4	79.7/56.6	78.6/56.2	82.1/56.7	80.0/54.8	84.0/60.8	84.5/61.3
MB	75.1/54.9	73.8/52.8	72.0/51.1	68.3/49.0	70.3/51.1	72.1/50.4	73.3/51.5	77.5/55.4	74.9/53.5
CM	76.2/55.6	75.2/52.9	74.7/52.3	75.2/52.7	69.7/50.9	74.3/53.2	73.6/53.2	81.5/58.9	80.2/58.0
BC	68.7/48.5	75.0/47.7	76.7/49.1	81.1/51.9	74.4/49.3	74.0/47.5	76.9/48.4	77.5/50.8	80.6/53.0
ALL	79.0/57.3	79.5/54.2	80.0/55.4	80.4/56.1	78.6/55.5	80.0/54.9	79.4/53.9	82.4/58.3	82.4/58.4

Table 2. Attribute-based precision rate and success rate (PR/SR) scores obtained by using different trackers on LasHeR dataset. The numbers with red colors indicate the best results.

	DAPNet [13]	MANet [8]	MaCNet [9]	CAT [5]	MANet++ [12]	DAFNet* [2]	FANet* [14]	mfDiMP* [10]	Student-Distill
Pub. Info.	ACM MM2019	ICCVW2019	Sensors2020	ECCV2020	TIP2021	ICCVW2019	TIV2021	ICCVW2019	2023
NO	69.8/47.9	67.2/46.3	74.0/51.7	65.4/43.0	63.6/40.7	66.2/46.2	70.2/47.6	81.3/64.3	85.2/66.0
PO	39.1/29.1	42.4/30.7	44.6/32.8	41.8/29.5	44.0/30.1	44.9/29.3	44.9/32.2	54.8/42.8	55.3/44.6
TO	32.5/24.5	35.0/26.0	38.6/29.2	36.1/26.0	35.4/25.4	36.0/27.2	39.4/28.6	47.7/37.0	48.7/40.1
HO	22.0/22.3	24.1/23.6	28.1/29.1	22.6/23.4	24.5/24.4	22.1/24.1	20.5/21.1	51.2/45.2	46.7/44.0
OV	33.9/31.3	32.1/34.9	34.8/36.7	26.0/23.0	28.0/22.0	45.2/37.3	25.7/24.4	57.1/49.8	45.2/40.7
LI	31.7/24.0	35.6/26.9	36.0/26.7	31.5/22.6	35.8/24.0	37.1/26.2	39.6/28.8	45.4/36.5	47.1/37.6
HI	51.3/35.3	47.3/34.4	52.0/37.4	52.5/35.7	53.3/34.7	52.2/34.7	53.7/36.2	67.8/52.6	65.6/53.7
AIV	16.2/12.6	14.5/14.8	17.3/15.6	22.6/19.0	18.8/15.8	17.2/14.8	19.7/16.8	31.7/29.3	38.8/34.0
LR	38.9/25.2	45.8/28.5	43.9/28.0	42.4/25.2	47.4/26.8	44.2/27.9	45.5/27.7	48.7/34.5	50.3/35.3
DEF	40.9/32.8	37.4/32.1	41.4/34.0	38.3/30.6	39.4/30.8	45.9/36.4	46.4/37.3	59.3/47.1	59.0/48.1
BC	35.8/28.1	38.3/30.2	42.2/31.9	39.8/29.8	43.6/31.4	42.9/32.9	41.2/30.7	52.0/40.3	53.6/42.3
SA	35.1/26.6	38.0/27.9	40.8/30.4	37.426.5	41.1/27.9	41.0/30.2	39.9/29.0	50.5/39.5	49.4/39.3
TC	36.0/26.1	38.6/27.3	39.8/28.7	37.0/26.2	40.1/26.8	40.7/29.0	41.8/29.1	50.7/38.8	51.6/40.1
MB	32.4/26.2	38.9/27.9	40.4/29.8	39.8/26.6	39.7/26.6	38.1/27.2	41.5/28.5	49.7/38.5	50.4/39.2
CM	38.7/28.8	42.8/31.2	46.7/33.9	41.9/29.4	42.2/29.4	44.8/32.6	44.3/32.1	56.2/43.0	56.8/44.2
FL	33.1/22.0	30.2/19.4	34.6/22.2	38.7/22.6	37.8/21.6	33.7/27.0	35.3/25.8	49.1/38.4	51.6/40.2
FM	37.8/28.9	41.0/30.6	43.7/33.0	39.9/29.1	41.1/28.9	44.1/32.5	43.5/31.9	57.1/45.0	56.9/45.2
SV	43.4/31.4	46.0/32.9	48.0/34.8	44.4/30.7	46.4/31.1	47.4/34.0	48.0/34.1	58.5/45.9	59.2/46.8
ARC	32.9/26.3	35.6/27.0	36.0/28.5	32.5/24.4	35.5/25.7	34/26.8	35.5/27.2	52.6/42.5	53.9/43.9
ALL	43.1/31.4	45.5/32.6	48.2/35.0	45.0/31.4	46.7/31.4	48.0/34.5	44.1/34.3	58.3/45.6	59.0/46.4

work and competitive learning. *Sensors*, 20(2):393, 2020. 1, 2

[10] Lichao Zhang, Martin Danelljan, Abel Gonzalez-Garcia, Joost van de Weijer, and Fahad Shahbaz Khan. Multi-modal fusion for end-to-end RGB-T tracking. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2019. 1, 2

[11] Pengyu Zhang, Jie Zhao, Chunjuan Bo, Dong Wang, Huchuan Lu, and Xiaoyun Yang. Jointly modeling motion and appearance cues for robust rgb-t tracking. *IEEE Transactions on Image Processing*, 30:3335–3347, 2021. 1, 2

[12] Pengyu Zhang, Jie Zhao, Chunjuan Bo, Dong Wang, Huchuan Lu, and Xiaoyun Yang. Jointly modeling motion and appearance cues for robust rgb-t tracking. *IEEE Transactions on Image Processing*, 30:3335–3347, 2021. 1, 2

[13] Yabin Zhu, Chenglong Li, Bin Luo, Jin Tang, and Xiao Wang. Dense feature aggregation and pruning for RGBT tracking. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 465–472, 2019. 1, 2

[14] Y. Zhu, C. Li, J. Tang, and B. Luo. Quality-aware feature aggregation network for robust rgbt tracking. *IEEE Transactions on Intelligent Vehicles*, 6(1):121–130, 2021. 1, 2