– Supplementary Material –
Visible-Thermal UAV Tracking: A Large-Scale Benchmark and New Baseline

1. Attribute Description of VTUAV

We summarize the attribute description of VTUAV, which is shown in Tab. 1. We state that comprehensive attribute annotation can fully exploit the effectiveness of attribute-aware trackers.

Table 1. Attribute description in VTUAV.

Name	Attribute description
TB	*Target Blur* - Target is blurry caused by illumination or motion.
CM	*Camera Movement* - Obvious camera motion occurs.
EI	*Extreme Illumination* - The target is in low or high light condition.
DEF	*Deformation* - The target is in no-rigid movement
PO	*Partial Occlusion* - The target is partial occluded by surroundings or some portion of target leaves the view.
FO	*Full Occlusion* - The target is fully occluded by surroundings
SV	*Scale Variation* - The ratio of current bounding box and initial bounding box is out of range $\tau \in [0.5, 2]$
TC	*Thermal cluttering* - The background near the target has the similar temperature as the target.
FM	*Fast Moving* - The offset of adjacent frames is larger than 20 pixels.
BC	*Background Clustering* - The background near the target has the similar color or texture as the target.
OV	*Out-of-View* - The target is completely missing in current view.
LR	*Low Resolution* - The area of bounding box is less than 400.
TVS	*Thermal-Visible Separation* - The bounding boxes in visible and thermal images have no overlap.

2. Attribute-based Comparison on VTUAV

We conduct attribute-specific analysis for both short-term and long-term trackers. As shown in Fig. 1, HMFT obtains the satisfying performance on all attributes, except fully occlusion, which is inferior to FSRPN [1]. This phenomenon may stem from HMFT being an online tracker, where the absence of target involves the learning noise of filter. While FSRPN, which is a Siamese-based offline method, relies on the initial template, achieving a better re-detection ability. As shown in Fig. 2, the proposed long-term tracker HMFT_LT shows its effectiveness on most of challenges, which outperforms the short-term version HMFT, with a large margin. This indicates that global mechanism achieves more than 10% promotion in both MSR and MPR when the target is in out-of-view or fully occlusion. All the short-term trackers perform poor performance on TVS, TC and OV, which indicates the necessity of global detection component.

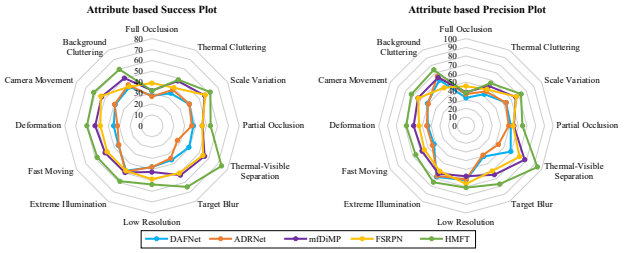

Figure 1. Attribute-based comparison on the short-term subset.

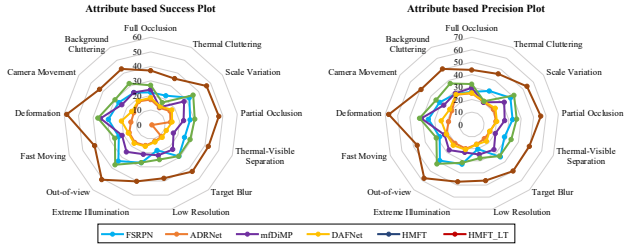

Figure 2. Attribute-based comparison on the long-term subset.

3. Mask Evaluation

We compare three short-term trackers (SiamMask [3], D3S [2] and AlphaRefine [5]) and two VOS methods (TVOS [6] and RANet [4]), shown as Tab. 2. All short-term trackers can output a mask for target with the initialization of bounding box. For VOS methods, we initialize the algorithms with the initial mask. VOS methods show inferior performance to short-term trackers with segmentation mask prediction. The reason is that short-term trackers predict the mask in a small search region, which avoid distractors with similar semantics, achieving better results.

4. Analysis on Different Alignment Methods

We apply image alignment for the initial frame of all the sequences. we compare the center distance among previ-

Table 2. Mask evaluation results on the proposed dataset.

Method	$\mathcal{J}$	$\mathcal{F}$	FPS
RANet [4]	0.322	0.454	5.0
TVOS [6]	0.369	0.510	21.7
SiamMask [3]	0.529	0.610	**57.2**
D3S [2]	0.534	0.607	18.4
AlphaRefine [5]	**0.599**	**0.719**	14.2

ous datasets (GTOT and RGBT234), shown in Tab. 3. We show that our alignment method is also effective for image registration. In the future, trackers can design an alignment module for better results.

Table 3. Comparison among previous datasets on center distance of bounding boxes in visual-thermal modalities.

	GTOT		RGBT234		VTUAV	
	Mean	Median	Mean	Medsian	Mean	Median
Dist.	2.74	2.81	5.19	4.54	10.99	8.83

5. Visualization Results on Long-term Subset

As shown in Fig. 3, we reveal the effectiveness of our HMFT trackers on long-term subset of VTUAV. Compared with existing long-term trackers, our tracker can handle all the challenging cases in various conditions, and show its great potential on re-detection ability.

6. Failure Cases

As shown in Fig. 4, we provide the limitation of HMFT tracker, HMFT often suffers target missing when the target is in heavy occlusion or full occlusion, leading to an inferior performance on fully occlusion attributes (which is shown in Figure 4 in the submission). When the target is occluded, our tracker has limited ability on re-detection. The reason is that the noisy samples cause biased filter learning in the online learning process, while Siamese-based methods in an offline manner get rid of this issue. Our long-term version (HMFT_LT) with a global tracker further improves our method in this challenge. In the future, development, such as the adaptive updating mechanism, can be accomplished to achieve better performance.

References

[1] Matej Kristan, Jiri Matas, Ales Leonardis, Michael Felsberg, and et al. The seventh visual object tracking VOT2019 challenge results. In *IEEE International Conference on Computer Vision Workshop*, pages 1–36, 2019. 1

Figure 3. Visualization of HMFT and HMFT_LT on long-term subset of VTUAV.

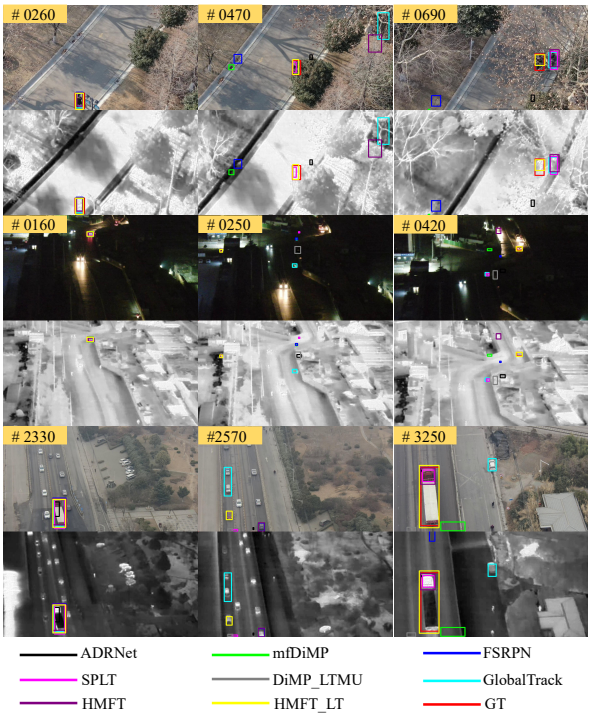

Figure 4. Failure analysis of the proposed HMFT.

[2] Alan Lukezic, Jiri Matas, and Matej Kristan. D3S-A discriminative single shot segmentation tracker. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 7133–

7142, 2020. 1, 2

[3] Qiang Wang, Li Zhang, Luca Bertinetto, Weiming Hu, and Philip HS Torr. Fast online object tracking and segmentation: A unifying approach. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1328–1338, 2019. 1, 2

[4] Ziqin Wang, Jun Xu, Li Liu, Fan Zhu, and Ling Shao. Ranet: Ranking attention network for fast video object segmentation. In *IEEE International Conference on Computer Vision*, pages 3978–3987, 2019. 1, 2

[5] Bin Yan, Xinyu Zhang, Dong Wang, Huchuan Lu, and Xiaoyun Yang. Alpha-refine: Boosting tracking performance by precise bounding box estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5289–5298, 2021. 1, 2

[6] Yizhuo Zhang, Zhirong Wu, Houwen Peng, and Stephen Lin. A transductive approach for video object segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 6949–6958, 2020. 1, 2