Template-based Uncertainty Multimodal Fusion Network for RGBT Tracking

Supplementary Material

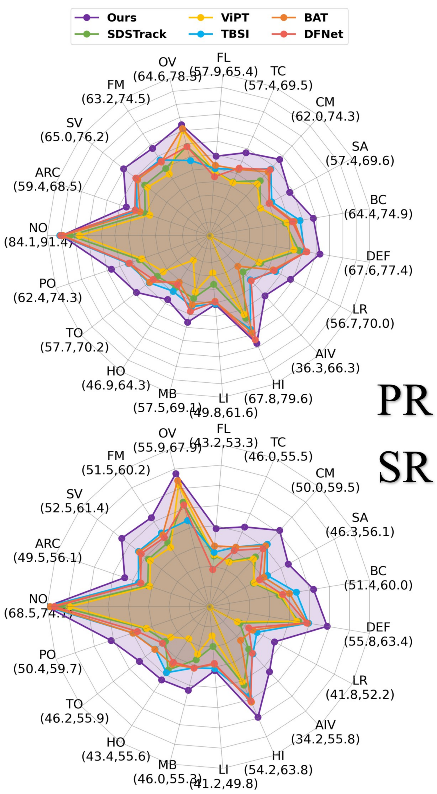

Figure 1: Precision Rate (PR) and Success Rate (SR) of challenge attributes on the LasHeR dataset.

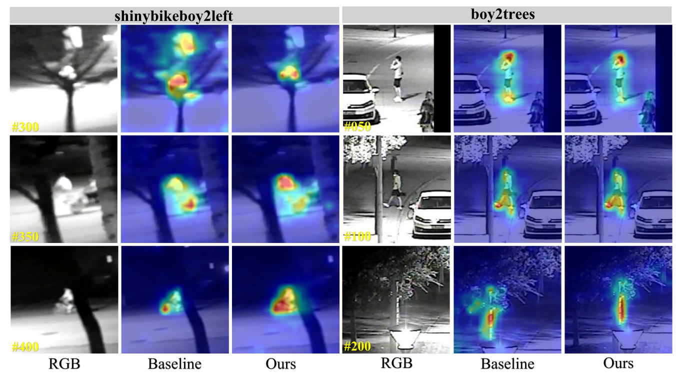

Figure 2: Comparison of attention maps between the baseline and our method on two sequences from the LasHeR dataset.

1 Supplementary Experiments

1.1 Comparisons With State-of-the-Art Methods

Evaluation on LasHeR dataset. The LasHeR dataset includes 19 additional attribute annotations, which are as follows: no occlusion (NO), partial occlusion (PO), total occlusion (TO), high occlusion (HO), out-of-view (OV), low illumination (LI), high illumination (HI), abrupt illumination variation (AIV), low resolution (LR), deformation (DEF), background clutter (BC), similar appearance (SA), thermal crossover (TC), motion blur (MB), camera movement (CM), fast motion (FM), scale variation (SV), and aspect ratio change (ARC). To validate the effectiveness of the proposed method, we present a comparison of experimental results with various state-of-the-art RGBT trackers on different attribute subsets, as shown in Figure 1. The results show that our method performs best in most of the challenge subsets, which demonstrates the great potential of the proposed method in various complex tracking scenarios.

Visualization of Attention Map. To validate the effectiveness of the proposed method, we visualize the attention maps of the baseline and our method, as shown in Figure 2. It can be observed that the baseline is more easily disturbed by the surrounding environment. For example, in the sequence *boy2trees*, when the target is occluded, the attention of the baseline becomes somewhat scattered, whereas our method's attention is more focused.

Table 1: Ablation study of loss weights on LasHeR dataset.

λ_t	λ_3	PR	NPR	SR
0.1	0.1	74.3	70.6	59.7
0.1	0.05	74.9	71.1	60.2
0.1	0.02	75.7	72.1	60.9
0.1	0.01	**76.4**	**72.3**	**61.4**
0.05	0.01	74.3	70.7	59.9
0.01	0.01	74.2	70.2	59.5

Inserting Layers	RGBT234		LasHeR		
	MPR	MSR	PR	NPR	SR
[4, 7, 10]	89.3	66.3	74.9	71.0	59.8
[10, 11, 12]	**90.8**	**67.8**	**76.4**	**72.7**	**61.4**

Table 2: Inserting layers of the proposed UMFM.

1.2 Ablation experiment on different weights of the loss function

To explore the impact of different loss weights on the model's performance, we conduct experiments with various weights, and the results are shown in Table 1. We conduct experiments by varying the weights of the two added loss functions, while keeping the baseline loss weight unchanged. It can be observed that the overall performance is optimal when the maximum value of λ_t is set to 0.1 and λ_3 is set to 0.01.

In addition, to further evaluate the effectiveness of inserting UMFM into the last three layers, we conduct experiments by inserting UMFM into the 4th, 7th, and 10th layers, similar to some previous works. The experimental results are shown in Table 2. When UMFM is inserted into the last three layers, the model achieves the best performance. We analyze that this is because the last three layers extract high-level semantic information, which facilitates modeling the uncertainty of the modalities and enables robust multimodal fusion.