

Supplementary Materials: Breaking Modality Gap in RGBT Tracking: Coupled Knowledge Distillation

Anonymous Authors

In the supplementary material we present the full version of the ablation study section. In addition, we add three experiments with chapter names marked in blue, including the impact of distillation layers, robustness performance, and tracking results visualization, thus to demonstrate the advantages of CKD in more detail.

0.1 Ablation Study

To verify the effectiveness of the proposed method, several ablation studies are performed on RGBT234 and LasHeR datasets.

Table 1: Ablation study on the main components of CKD.

	Pretrained model	RGBT234		LasHeR	
		PR	SR	PR	SR
baseline	SOT	86.4	64.5	67.8	54.0
w/ SD	SOT	86.4	65.0	68.9	54.5
w/ SD CD	SOT	87.4	65.5	71.6	56.9
w/ SD CD MM	SOT	88.6	66.1	72.3	57.4
w/ SD CD MM	DropMAE	90.4	67.8	73.1	58.0

0.1.1 Component Analysis. In Table 1, we conduct ablation studies on RGBT234 and LasHeR datasets to verify the effectiveness of different designed modules in CKD. Our baseline structure is the same as CKD, along with consistent training data and task losses, to fairly verify the effectiveness of the proposed components.

w/ SD denotes the baseline equipped with style distillation, which achieves a certain improvement, surpassing the baseline by 0.5% in SR on RGBT234, and 1.1%/0.5% in PR/SR on LasHeR. The experiment shows that aligning styles between modalities is effective, but there are limitations.

w/ SD CD indicates that adding content distillation to **w/ SD** results in significant performance improvements, achieving PR/SR improvements of 1.0% /1.0% over baseline on RGBT234 and achieving PR/SR improvements of 3.8%/2.9% over baseline on LasHeR. The experiment shows that keeping modality content representation stable is crucial in performing style distillation between different modalities. In other words, unconstrained style distillation could harm modality content representation, which could explain the limitations of **w/ SD**.

w/ SD CD MM represents that adding masked modeling to **w/ SD CD**, which further improves performance by 1.2%/0.6% in PR/SR on RGBT234, and 0.7%/0.5% in PR/SR on LasHeR. The experiment demonstrates the effectiveness of the masked modeling strategy. In addition, it can be observed that the mask modeling strategy that seamlessly integrates with style and content distillation improves by 2.2%/1.6% in PR/SR on RGBT234, and 4.2%/3.4% in PR/SR on LasHeR over baseline.

0.1.2 Impact of Pretrained model. We also explore the DropMAE [7] pretrained model trained on the Kinetics700 dataset [2] as our

Table 2: Ablation study on the different elimination scheme.

	RGBT234		LasHeR		MACs(G)	FPS
	PR	SR	PR	SR		
CKD _{slow}	90.4	67.8	73.1	58.0	57.802	84.8
w/ CE [10]	88.7	66.5	73.0	58.0	42.735	96.4
w/ MCE	90.0	67.4	73.2	58.1	42.735	96.4

pretrained model, which further achieves significant performance gains. Compared to the "SOT" pretrained model usually exploited by existing RGBT tracking methods [1, 3–6, 8, 11], DropMAE can bring superior performance to RGBT tracking. The experiment provides insights to further improve RGBT performance.

0.1.3 Effectiveness of token elimination strategy. To verify the effectiveness of the proposed multi-modal candidate token elimination (MCE) strategy, we evaluate different token elimination methods in Table 2. Here, CKD_{slow} represents the CKD method without a token elimination strategy, but it is still faster than existing RGBT trackers.

w/ CE [10] indicates that the two student branches individually apply the candidate token elimination strategy, following [10]. However, although CE brings an improvement in tracking efficiency, it also causes a significant performance drop.

w/ MCE indicates that the two student branches follow the proposed multi-modal candidate token elimination strategy for collaborative token elimination. It can be seen that MCE achieves a balance between tracking efficiency and accuracy. In particular, the introduction of MCE improves the tracking speed by 13.7% and reduces MACs by 35.3%, and also improves tracking performance on LasHeR. Although tracking performance on RGBT234 is slightly degraded, it is consistently better than existing RGBT trackers.

0.1.4 Hyper-parameter sensitivity analysis. We analyze the parameter sensitivity as follows.

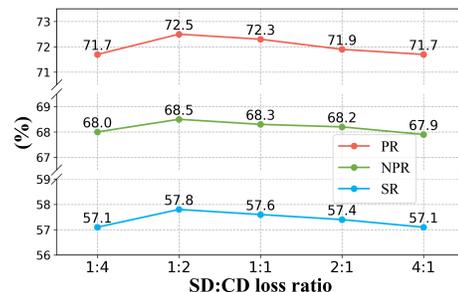


Figure 1: Ablation study of loss weights on LasHeR dataset.

Impact of loss weights. We explore the influence of different loss weights between style and distillation losses on CKD performance

in Figure 1. We set five different loss weight ratios, to explore RGBT tracker in the optimized orientation of two kinds of losses. From Figure 1 it can be observed that the two kinds of losses in CKD are robust to the hyperparameters for these weights. In addition, it can be found that appropriately increasing the importance of style loss in CKD brings better performance.

Table 3: Ablation study on different masked ratios.

	RGBT234		LasHeR		
	PR	SR	PR	NPR	SR
CKD w/ mask 0%	87.4	65.5	71.6	67.5	56.9
CKD w/ mask 25%	88.6	66.1	72.3	68.1	57.4
CKD w/ mask 50%	88.2	65.1	71.4	67.2	56.9
CKD w/ mask 75%	88.2	64.3	70.6	66.7	56.4

Impact of masked ratios. As shown in Table 3, we analyze the influence of different masked ratios on masked modeling strategy in CKD. It can be observed that the performance of CKD is always improved after the introduction of mask modeling, but the performance decreases slightly with the increase of mask modeling. The experiment shows that 25% mask ratio in CKD is best.

Table 4: Ablation study on the different distillation layers.

	RGBT234		LasHeR		
	PR	SR	PR	NPR	SR
last 1 layer	87.6	65.6	71.5	67.2	56.8
last 6 layers	88.2	65.5	71.9	67.6	57.1
all layers (12)	88.6	66.1	72.3	68.1	57.4

Impact of distillation layers. In Table 4, we also explore the impact of distillation with different number of layers on the performance of CKD. It can be observed that the performance of CKD is increases slightly with the increase of distillation layers. The experiment shows that the current setting of all layers distillation is optimal.

0.1.5 Analysis of feature decoupling scheme. In Table 5, we construct several variants to verify the effectiveness of feature decoupling scheme.

baseline w/ IN denotes the introduction of instance normalization in both student branches, which performs tracking with only content features. The scheme exhibits some performance degradation compared to baseline. It shows that modality style features certain discriminative information, which can lead to performance loss when directly dropped, and thus it is key to find common modality styles.

baseline w/ FD represents the introduction of non-decoupled feature distillation (FD) only between two student branches. The scheme shows some performance decrease compared to baseline, suggesting that the non-decoupled distillation scheme may harm the modal content representation, thus limiting performance.

baseline w/ SD is to perform distillation only in the style features between two student branches, which achieves higher performance compared to the non-decoupled distillation scheme. The experiment further verifies that performing distillation for all modal features

Table 5: Ablation study on the feature decoupling scheme.

	RGBT234		LasHeR		
	PR	SR	PR	NPR	SR
baseline	86.4	64.5	67.8	64.3	54.0
baseline w/ IN	85.6	63.7	67.1	63.2	53.4
baseline w/ FD	85.2	63.8	67.2	63.4	53.7
baseline w/ SD	86.4	65.0	68.9	64.3	54.5
baseline w/ CKD	87.4	65.5	71.6	67.5	56.9

is unnecessary, and also demonstrates that pursuing inter-modal feature style consistency can effectively mitigate modality gap.

baseline w/ CKD is the coupled distillation scheme proposed in this paper, which significantly improves the performance on both datasets, thus further demonstrating the importance of feature decoupling scheme. Moreover, in comparison with the style distillation scheme, it can be observed that even if only modal style features are distilled, it is difficult to completely avoid causing harm to the modal content representation, which leads to sub-optimal performance.

Table 6: Ablation study on the different missing ratios.

Missing challenge	Tracker	RGBT234		LasHeR		
		PR	SR	PR	NPR	SR
w/o RGB (50%)	baseline	80.3	58.0	59.0	54.2	46.9
	CKD	85.4	61.4	65.6	60.6	51.9
w/o TIR (50%)	baseline	84.1	62.6	63.2	59.4	50.5
	CKD	85.0	63.7	68.1	63.8	54.0
w/o RGB (80%)	baseline	73.5	52.2	53.0	48.3	42.3
	CKD	81.1	56.9	60.1	55.1	47.5
w/o TIR (80%)	baseline	81.8	60.9	60.0	56.3	48.1
	CKD	82.6	61.6	64.4	59.7	50.8

0.1.6 Robustness performance. In Table 6, we also explore the robustness of CKD with different modality miss ratios. We adopt a random missing strategy for the test dataset, and copy the non-missing modality data as the input compensation for missing modality. We construct two different miss ratios to more fully compare the performance of CKD and baseline methods when encountering the same miss scenario. It can be observed that the performance of baseline shows a significant decline while the performance of CKD is decreases slightly with the increase of missing ratios. The experiment shows that our method is robust even in the missing challenge.

In addition, it can be seen that there are significant differences in the importance of different modalities. For example, in the experiment with 80% missing ratio, the baseline method achieves 8.3% (PR) performance difference between different missing modalities in RGBT234, while the difference is only 1.5% (PR) in CKD. The experiment indicates that CKD effectively eliminates the gap between modalities, thus bridging the performance difference of different modalities.

0.1.7 Visual analysis. In this section, we design three different visualization strategies to show the advantage of CKD, including

feature visualization, tracking results visualization and token eliminate visualization.

Feature visualization. In Figure 2, we visualize and compare the RGB and TIR content features extracted by the models trained with different distillation techniques and the RGB and TIR content features extracted by the models not trained with distillation methods. We also show their similarity relationships in the second and four rows. Moreover, we present the distribution of the two modality features between different models. Although the feature distillation (FD) approach significantly narrows the inter-modal feature distance between two modalities, as shown in (o) and (r), it also leads to a perceptible difference between its feature content (g)/(n) and the original feature content (a)/(h), which is likely to harm the modality content representation. By adopting style distillation (SD), we can observe that SD achieves the preservation of feature content (c)/(j) while reducing the inter-modal gap (q). However, our visualization of the cosine similarity (f)/(m) between (c)/(j) and (a)/(h) reveals that they share a low similarity, suggesting that imperceptible variations in content features still occur in SD. Finally, the proposed CKD method achieves a good balance between inter-modal gap elimination and modality content representation preservation, as shown in (e)/(l) and (p).

Tracking results visualization. In Figure 3, we visualize the tracking results of CKD and other advanced RGBT trackers, including SDSTrack [4], TBSI [5], ViPT [11] and APFNet [9]. In four representative examples, it can be seen that our CKD method can achieve far better tracking performance than existing algorithms under different challenges, including (a) occlusion challenge, (b) transparent target challenge, (c) background clutters challenge, and (d) similar object interference challenge. The experiment demonstrates that our method can fully cooperative information of two modalities to handle challenging scenarios by breaking modality gap.

Token eliminate visualization. In Figure 4, we visualize the eliminate results of multi-modal candidate token eliminate (MCE) and original single-modal candidate token eliminate CE [10]. In first row, it can be observed that the CE strategy introduces inaccurate elimination results in target region of RGB modality, which may cause performance degradation. In contrast, MCE can provide correct and consistent elimination results in both modalities. Similar situation occurs in the fourth row. In the TIR modality, the CE strategy drops the target region's token, while the MCE keeps the target region's token.

REFERENCES

- [1] Bing Cao, Junliang Guo, Pengfei Zhu, and Qinghua Hu. 2024. Bi-directional Adapter for Multi-modal Tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*.
- [2] Joao Carreira, Eric Noland, Chloe Hillier, and Andrew Zisserman. 2019. A short note on the kinetics-700 human action dataset. *arXiv preprint arXiv:1907.06987* (2019).
- [3] Lingyi Hong, Shilin Yan, Renrui Zhang, Wanyun Li, Xinyu Zhou, Pinxue Guo, Kaixun Jiang, Yiting Chen, Jinglun Li, Zhaoyu Chen, et al. 2024. OneTracker: Unifying Visual Object Tracking with Foundation Models and Efficient Tuning. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*.
- [4] Xiaojun Hou, Jiazheng Xing, Yijie Qian, Yaowei Guo, Shuo Xin, Junhao Chen, Kai Tang, Mengmeng Wang, Zhengkai Jiang, Liang Liu, et al. 2024. SDSTrack: Self-Distillation Symmetric Adapter Learning for Multi-Modal Visual Object Tracking. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*.
- [5] Tianrui Hui, Zizheng Xun, Fengguang Peng, Junshi Huang, Xiaoming Wei, Xiaolin Wei, Jiao Dai, Jizhong Han, and Si Liu. 2023. Bridging Search Region Interaction

With Template for RGB-T Tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13630–13639.

- [6] Hongyu Wang, Xiaotao Liu, Yifan Li, Meng Sun, Dian Yuan, and Jing Liu. 2024. Temporal Adaptive RGBT Tracking with Modality Prompt. In *Proceedings of the AAAI Conference on Artificial Intelligence*.
- [7] Qiangqiang Wu, Tianyu Yang, Ziquan Liu, Baoyuan Wu, Ying Shan, and Antoni B Chan. 2023. Dropmae: Masked autoencoders with spatial-attention dropout for tracking tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14561–14571.
- [8] Zongwei Wu, Jilai Zheng, Xiangxuan Ren, Florin-Alexandru Vasluianu, Chao Ma, Danda Pani Paudel, Luc Van Gool, and Radu Timofte. 2024. Single-Model and Any-Modality for Video Object Tracking. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*.
- [9] Yun Xiao, Mengmeng Yang, Chenglong Li, Lei Liu, and Jin Tang. 2022. Attribute-based progressive fusion network for rgbt tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 2831–2838.
- [10] Botao Ye, Hong Chang, Bingpeng Ma, Shiguang Shan, and Xilin Chen. 2022. Joint feature learning and relation modeling for tracking: A one-stream framework. In *European Conference on Computer Vision*. Springer, 341–357.
- [11] Jiawen Zhu, Simiao Lai, Xin Chen, Dong Wang, and Huchuan Lu. 2023. Visual prompt multi-modal tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9516–9526.

291
292
293
294
295
296
297
298
299
300
301
302
303
304
305
306
307
308
309
310
311
312
313
314
315
316
317
318
319
320
321
322
323
324
325
326
327
328
329
330
331
332
333
334
335
336
337
338
339
340
341
342
343
344
345
346
347
348

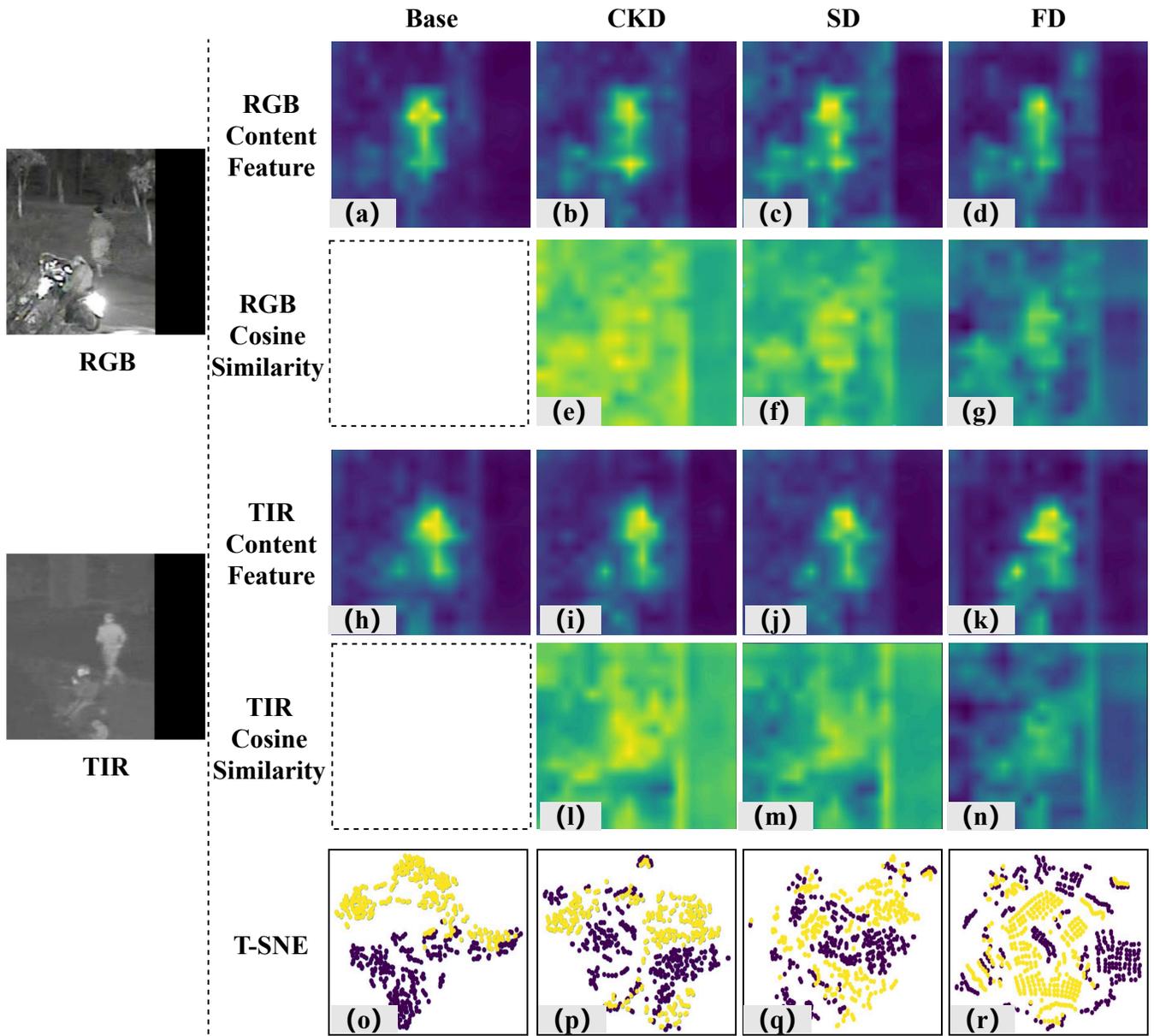
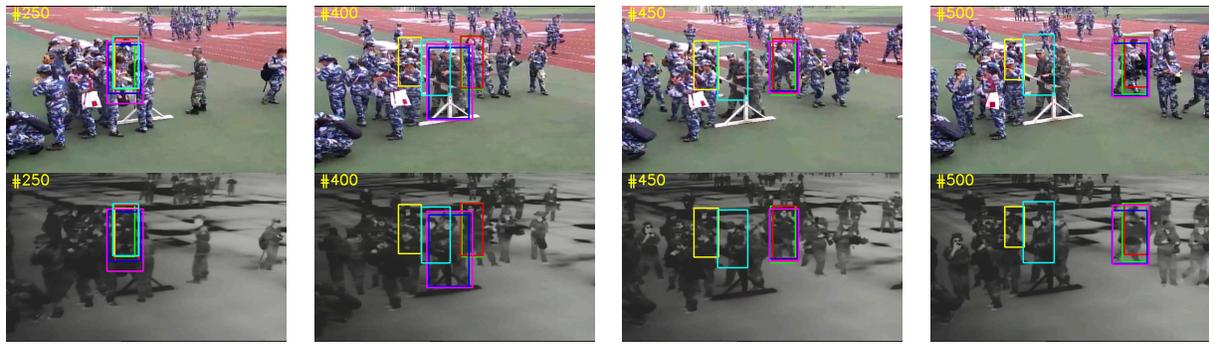
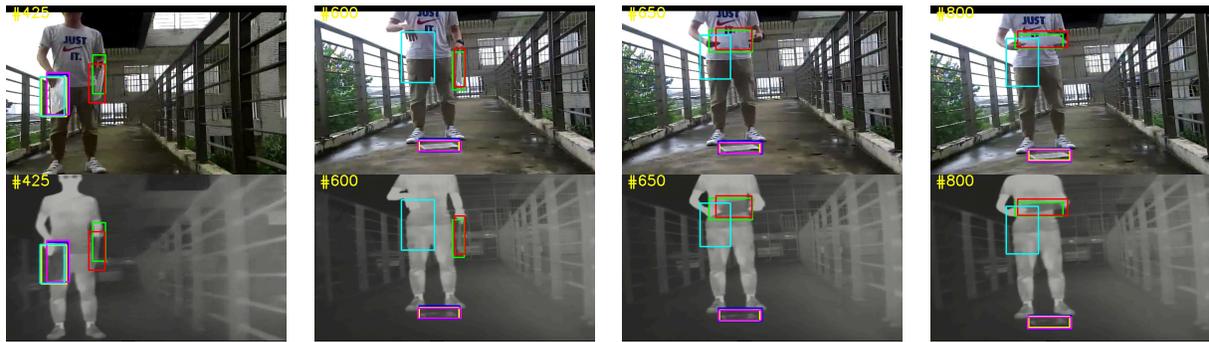


Figure 2: Comparison of feature maps and T-SNE visualizations for different distillation methods. For T-SNE maps, they have the same scale of axes. The hotter color in the first and third rows indicates more salient features, while in the second and four rows the hotter color indicates more similar between the non-distilled (Base features) and distilled features, and vice versa. In the five row, the yellow and purple color indicate the features of RGB and TIR modalities respectively.



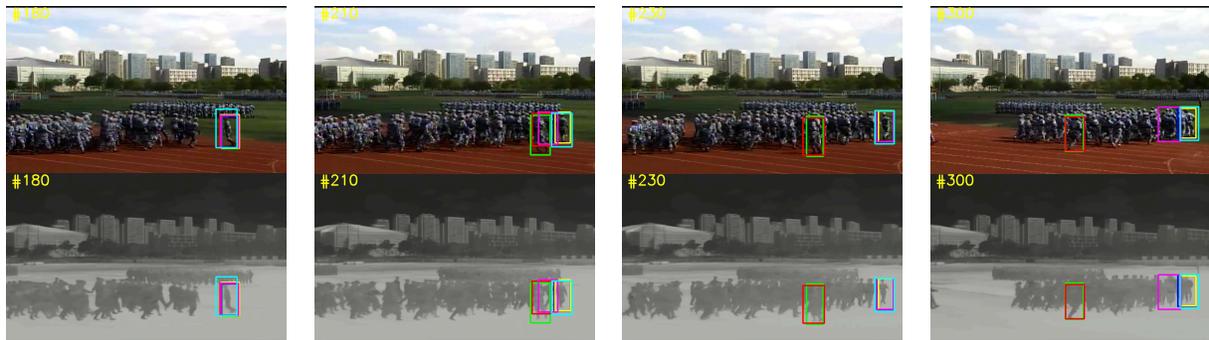
(a) blkhairgirltakingblkbag



(b) foldedfolderatlefthand



(c) girlrightthewautress



(d) leftboyoutofthetroop

GT
 CKD
 SDSTrack
 TBSI
 ViPT
 APFNet

Figure 3: Comparison of tracking results for different RGBT tracking methods.

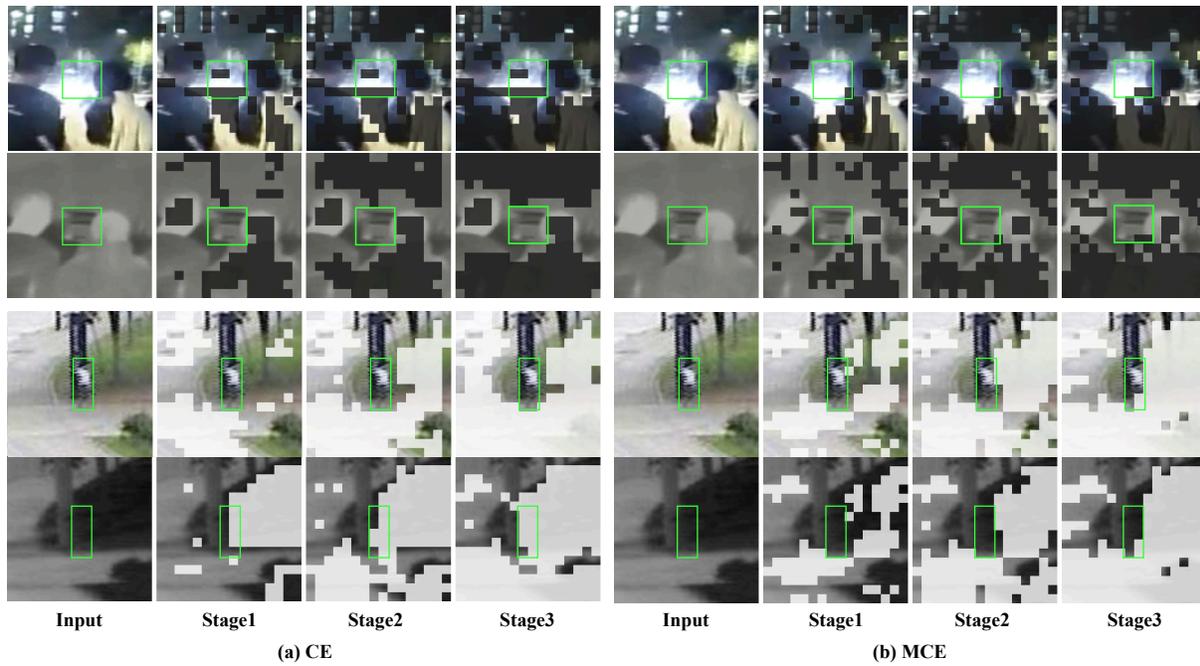


Figure 4: Comparison of different candidate token eliminate methods. We choose two representative examples *carcominginlight* in the first two rows and *child1* in the third four rows. Here, we adopt a black-gray mask in the first sequence to clearly show the discarded tokens in the RGB modality, and a white-gray mask in the second sequence to more clearly show the discarded tokens in the thermal modality.