Progressive Multi-cue Alignment for Unaligned RGBT Tracking

Supplementary Material

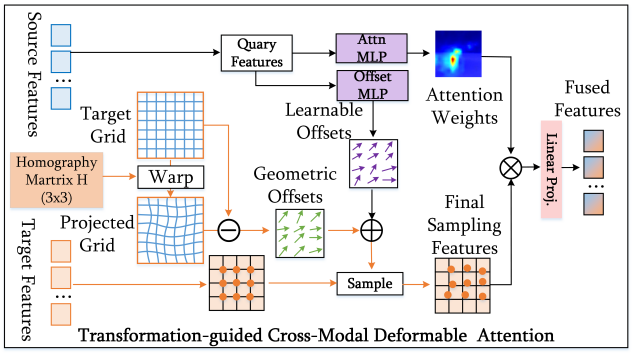


Figure 8. Overview of the proposed TCMDA.

7. Additional Experimental Results

Overall Performance on Our MUART244. To comprehensively evaluate our MUART dataset, we select 10 state-of-the-art trackers, including SDSTrack [9], ViPT [54], TBSI [13], BAT [1], AINet [30], Un-Track [41], AFter [29], OSTrack [45], CAFormer [44], and SUTrack [3]. We first train all methods on the LasHeR-Unaligned training set and then uniformly evaluate their performance on MUART244. As shown in Fig. 10, existing RGBT trackers experience a significant performance drop when confronted with large-scale cross-modal misalignment. In particular, although CAFormer fuses multimodal information through modality modulation, the severe misalignment causes modal attention to mismatch and the fusion mechanism to fail, preventing the tracker from robustly following the target. In contrast, AINet aggregates multi-level visual features and performs differential interactions between hidden features across modalities, enabling it to achieve higher accuracy than other RGBT trackers. Our proposed method progressively eliminates spatial misalignment through a multi-stage alignment strategy and addresses severe offset cases via an offset-contrastive updating mechanism, which alleviates the issue of infrared sampling missing the target under large displacement. As illustrated in Fig. x, our approach achieves 62.7%/55.9%/45.8% in PR, NPR, and SR, respectively. Compared with SUTrack, our method improves PR/NPR/SR by 13.2%/15.0%/12.3%. Relative to AINet, the improvements reach 5.4%/5.5%/4.7% on the same metrics.

Effect of TCMDA. The detail pipeline of TCMDA is shown in Fig. 8. As discussed in Sec. 3.4, Shi et al. demonstrated that directly aligning features or pixels may lead to suboptimal results in certain scenarios. To address this issue, we introduce the TCMDA module, which uses the predicted offsets as alignment priors to perturb the initial sampling grid, thereby reducing the noise caused by feature deformation (as illustrated in Fig. 9). Furthermore, as shown in Table 6, progressively introducing geometric alignment and deformable attention significantly enhances the fusion performance. Compared with the Baseline, directly applying CDA fails to learn the complex cross-modal offset relationships effectively and thus yields limited improvement. However, when combined with the predicted feature warping, the performance increases substantially. The final TCMDA achieves gains of +2.9%/+2.3%/+2.1% on PR, NPR, and SR, respectively, demonstrating that the integration of explicit transformation guidance and adaptive sampling provides the most effective fusion strategy.

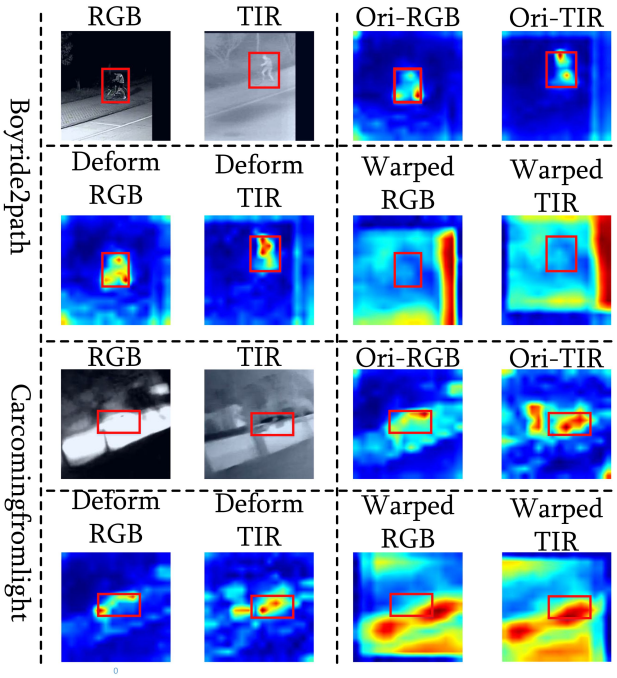


Figure 9. This figure visualizes the features from the final layer of the model. Ori denotes the baseline, Warped represents the features shifted using the directly predicted offsets [20], and Deform corresponds to our TCMDA.

Effect of Our Alignment Strategy. As discussed in Sec. 4, to demonstrate the effectiveness of our alignment strategy, we compare it with several mainstream alignment methods [20, 39, 48] in Tab. 7. First, compared with the IMF [39], our approach exhibits stronger alignment robustness and achieves significant improvements on both datasets. Relative to GFNet [48], which per-

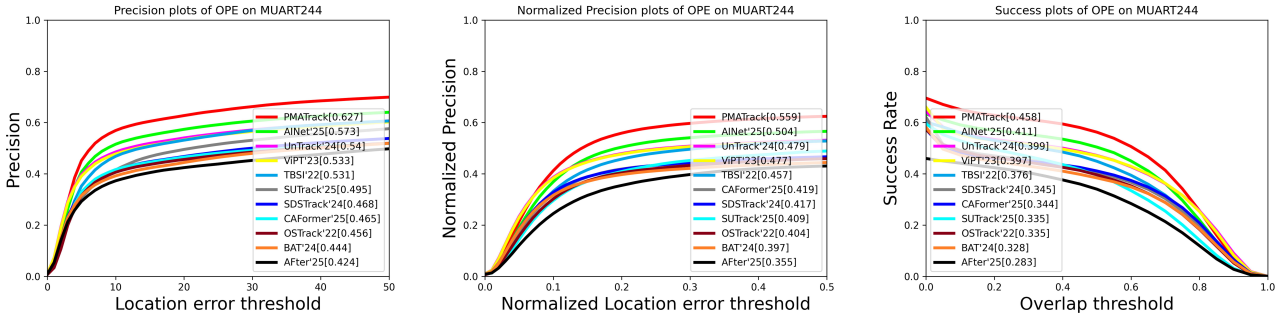


Figure 10. Evaluation results of precision, normalized rrecision , and success rate on the MUART244 dataset, with representative scores provided in the legend.

Table 6. Comparison of different fusion strategies. CDA denotes Deformable Attention [55] across modality, while Warp applies the predicted alignment matrix to warp features [20] prior to further processing.

Fusion Strategy	LasHeR-Unaligned		
	PR	NPR	SR
Baseline	61.5	56.4	48.5
Baseline-CDA	61.7	56.3	48.4
Ours-Warp&CDA	62.5	57.1	49.1
Ours-TCMDA	64.4	58.7	50.6

forms alignment matching using the DiNOv2-Large model, our method improves PR/NPR/SR by +2.2%/+1.9%/+1.7% and +4.0%/+4.8%/+3.7% on the two datasets, respectively. Moreover, our difficulty-aware multi-cue expert mechanism also provides a substantial speed advantage, increasing the FPS by +20.4 compared with GFNet+Baseline. These results validate the robustness of our progressive alignment strategy as well as the efficiency–accuracy balance achieved by DMAE.

We constructed stronger baselines, AINet + Homography Regression and AINet + STN, and reported the corresponding results 8. The comparison shows that augmenting AINet with a single geometric regression/warping module still falls short of our overall framework based on progressive decomposition, difficulty-aware routing, and TCMDA-guided fusion.

8. Overview of Our MUART244 Dataset

As described in Sec. 4.1, we provide a detailed discussion of the dataset construction process along with an in-depth analysis.

8.1. Video Collection

Our UAV sequences were collected by professional drone pilots operating advanced aerial platforms such as the DJI Matrice 300 RTK and DJI Mavic 3T, both equipped with Zenmuse H20T thermal-visible sensor modules, to capture RGBT videos across diverse locations and scenes. The ground-view data were acquired using a FLIR SC620 system equipped with paired CCD and thermal infrared cameras. Due to parallax and differences in the fields of view between the two cameras, the same object captured in paired images inevitably exhibits positional offsets and scale discrepancies, as illustrated in Fig. 13. Consequently, the misalignment in MUART244 originates from real-world camera imaging conditions and faithfully reflects practical application scenarios. MUART244 also avoids labour-intensive preprocessing steps such as manual cropping and rescaling.

We preserve the original resolutions of the dual sensors: the RGB images range from 1600×1200 to 3840×2160, while the infrared images range from 640×512 to 1280×1024. These settings reflect real-world ground-camera and UAV configurations and introduce critical challenges for cross-resolution feature learning and alignment. In total, we obtain MUART244, which consists of 244 video pairs and 205,000 RGBT image pairs, with an average of 844 frames per video, as summarised in Table 9.

8.2. Annotation

The MUART244 dataset was annotated at the frame level by 20 professional annotators and subsequently underwent multi-stage verification by an additional 5 experts to ensure annotation quality. Owing to the inherent misalignment and resolution differences between modalities, each modality was annotated and inspected independently with meticulous care. In addition, sequence-level annotations include 22 challenge attributes, integrating 15 common attributes from previous datasets and 7 newly introduced attributes specific to our dataset, as illustrated in Tab. 10 and Fig. 14.

8.3. Statistical Analysis of Our MUART244 Dataset.

Diverse Object Classes and platform Our MUART244 dataset offers diverse object categories and multi-platform

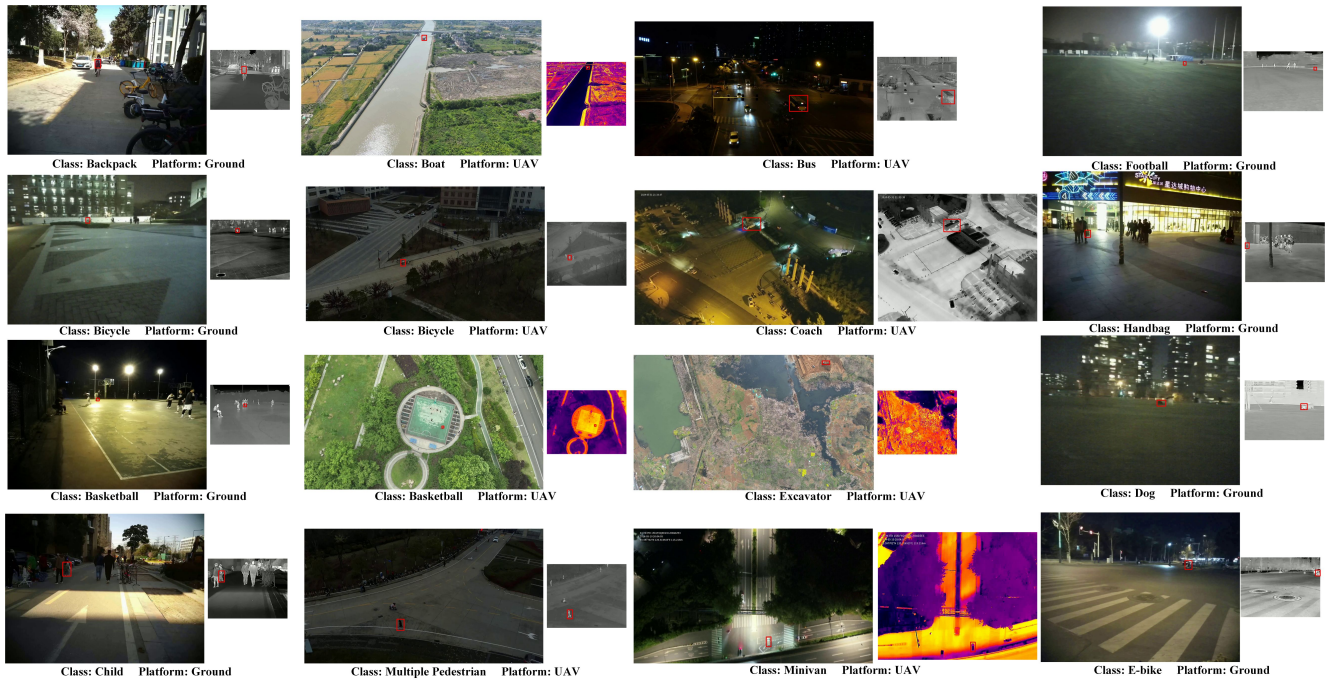
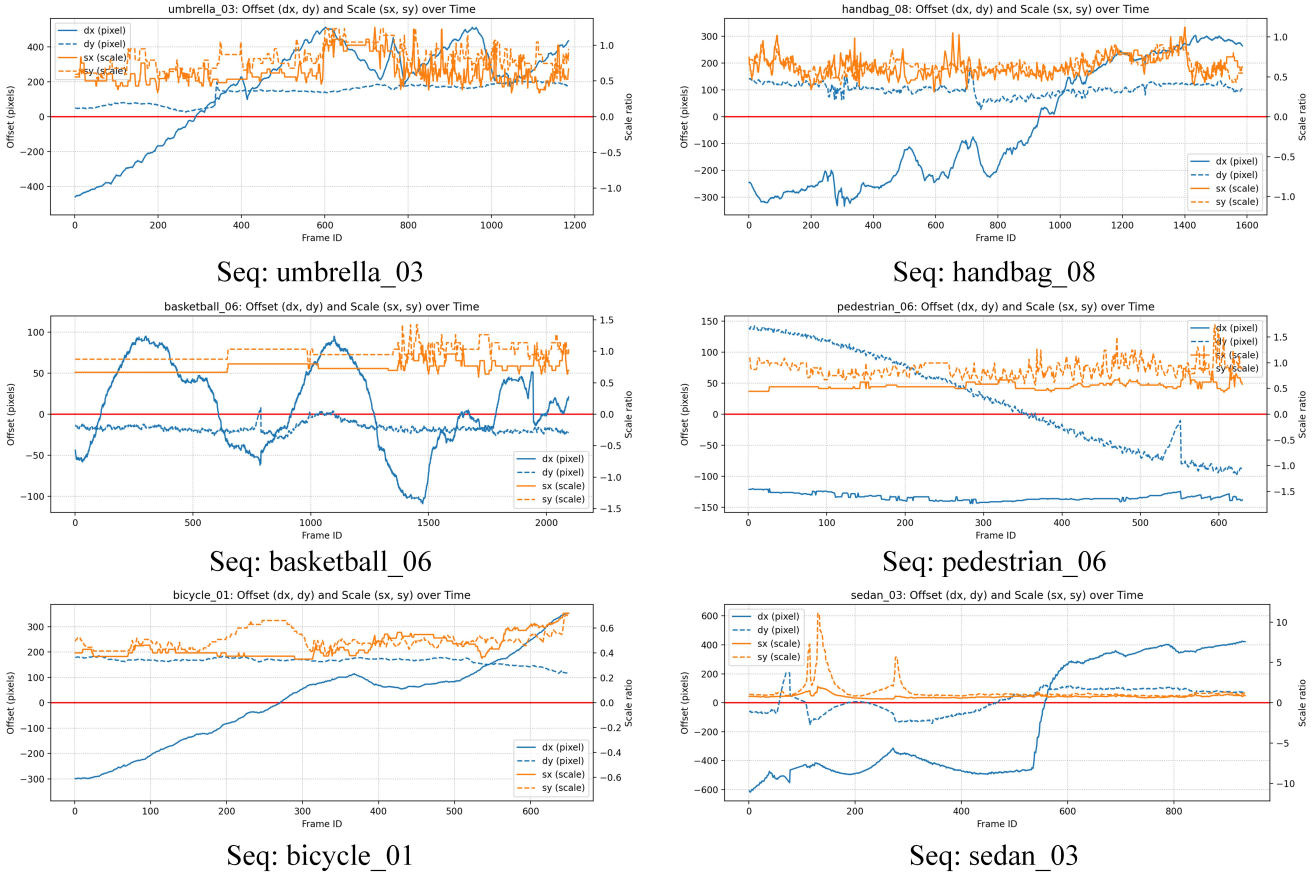


Figure 11. Visualization of selected target types and scenes from the dataset.



Seq: umbrella_03 Seq: handbag_08

Seq: basketball_06 Seq: pedestrian_06

Seq: bicycle_01 Seq: sedan_03

Figure 12. Temporal variation of center offsets (d_x, d_y) and scale differences (s_x, s_y) for several sequences in the dataset.

Table 7. Comparison of different multi-modal image alignment model on LasHeR-Unaligned dataset and MUART244 dataset.

Alignmet Model	Pub. Info.	LasHeR-Unaligned [19]			MUART			FPS↑
		PR↑	NPR↑	SR↑	PR↑	NPR↑	SR↑	
Baseline [40]	CVPR'23	61.5	56.4	48.5	57.3	50.4	41.1	**44.4**
IMF-Reg [39]	TCSVT'24	61.6	56.5	48.5	57.5	50.1	41.2	24.2
BSAFusion-Reg [20]	AAAI'25	61.8	56.5	48.6	58.1	50.7	41.7	25.7
GFNet [48]	CVPR'25	62.2	56.8	48.9	58.7	51.1	42.1	7.6
Ours	-	**64.4**	**58.7**	**50.6**	**62.7**	**55.9**	**45.8**	28.0

Table 8. Comparison on MUART244 (PR/NPR/SR, %).

Method	PR	NPR	SR
Ours	62.7	55.9	45.8
Ours Baseline	58.7	51.1	42.1
AINet [30]	57.3	50.4	41.1
AINet + Homography	59.6	52.7	42.9
AINet + STN	59.5	52.3	42.6

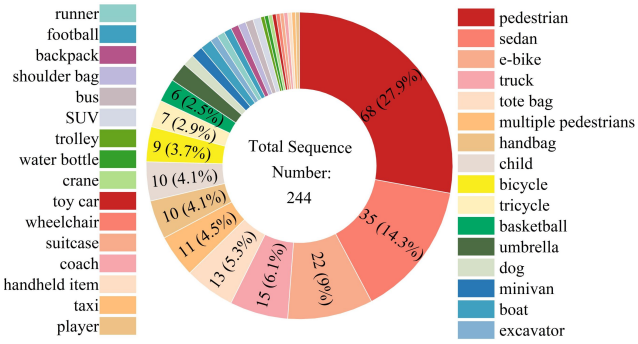


Figure 13. Visualization of the distribution of the 26 target categories in our MUART244 dataset.

perspectives, substantially enhancing current benchmarking standards. As shown in the Fig. 13 and Fig. 11, it contains 26 object categories and exhibits pronounced differences between ground and UAV viewpoints, covering a wide range of environments and target scales.

Offset variations in the dataset. As shown in Fig. 2, we visualize the temporal evolution of the four offset parameters for several sequences in our dataset. Since the only available registration parameters are the initial-frame bounding boxes of the two modalities, the 3×3 homography matrix can be reduced to four parameters $[d_x, d_y, s_x, s_y]$. We observe that in some sequences, for example, *umbrell_03*, the initial d_x is around -400, but as the target continues moving to approximately frame 600, d_x flips to +400. This confirms our claim in Sec. 1 that relying solely on the initial registration parameters is insufficient for accurately aligning subsequent frames. Such behaviour is not limited to $[dx, d_y]$. In the *pedestrian_06* sequence, it exhibits a similar reversal, often caused by fast target motion or the target crossing the image centre. In sequences such as *sedan_03*, the scale parameters also undergo large fluctuations, typically due to significant camera motion.

Moreover, many sequences display center offsets that evolve in the opposite direction of the initial offset, which is a key reason why using the initial offset for alignment leads to performance degradation in Sec. 4.5. Therefore, when the target exhibits large motion, high speed, or when the camera undergoes noticeable movement, the initial offset is insufficient for accurately aligning subsequent frames, making online offset correction essential.

Highly variable scales and area ratios. As shown in Fig. 15, we visualize the performance of several trackers under different ranges of average center offsets and average target area ratios, where the offset and area-ratio intervals are partitioned following the scheme in Fig. 1. The compared trackers include BAT, SUTrack, CAFormer, and AINet. We observe that across the vast majority of offset and scale conditions, our tracker consistently outperforms existing methods, demonstrating that our divide-and-conquer progressive alignment strategy effectively decouples cross-modal alignment parameters and achieves both efficient and robust alignment. In certain extreme cases, such as when the offset becomes very large, our method, which relies more heavily on multimodal information during training, may encounter insufficient offset correction under severe misalignment. This may cause the infrared modality to fail in registration and introduce additional noise.

References

[1] Bing Cao, Junliang Guo, Pengfei Zhu, and Qinghua Hu. Bidirectional adapter for multimodal tracking. In *Proceedings of the AAAI conference on artificial intelligence*, pages 927–935, 2024. 5, 6, 1

[2] Junren Chen, Rui Chen, Wei Wang, Junlong Cheng, Lei Zhang, and Liangyin Chen. Tinyu-net: Lighter yet better u-net with cascaded multi-receptive fields. In *International Conference on Medical Image Computing and Computer-*

Table 9. Statistics comparison among existing RGBT tracking datasets.

Benchmark	Train. Seq.	Test. Seq.	Avg. Frame	Resolution	Class Number	Challenge Number	Multi Platform	Modality Unaligned	Year
GTOT [15]	-	50	157	384×288	9	7	×	×	2016
RGBT210 [16]	-	210	498	630×460	22	12	×	×	2017
RGBT234 [17]	-	234	498	630×460	22	12	×	×	2019
LasHeR [19]	979	245	600	630×480	32	19	×	×	2021
VTUAV [50]	250	250	3329	1920×1080	13	13	×	×	2022
HiAl [43]	100	50	-	1920×1080	9	12	×	×	2023
LasHeR Unaligned [19]	748	205	694	631×481&631×481 960×576&960×576	32	19	×	✓	2021
MUART244 (Ours)	-	244	844	1600×1200&640×480 3840×2160&1280×1024	26	22	✓	✓	2025

Table 10. List and description of 22 attributes in MUART244.

Attribute	Definition
NO	No Occlusion - Target is not occluded.
PO	Partial Occlusion - Target is partially occluded
TO	Total Occlusion - Target is totally occluded
OV	Out-of-View - Target leaves view of camera
LI	Low Illumination - The illumination is low
HI	High Illumination - The illumination is high
AIV	Abrupt Illumination Variation - The illumination of the target changes significantly.
LR	Low Resolution - The resolution in the target region is low.
BC	Background Clutter - Background information is messy
SA	Similar Appearance - Other objects with similar appearance around target
TC	Thermal Crossover - Similar temperature between target and background
CM	Camera Moving - The target object is captured by moving the camera
FM	Fast Motion - Motion between adjacent frames is larger than 20 pixels
SV	Scale Variation - Ratio of current bounding box is out of range [0.5,2] compared to first frame
ARC	Aspect Ratio Change - Ratio of current bounding box aspect is out of range [0.5,2] compared to first frame
TOV	TIR Modality Target Out-of-View - TIR Modality Target leaves view of camera
HM	Horizontal Motion - UAV moves in the horizon direction
VM	Vertical Motion - UAV moves in the vertical direction
RM	Rotation Movement - UAV rotation results in lens rotation
PAC	Pitch Angle Change - Change in angle of airbrone lens relative to horizontal
FG	Foggy - Foggy Weather
CZ	Camera Zoom - Focal length of sensors change

Assisted Intervention, pages 626–635. Springer, 2024. 4

[3] Xin Chen, Ben Kang, Wanting Geng, Jiawen Zhu, Yi Liu, Dong Wang, and Huchuan Lu. Sutrack: Towards simple and unified single object tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2239–2247, 2025. 1, 6

[4] Yifei Deng, Guohao Wang, Chenglong Li, Wei Wang, Cheng Zhang, and Jin Tang. Collaborative license plate recognition via association enhancement network with auxiliary learning and a unified benchmark. *IEEE Transactions on Multimedia*, 2024. 1

[5] Yifei Deng, Chenglong Li, Zhengyu Chen, Zihen Xu, and Jin Tang. Decoupled cross-modal alignment network for text-rgbt person retrieval and a high-quality benchmark. *Information Fusion*, page 103948, 2025. 1

[6] Yifei Deng, Chenglong Li, Futian Wang, and Jin Tang. Learning hierarchical cross-modal association with intra-modal context for text-image person retrieval. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 2723–2731, 2025. 1

[7] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3

[8] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *First conference on language modeling*, 2024. 3

[9] Xiaojun Hou, Jiazheng Xing, Yijie Qian, Yaowei Guo, Shuo Xin, Junhao Chen, Kai Tang, Mengmeng Wang, Zhengkai

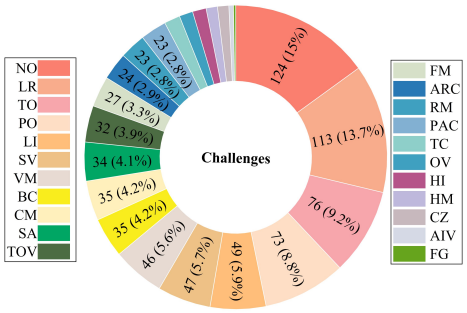


Figure 14. Visualization of the distribution of the 22 challenge factors in the dataset

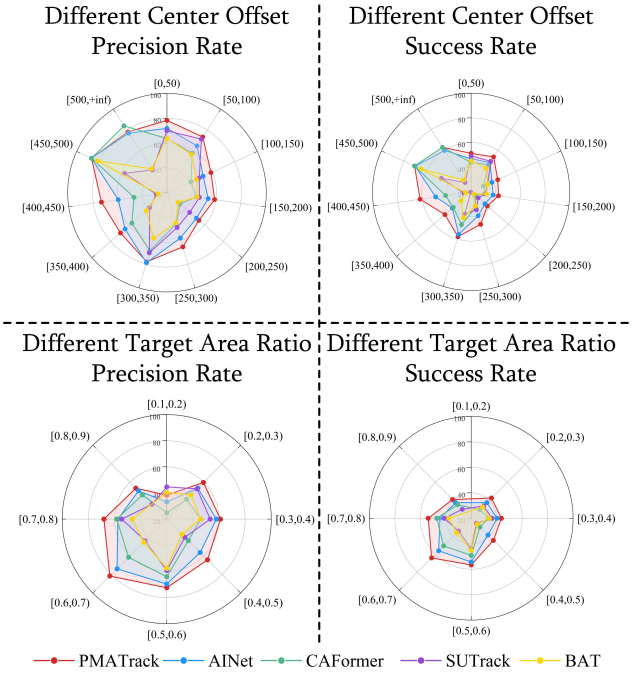


Figure 15. Performance comparison with state-of-the-art methods on subsets with different center offsets and target scales.

Jiang, Liang Liu, et al. Sdstrack: Self-distillation symmetric adapter learning for multi-modal visual object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26551–26561, 2024. 6, 1

[10] Xiantao Hu, Bineng Zhong, Qihua Liang, Shengping Zhang, Ning Li, and Xianxian Li. Toward modalities correlation for rgb-t tracking. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(10):9102–9111, 2024. 1

[11] Xiantao Hu, Ying Tai, Xu Zhao, Chen Zhao, Zhenyu Zhang, Jun Li, Bineng Zhong, and Jian Yang. Exploiting multimodal spatial-temporal patterns for video object tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3581–3589, 2025. 1

[12] Xiantao Hu, Bineng Zhong, Qihua Liang, Liangtao Shi, Zhiyi Mo, Ying Tai, and Jian Yang. Adaptive perception for unified visual multi-modal object tracking. *IEEE Transactions on Artificial Intelligence*, 2025. 1

[13] Tianrui Hui, Zizheng Xun, Fengguang Peng, Junshi Huang, Xiaoming Wei, Xiaolin Wei, Jiao Dai, Jizhong Han, and Si Liu. Bridging search region interaction with template for rgb-t tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13630–13639, 2023. 5, 6, 1

[14] Bo Li, Fengguang Peng, Tianrui Hui, Xiaoming Wei, Xiaolin Wei, Lijun Zhang, Hang Shi, and Si Liu. Rgb-t tracking with template-bridged search interaction and target-preserved template updating. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 5

[15] Chenglong Li, Hui Cheng, Shiyi Hu, Xiaobai Liu, Jin Tang, and Liang Lin. Learning collaborative sparse representation for grayscale-thermal tracking. *IEEE Transactions on Image Processing*, 25(12):5743–5756, 2016. 5

[16] Chenglong Li, Nan Zhao, Yijuan Lu, Chengli Zhu, and Jin Tang. Weighted sparse representation regularized graph learning for rgb-t object tracking. In *Proceedings of the 25th ACM International Conference on Multimedia*, pages 1856–1864, 2017. 5

[17] 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

[18] 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. 6

[19] Chenglong Li, Wanlin Xue, Yaqing Jia, Zhichen Qu, Bin Luo, Jin Tang, and Dengdi Sun. Lasher: A large-scale high-diversity benchmark for rgbt tracking. *IEEE Transactions on Image Processing*, 31:392–404, 2021. 2, 6, 7, 4, 5

[20] Huafeng Li, Dayong Su, Qing Cai, and Yafei Zhang. Bsafusion: A bidirectional stepwise feature alignment network for unaligned medical image fusion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4725–4733, 2025. 1, 2, 4

[21] Kunchi Li, Hongyang Chen, Jun Wan, and Shan Yu. Ckdf-v2: effectively alleviating representation shift for continual learning with small memory. *IEEE Transactions on Neural Networks and Learning Systems*, 2025. 2

[22] Kunchi Li, Chaoyue Ding, Jun Wan, and Shan Yu. Enhance the old representations' adaptability dynamically for exemplar-free continual learning. *Neurocomputing*, 639: 130286, 2025. 2

[23] Lei Liu, Chenglong Li, Yun Xiao, and Jin Tang. Quality-aware rgbt tracking via supervised reliability learning and weighted residual guidance. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 3129–3137, 2023. 3

[24] Lei Liu, Chenglong Li, Aihua Zheng, Jin Tang, and Yanping Xiang. Non-aligned rgbt tracking via joint temporal-iterated homography estimation and multimodal transformer fusion. In *Computer and Information Science and Engineering: Volume 16*, pages 17–32. Springer, 2024. 1, 3, 6

[25] Cheng Long Li, Andong Lu, Ai Hua Zheng, Zhengzheng Tu, and Jin Tang. Multi-adapter rgbt tracking. In *Proceedings of*

the IEEE/CVF international conference on computer vision workshops, pages 0–0, 2019. 6

[26] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

[27] Andong Lu, Chenglong Li, Yuqing Yan, Jin Tang, and Bin Luo. Rgbt tracking via multi-adapter network with hierarchical divergence loss. *IEEE Transactions on Image Processing*, 30:5613–5625, 2021. 6

[28] Andong Lu, Cun Qian, Chenglong Li, Jin Tang, and Liang Wang. Duality-gated mutual condition network for rgbt tracking. *IEEE Transactions on Neural Networks and Learning Systems*, 36(3):4118–4131, 2022. 6

[29] Andong Lu, Wanyu Wang, Chenglong Li, Jin Tang, and Bin Luo. After: Attention-based fusion router for rgbt tracking. *IEEE Transactions on Image Processing*, 2025. 6, 1

[30] Andong Lu, Wanyu Wang, Chenglong Li, Jin Tang, and Bin Luo. Rgbt tracking via all-layer multimodal interactions with progressive fusion mamba. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5793–5801, 2025. 3, 6, 1, 4

[31] Christoph Mayer, Martin Danelljan, Goutam Bhat, Matthieu Paul, Danda Pani Paudel, Fisher Yu, and Luc Van Gool. Transforming model prediction for tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8731–8740, 2022. 6

[32] Qiushi Nie, Xiaoqing Zhang, Chuan Chen, Zhixuan Zhang, Yan Hu, and Jiang Liu. Reparameterized multi-scale transformer for deformable retinal image registration. *Machine Intelligence Research*, 22(3):524–538, 2025. 3

[33] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 3

[34] Song-Liang Pan, Kunchi Li, Da-Han Wang, Xu-Yao Zhang, Jiantao Liu, and Shunzhi Zhu. Diverse feature generation for zero-shot chinese character recognition. *Expert Systems with Applications*, page 129442, 2025. 2

[35] Shuwei Shi, Jinjin Gu, Liangbin Xie, Xintao Wang, Yujiu Yang, and Chao Dong. Rethinking alignment in video super-resolution transformers. *Advances in Neural Information Processing Systems*, 35:36081–36093, 2022. 5

[36] Xinrui Song, Xuanang Xu, and Pingkun Yan. Dino-reg: General purpose image encoder for training-free multi-modal deformable medical image registration. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 608–617. Springer, 2024. 3

[37] Linfeng Tang, Qinglong Yan, Xinyu Xiang, Leyuan Fang, and Jiayi Ma. C2rf: Bridging multi-modal image registration and fusion via commonality mining and contrastive learning. *International Journal of Computer Vision*, pages 1–19, 2025. 3

[38] Zhangyong Tang, Tianyang Xu, Xiaojun Wu, Xue-Feng Zhu, and Josef Kittler. Generative-based fusion mechanism for multi-modal tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5189–5197, 2024. 6

[39] Di Wang, Jinyuan Liu, Long Ma, Risheng Liu, and Xin Fan. Improving misaligned multi-modality image fusion with one-stage progressive dense registration. *IEEE Transactions on Circuits and Systems for Video Technology*, 34 (11):10944–10958, 2024. 1, 4

[40] Qiangqiang Wu, Tianyu Yang, Ziquan Liu, Baoyuan Wu, Ying Shan, and Antoni B Chan. Dropmae: Masked autoencoders with spatial-attention dropout for tracking tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14561–14571, 2023. 6, 4

[41] Zongwei Wu, Jilai Zheng, Xiangxuan Ren, Florin-Alexandru Vasluianu, Chao Ma, Danda Pani Paudel, Luc Van Gool, and Radu Timofte. Single-model and any-modality for video object tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19156–19166, 2024. 6, 7, 1

[42] Yun Xiao, Mengmeng Yang, Chenglong Li, Lei Liu, and Jin Tang. Attribute-based progressive fusion network for rgbt tracking. In *Proceedings of the AAAI conference on artificial intelligence*, pages 2831–2838, 2022. 6

[43] Yun Xiao, Dan Cao, Chenglong Li, Bo Jiang, and Jin Tang. A benchmark dataset for high-altitude uav multi-modal tracking. *Journal of Image and Graphics*, 30:361–374, 2025. 5

[44] Yun Xiao, Jiacong Zhao, Andong Lu, Chenglong Li, Bing Yin, Yin Lin, and Cong Liu. Cross-modulated attention transformer for rgbt tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8682–8690, 2025. 1, 6

[45] Botao Ye, Hong Chang, Bingpeng Ma, Shiguang Shan, and Xilin Chen. Joint feature learning and relation modeling for tracking: A one-stream framework. In *European conference on computer vision*, pages 341–357. Springer, 2022. 5, 6, 1

[46] Junchen Yu, Si-Yuan Cao, Runmin Zhang, Chenghao Zhang, Zhu Yu, Shujie Chen, Bailin Yang, and Hui-Liang Shen. Sshnet: Unsupervised cross-modal homography estimation via problem reformulation and split optimization. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 16685–16694, 2025. 3

[47] Hui Zhang, Lei Zhang, Li Zhuo, and Jing Zhang. Object tracking in rgb-t videos using modal-aware attention network and competitive learning. *Sensors*, 20(2):393, 2020. 6

[48] Kaining Zhang, Yuxin Deng, Jiayi Ma, and Paolo Favaro. Adapting dense matching for homography estimation with grid-based acceleration. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 6294–6303, 2025. 3, 1, 4

[49] Pengyu Zhang, Dong Wang, Huchuan Lu, and Xiaoyun Yang. Learning adaptive attribute-driven representation for real-time rgb-t tracking. *International Journal of Computer Vision*, 129(9):2714–2729, 2021. 6

[50] Pengyu Zhang, Jie Zhao, Dong Wang, Huchuan Lu, and Xiang Ruan. Visible-thermal uav tracking: A large-scale benchmark and new baseline. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8886–8895, 2022. 1, 5

[51] Pengyu Zhang, Simiao Lai, Dong Wang, and Huchuan Lu.

Motion-guided visual tracking. *Machine Intelligence Research*, 22(5):983–998, 2025. 2

[52] Tianlu Zhang, Xiaoyi He, Qiang Jiao, Qiang Zhang, and Jungong Han. Amnet: Learning to align multi-modality for rgb-t tracking. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(8):7386–7400, 2024. 1, 3

[53] Xiaowei Zhao, Chenglong Li, Jin Tang, and Chuanfu Li. Learning with explicit topological priors for chest x-ray rib segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 300–309. Springer, 2025. 4

[54] Jiawen Zhu, Simiao Lai, Xin Chen, Dong Wang, and Huchuan Lu. Visual prompt multi-modal tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9516–9526, 2023. 6, 1

[55] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. In *International Conference on Learning Representations*, 2021. 5, 2

[56] Yabin Zhu, Chenglong Li, Jin Tang, and Bin Luo. Quality-aware feature aggregation network for robust rgbt tracking. *IEEE Transactions on Intelligent Vehicles*, 6(1):121–130, 2020. 6