Supplementary Materials for MSITrack: A Challenging Benchmark for Multispectral Single Object Tracking

1 Dataset Construction

1.1 Data Acquisition.

Collection Equipment. To construct our dataset, we utilized the MAIA[7] snapshot multispectral camera, a device capable of simultaneously capturing both visible and near-infrared wavelengths at a resolution of 1280 × 960 pixels. The MAIA system comprises nine sensor arrays, including eight single-channel sensors and one RGB sensor. Each sensor is fitted with a band-pass filter, providing spectral sensitivity across the 395–950 nm range. Specifically, channels S1–S8 correspond to the following spectral bands: 395–450, 455–520, 525–575, 580–625, 630–690, 705–745, 750–820, and 825–950 nm.

Data Size. Initially, more than 290k raw frames were captured. Following a thorough evaluation of each scene's relevance for visual tracking, we manually filtered and trimmed the most valuable segments, discarding irrelevant or redundant content. The final dataset comprises 129k carefully selected frames organized into 300 video sequences, with an average length of 431 frames per sequence. As shown in Fig. 2 , MSITrack covers more diverse, more natural, and more challenging scenes. The MSITrack dataset are publicly available at: https://github.com/Fengtao191/MSITrack.

Data Processing. All acquired multispectral images underwent precise registration through geometric and radiometric correction. First, geometric correction was applied using the intrinsic calibration parameters of each sensor to produce distortion-free images. Subsequently, band-wise image registration was performed with respect to a reference image, ensuring pixel-level alignment across all spectral channels. Finally, radial radiometric correction was carried out to mitigate boundary effects—typically observed as darkened pixels—caused by lens curvature. To ensure consistency, image resolution was standardized from the original 1280 × 960 to 1200 × 900 pixels, eliminating invalid edge regions resulting from minor discrepancies in the field of view among sensors. The camera outputs raw-format files, which we converted into the widely used MAT format for downstream processing.

Annotation. For visualization and high-precision annotation, we employed the RGB images captured by the multispectral camera. These images are spatially aligned with the corresponding multispectral data, substantially improving both annotation efficiency and accuracy. We use DarkLabel as an annotation tool. It is available at https://github.com/darkpgmr/DarkLabel.

Privacy Protection. Notably, all personally identifiable information—such as faces and license plates—captured during the collection of ground-level scenes has been fully anonymized using mosaic blurring, ensuring that MSITrack can be safely and ethically used by other researchers.

1.2 Details of Object Classes

MSITrack covers 55 classes, aiming to facilitate the development of universal and general multispectral object tracking. All these fine categories are verified by the expert to ensure that the sequences in this class are suitable for tracking. Fig. 1 displays the category organization in MSITrack. In order to enable better understanding of MSITrack, below we present each object category and the number of sequences in the format of "Class (# videos)", e.g., "Person (53)", that means object category of "Person" with 53 videos, arranged from largest to smallest as follows: Person (53), Car (33), Bicycle (27), Motorcycle (22), Boat (21), Duck (14), Children (10), Goose (6), Athlete (5), Dog (5), Flamingo (5), Instrument (5), Mandarin Duck (5), Pelicans (5), Pigeon (5), Ball (4), Giraffe (4), Meerkat (4), Monkey (4), Ruddy Shelduck (4), Anteater (3), Black Swan (3), Gorillas (3), Peacock (3), Siberian roe deer (3), Tiger (3), Alpaca (2), Bamboo Rat (2), Central American tapir (2), Cheetah (2), Deliveryman (2), Elephant (2), Elk (2), Goat (2), Magpie (2), Ring-tailed lemur (2), Swan (2), Trolley (2), Black Muntjac (1), Burro (1), Cat (1), Garbage Shovel (1), Goldfish (1), Horse (1), Kangaroo (1), Lion (1), Pig (1), Porcupine (1), Red Goral (1), Rhea Americana (1), Small Muntjac (1), Striped Mongoose (1), Yak (1), Yellow-billed Stork (1), Zebra (1).

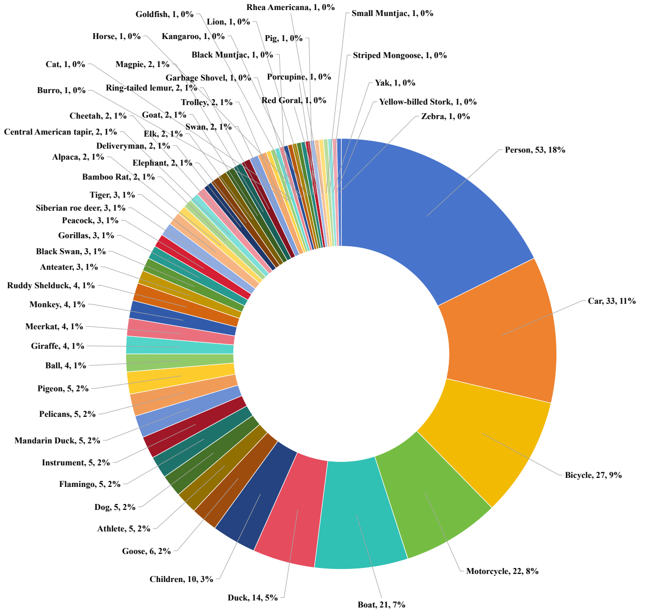

Figure 1: Category organization of MSITrack. Please zoom in.

1.3 Statistic of Video Length

In Fig. 3 , we show distribution of video length on MSITrack. Please notice, MSITrack has an average video length of 431 frames, a maximum video length of 2251 frames and a minimum video length of 55 frames, mainly focusing on challenging tracking.

Figure 2: Diverse and natural scenes in MSITrack (partial)

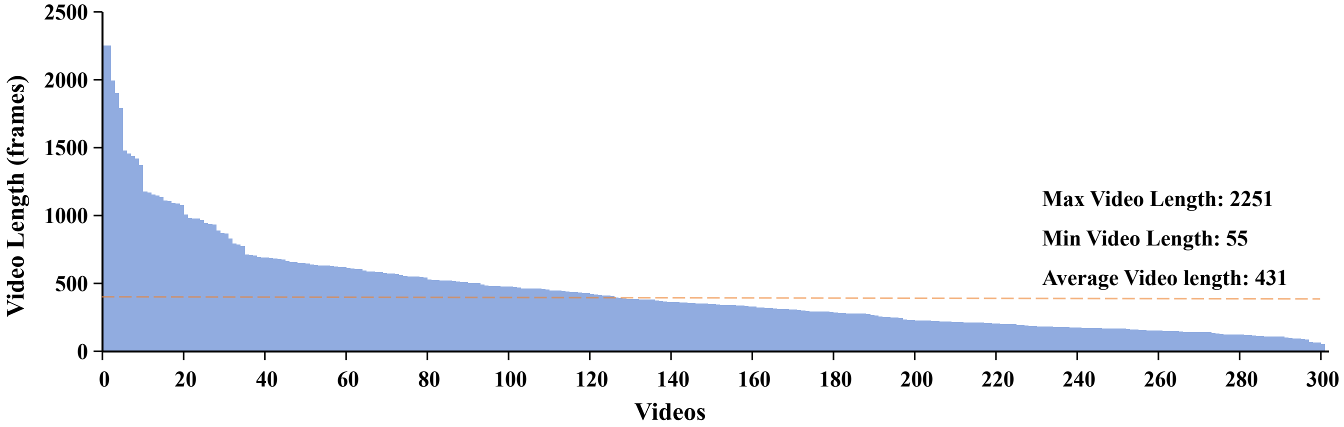

Figure 3: Distribution of sequence length on MSITrack. The average video length of MSITrack is 431 frames (about 86 seconds).

2 Experiments

2.1 Experimental Settings

We trained and evaluated a range of state-of-the-art trackers on the proposed Multispectral Single Object Tracking (MSITrack) dataset, with each tracker fine-tuned to achieve optimal performance. For RGB-based evaluations, we employed the AdamW optimizer with an initial learning rate of 4×10^{-4}. The search region was resized to 256 × 256 pixels and the template region to 128 × 128 pixels. Training was performed for 50 epochs with a batch size of 16, and the learning rate was reduced by a factor of 10 after 30 epochs. All experiments were conducted on an NVIDIA RTX 3090 GPU.

For MSI-based evaluations, all eight spectral channels were used. The only modification made to the trackers during training and testing was adjusting the number of input channels from 3 to 8. All other experimental settings remained identical to those used in the RGB-based configuration.

2.2 Pre-trained Weights

For evaluations based on RGB inputs, we employ RGB images captured from the same viewpoint using the RGB sensor embedded in the multispectral camera. The pre-trained weights for these RGB-based trackers follow the official configurations provided by their respective implementations, with no modifications.

Given the differences in spectral resolution and wavelength coverage across datasets, there are currently no publicly available pre-trained parameters tailored for MSI-based tasks. To address this gap, we instead initialize the network with ImageNet-trained

Table 1: Comparison with state-of-the-art trackers in AUC across each challenge attribute. The top two results are highlighted in red and blue.

Method	SV	POC	FOC	BC	IV	LR	DEF	FM	OV	SOB	SC
SGLATrack[12]	39.7	32.1	26.2	21.0	33.8	36.5	30.8	28.0	34.1	25.7	37.5
HIPTrack[1]	38.9	34.7	27.3	21.4	35.7	39.3	29.2	29.2	33.5	29.5	27.8
OSTrack$_{256}$[13]	44.1	37.9	36.1	30.5	44.9	42.9	38.3	38.3	42.8	33.1	42.9
OSTrack$_{384}$[13]	44.2	38.9	36.1	27.8	41.3	45.2	38.2	37.3	41.9	35.4	43.7
LMTrack[11]	47.8	40.5	34.6	33.5	44.0	40.9	40.7	36.6	**58.0**	29.4	40.1
ZoomTrack[6]	45.4	40.9	39.9	29.3	42.4	47.7	39.1	**40.5**	32.8	38.0	42.6
NeighborTrack[2]	45.7	41.3	39.2	31.4	42.6	46.7	41.5	37.1	32.7	38.3	40.8
EVPTrack[9]	47.1	42.3	37.2	35.8	46.8	46.8	**44.6**	39.8	**55.5**	**39.9**	**45.8**
AQATrack[10]	48.5	43.2	40.0	**36.1**	**48.0**	45.6	42.8	39.6	41.2	36.7	45.1
UNTrack[8]	**51.4**	**44.6**	**40.7**	34.0	43.8	**53.5**	37.0	40.2	39.3	**40.3**	41.3
GRM[4]	**50.3**	**43.4**	**43.5**	**36.7**	**50.3**	**48.5**	**44.3**	**42.0**	45.9	39.0	**46.9**

Table 2: Comparison between RGB-based and MSI-based models on MSITrack.

Method	Input	AUC	$SR_{0.5}$	$SR_{0.75}$	Pre	Pre_N
GRM[4]	RGB	42.5	52.1	35.8	55.2	51.8
	MSI	**47.8 (↑5.3)**	**58.1 (↑6.0)**	**40.8 (↑5.0)**	**63.0 (↑7.8)**	**57.8 (↑6.0)**
AQATrack[10]	RGB	43.4	53.1	38.4	56.2	52.5
	MSI	**47.1 (↑3.7)**	**56.7 (↑3.6)**	**44.7 (↑6.3)**	**58.5 (↑2.3)**	**56.9 (↑4.4)**
OSTrack$_{256}$[13]	RGB	38.7	47.2	32.6	51.0	46.8
	MSI	**42.0 (↑3.3)**	**50.9 (↑3.7)**	**34.2 (↑1.6)**	**55.7 (↑4.7)**	**50.5 (↑3.7)**

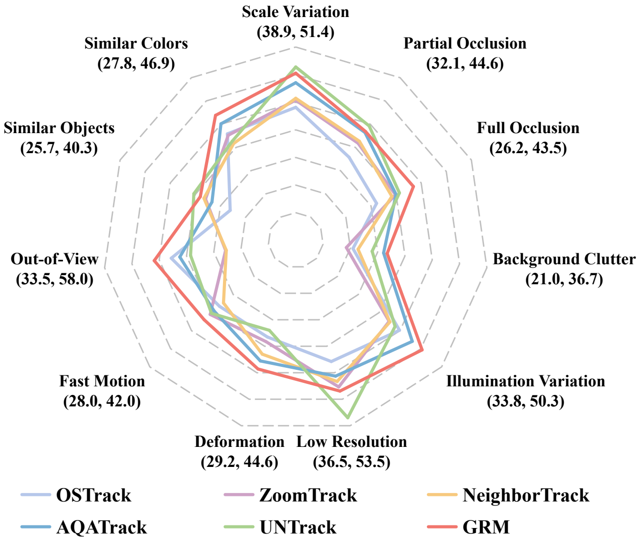

Figure 4: Comparison with state-of-the-art trackers in terms of AUC across each challenge attribute on MSITrack dataset, with the lowest and highest performance values marked.

weights[3, 5] and employ a simple yet effective parameter reconstruction strategy. The strategy aims to extend RGB-based pretrained parameters to MSI-based vision tasks through replication. Specifically, for each MSI channel, the corresponding weights are inherited from the RGB channel whose central wavelength is closest to that of the MSI channel.

2.3 Experimental Results

Attribute-Based Performance. To facilitate a more detailed comparison of tracker performance under MSI input, we conducted a comprehensive AUC-based evaluation across 11 challenging attributes. As shown in Tab. 1 and Fig. 4, UNTrack, GRM and AQATrack achieved the top three results on the two most frequently occurring attributes—POC and SV (ranked by the number of videos per attribute)—with respective AUC scores of 0.514/0.446, 0.503/0.434 and 0.485/0.432. These rankings align closely with their overall performance. Notably, UNTrack and GRM also ranked as the top two trackers for the more difficult attributes, SOB and SC. A particularly interesting finding concerns the LR attribute, which is generally regarded as a challenging scenario for tracking[8]. However, our results reveal a contrary trend. We attribute this to the enriched spectral information provided by MSI data, which enhances the tracker's ability to localize and follow small objects. This insight underscores the critical role of spectral information in improving tracking performance for small-scale targets.

Overall, these results indicate that despite recent progress, state-of-the-art trackers continue to face significant challenges when applied to multispectral video data. Further advancements are necessary to improve robustness and generalizability in multispectral object tracking.

Benefits of Multispectral Cues. To further quantify the advantages of spectral information in object tracking, we compared the performance of the same tracking algorithms using RGB input. As shown in Tab. 2, algorithms utilizing MSI input demonstrated significant improvements across all five performance metrics.

References

[1] Wenrui Cai, Qingjie Liu, and Yunhong Wang. 2024. Hiptrack: Visual tracking with historical prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 19258–19267.

[2] Yu-Hsi Chen, Chien-Yao Wang, Cheng-Yun Yang, Hung-Shuo Chang, Youn-Long Lin, Yung-Yu Chuang, and Hong-Yuan Mark Liao. 2023. Neighbortrack: Single object tracking by bipartite matching with neighbor tracklets and its applications to sports. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5139–5148.

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 248–255.

[4] Shenyuan Gao, Chunluan Zhou, and Jun Zhang. 2023. Generalized relation modeling for transformer tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 18686–18695.

[5] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. 2022. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 16000–16009.

[6] Yutong Kou, Jin Gao, Bing Li, Gang Wang, Weiming Hu, Yizheng Wang, and Liang Li. 2023. Zoomtrack: Target-aware non-uniform resizing for efficient visual tracking. *Advances in Neural Information Processing Systems* 36 (2023), 50959–50977.

[7] E Nocerino, M Dubbini, F Menna, F Remondino, M Gattelli, and D Covi. 2017. Geometric calibration and radiometric correction of the maia multispectral camera. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* 42 (2017), 149–156.

[8] Haolin Qin, Tingfa Xu, Tianhao Li, Zhenxiang Chen, Tao Feng, and Jianan Li. 2025. MUST: The First Dataset and Unified Framework for Multispectral UAV Single Object Tracking. *arXiv preprint arXiv:2503.17699* (2025).

[9] Liangtao Shi, Bineng Zhong, Qihua Liang, Ning Li, Shengping Zhang, and Xianxian Li. 2024. Explicit visual prompts for visual object tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 4838–4846.

[10] Jinxia Xie, Bineng Zhong, Zhiyi Mo, Shengping Zhang, Liangtao Shi, Shuxiang Song, and Rongrong Ji. 2024. Autoregressive queries for adaptive tracking with spatio-temporal transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 19300–19309.

[11] Chenlong Xu, Bineng Zhong, Qihua Liang, Yaozong Zheng, Guorong Li, and Shuxiang Song. 2025. Less is More: Token Context-aware Learning for Object Tracking. *arXiv preprint arXiv:2501.00758* (2025).

[12] Chaocan Xue, Bineng Zhong, Qihua Liang, Yaozong Zheng, Ning Li, Yuanliang Xue, and Shuxiang Song. 2025. Similarity-guided layer-adaptive vision transformer for UAV tracking. *arXiv preprint arXiv:2503.06625* (2025).

[13] 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.

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009