

MSITrack: A Challenging Benchmark for Multispectral Single Object Tracking

Tao Feng¹ Tingfa Xu^{1*} Haolin Qin¹ Tianhao Li¹ Shuaihao Han¹ Xuyang Zou¹

Zhan Lv¹ Jianan Li^{1*}

¹Beijing Institute of Technology

Abstract

Visual object tracking in real-world scenarios presents numerous challenges including occlusion, interference from similar objects and complex backgrounds — all of which limit the effectiveness of RGB-based trackers. Multispectral imagery, which captures pixel-level spectral reflectance, enhances target discriminability. However, the availability of multispectral tracking datasets remains limited. To bridge this gap, we introduce **MSITrack**, the largest and most diverse multispectral single object tracking dataset to date. MSITrack offers the following key features: (i) **More Challenging Attributes** – including interference from similar objects and similarity in color and texture between targets and backgrounds in natural scenarios, along with a wide range of real-world tracking challenges; (ii) **Richer and More Natural Scenes** – spanning 55 object categories and 300 distinct natural scenes, MSITrack far exceeds the scope of existing benchmarks. Many of these scenes and categories are introduced to the multispectral tracking domain for the first time; (iii) **Larger Scale** – 300 videos comprising over 129k frames of multispectral imagery. To ensure annotation precision, each frame has undergone meticulous processing, manual labeling and multi-stage verification. Extensive evaluations using representative trackers demonstrate that the multispectral data in MSITrack significantly improves performance over RGB-only baselines, highlighting its potential to drive future advancements in the field. The MSITrack dataset is publicly available at: <https://github.com/Fengtao191/MSITrack>.

Keywords

Multispectral Image; Single Object Tracking; Benchmark Dataset

1 Introduction

Object tracking is a fundamental problem in computer vision, with broad applications in surveillance and manufacturing. In real-world scenarios, visual trackers often encounter major challenges such as occlusion, interference from visually similar objects and low discriminability due to color or texture similarity between the target and background. These factors considerably complicate the task of visual object tracking.

Existing RGB-based tracking algorithms[11, 26, 32] primarily rely on spatial appearance features—such as color and texture—to detect and track objects. However, in complex scenarios involving occlusion or background similarity, these appearance cues can degrade severely, compromising the algorithm’s ability to reliably distinguish and track the target. As illustrated in Fig. 1(a), a meerkat is nearly indistinguishable from the background, demonstrating the limitations of appearance-based tracking methods. This highlights the need to explore feature dimensions beyond visual appearance

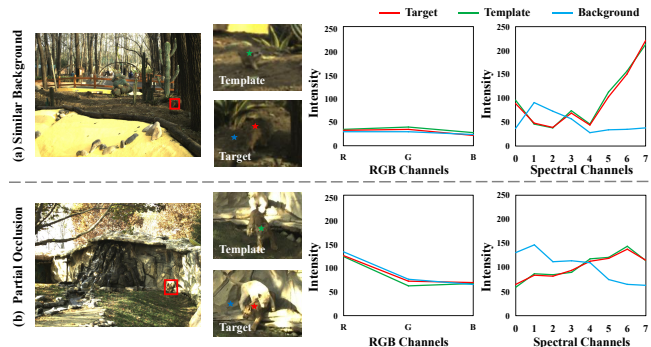


Figure 1: The target’s spectral information differs significantly from the background and aligns with the template’s spectral data, facilitating differentiation and localization.

in order to enhance object distinguishability and improve the accuracy and robustness of general-purpose tracking algorithms under real-world conditions.

Multispectral imaging (MSI) captures the spectral reflectance of objects at the pixel level across a wide range of wavelengths, thereby providing richer information on material composition and surface properties information[5, 10, 23] than RGB imagery. In challenging scenarios involving background interference or occlusion, MSI provides complementary spectral cues that preserve target identity even when spatial features are compromised. As shown in Fig. 1(b), a distinct spectral curve allows for effective target-background separation, even under severe occlusion. Compared to RGB data, multispectral imagery offers more discriminative spectral signatures[18], presenting a more robust solution for object tracking. However, progress in this area is hindered by the lack of large-scale, high-quality multispectral tracking datasets.

To address this limitation, we present MSITrack—the largest and most diverse multispectral single object tracking dataset. Collected using a snapshot multispectral camera, MSITrack captures real-world scenes under a wide range of temporal and environmental conditions. The dataset features three major characteristics:

(i) More Challenging Attributes. The collected sequences are intentionally designed to be highly challenging, capturing scenarios such as interference from similar objects and visual similarity between the target and background in terms of color and texture. These challenges closely mirror real-world conditions.

(ii) Richer and More Natural Scenes. To enhance the diversity and generalization capabilities of multispectral tracking, MSITrack includes 55 object categories and 300 unique natural scenes — substantially more than existing benchmarks such as HOT (20 categories and 21 scenes). Many of the newly introduced categories and

Table 1: Comparison of MSITrack with other datasets. – means no relevant information was provided.

Dataset	Modality	Channels	Task	Classes	Total Frames	Durations	Sequences	Resolution	Wavelength Range
OTB50[27]	RGB	3	General SOT	10	29k	16.4 min	50	-	None
VOT2017[12]	RGB	3	General SOT	24	21k	11.9 min	60	-	None
TrackingNet[20]	RGB	3	General SOT	27	509k	140 hours	>30K	-	None
LaSOT[4]	RGB	3	General SOT	70	3.52M	32.5 hours	1.4K	1280×720	None
GTOT[14]	RGBT	4	General SOT	9	7.8k	-	50	-	None
RGBT210[17]	RGBT	4	General SOT	22	104.7k	-	210	-	None
MUST[22]	MSI	8	UAV SOT	8	43k	143 min	250	1200×900	395-950nm
HOT[29]	MSI	16	General SOT	20	21k	14 min	50	512×256	470-620nm
MSITrack (Ours)	MSI	8	General SOT	55	129k	430 min	300	1200×900	395-950nm

scenes are appearing in the context of multispectral tracking for the first time, thereby expanding the coverage and applicability.

(iii) Larger Scale. The dataset consists of 300 videos comprising a total of 129k multispectral frames, each with a spatial resolution of 1200×900 pixels. It spans eight spectral bands ranging from 395 nm to 950 nm, covering the visible to near-infrared spectrum. These sequences represent 300 distinct natural scenes. All frames have been manually annotated through a meticulous process requiring over 1,300 labor hours, ensuring high-quality bounding boxes and providing a robust foundation for reliable evaluation.

The proposed dataset provides a solid foundation for advancing the field of multispectral single object tracking. To benchmark existing methods, we conducted a comprehensive evaluation of a wide range of recent state-of-the-art trackers. Experimental results consistently show that trackers leveraging multispectral input significantly outperform their RGB-based counterparts, particularly in complex scenarios where spatial appearance features are limited. These findings underscore the potential of MSITrack to support and inspire future advancements in this domain.

Our main contributions are as follows: (i) We present MSITrack, the largest and most diverse dataset for multispectral single object tracking to date; (ii) The dataset incorporates more challenging attributes and support the development of more robust tracking algorithms; (iii) We perform extensive experimental evaluations, establishing strong baselines to guide future research and comparative studies. All dataset resources and source code are publicly available to encourage further development and promote reproducibility.

2 Related Work

MSI-based Tracking Datasets. Several MSI datasets have recently been introduced for visual object tracking. The HOT[29] dataset is a ground-based collection acquired using a snapshot mosaic hyperspectral camera, comprising 50 videos across 20 object categories. Due to hardware limitations, its spectral coverage is confined to the 470–620 nm range, with a spatial resolution of 512×256 pixels. The MUST[22] dataset, by contrast, is designed specifically for aerial-view multispectral tracking. It includes 250 videos spanning 8 object categories and demonstrates the benefits of spectral data in airborne environments. All scenes in MUST are captured from a high-altitude top-down perspective over urban streets, resulting in small target sizes due to the elevated viewpoint. While the aforementioned datasets have significantly contributed to the progress of

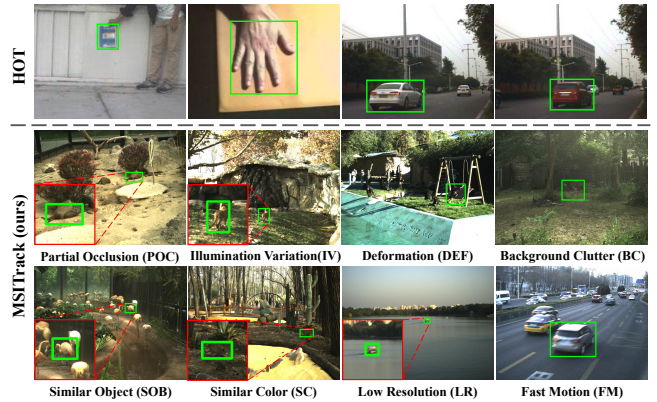


Figure 2: Comparison of scenes between MSITrack and HOT.

multispectral tracking, their overall scale remains relatively limited. To overcome this limitation, we propose MSITrack, a challenging dataset encompassing a broader and more diverse set of scenes.

RGB-based Tracking Datasets. Early tracking datasets were limited both in variety and scale. OTB-2013[27], the first single object tracking dataset, contained only 50 videos and 10 categories. TC-128[19] consists of 128 RGB videos, while VOT2017[12] introduced a series of challenges to better compare the performance of various trackers. TrackingNet[20] proposed a large-scale tracking dataset, including over 30k tracking sequences. LaSOT[4] contains 1,400 long sequences, and GOT-10k[8] expanded the tracking categories to 563 categories and 10k videos. In contrast, MSITrack introduces a complementary and discriminative spectral dimension that offers a more effective solution for object tracking.

RGBT-based Tracking Datasets. The GTOT[14] dataset comprises 50 RGBT videos, using visible light and thermal infrared images for tracking. Li et al.[17] introduced the RGBT210 dataset, which includes 210 RGBT videos featuring annotations for 12 challenge attributes. This dataset was subsequently expanded into RGBT234[15], which contains 234 RGBT videos, offering improved annotation quality and broader challenge coverage. However, thermal infrared imaging is susceptible to interference from external heat sources[13, 16, 34]. In contrast, multispectral imaging demonstrates stable performance under natural daylight conditions and is less prone to errors caused by thermal variations[9, 25, 33].

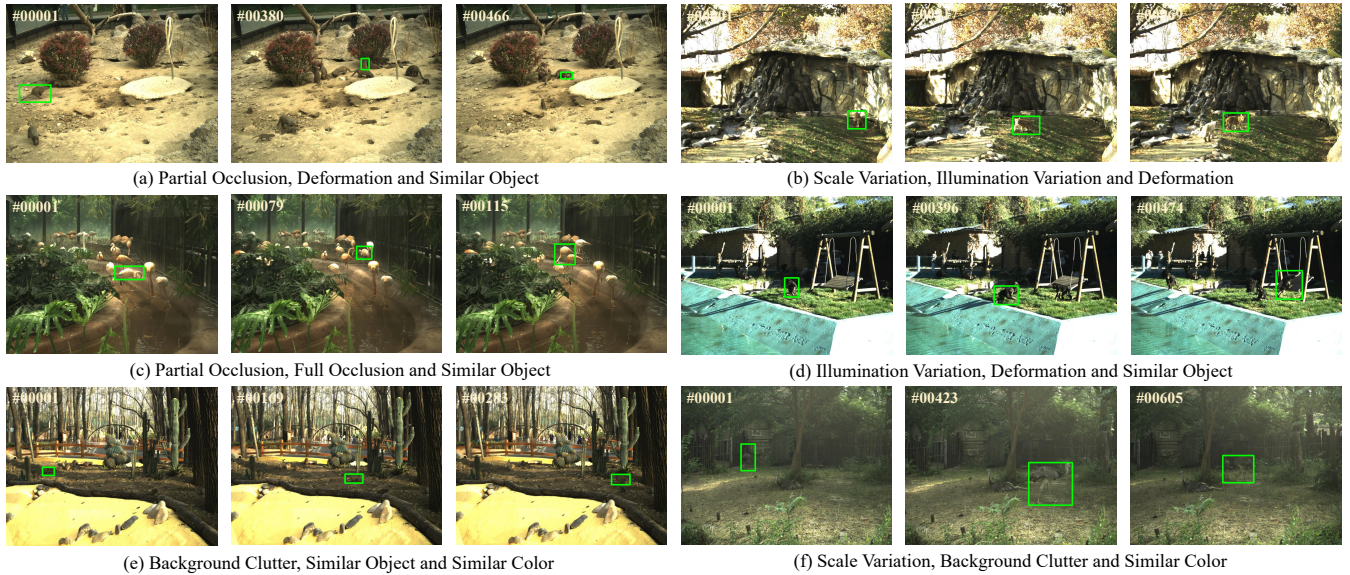


Figure 3: Visualization of several annotation examples in the proposed MSITrack.

3 MSITrack Dataset

3.1 Construction Principle

MSITrack is designed to serve as a comprehensive, large-scale and highly challenging benchmark tailored for general object tracking in real-world scenarios. To this end, the construction of MSITrack follows several key principles:

- **More Complex Real-World Challenges.** The dataset should capture extensive variations in background complexity, occlusion levels, and include more challenging attributes such as similar objects interference, color similarity between targets and backgrounds, offering comprehensive coverage of the real-world challenges.
- **More Diverse and Real-World Scenes.** A rich variety of tracking categories and natural scenes is critical for advancing generalizable multispectral tracking. MSITrack should encompass a variety of dynamic environments, including urban areas, forests, wetlands and lakes, and include more than 50 object categories, further broadening its applicability and relevance.
- **Larger Scale.** A primary goal of MSITrack is to deliver a large-scale multispectral dataset that supports both the training of deep learning models and robust performance evaluation. To this end, we target a large-scale dataset comprising 300 videos and over 100k multispectral images, enabling general-purpose object tracking across a broad range of conditions.

3.2 Dataset Overview

As shown in Fig. 4(a), MSITrack is currently the largest and most diverse multispectral single object tracking dataset, comprising 300 videos and over 129k frames of multispectral imagery. As illustrated in Fig. 4(c), MSITrack includes a wide range of tracking categories. In addition to commonly seen classes such as pedestrian, car and bicycle, the dataset features less frequently encountered objects such as meerkat, lion, gorilla, flamingo, elk, monkey and giraffe,

totaling 55 distinct categories. Many of these classes are introduced to the multispectral tracking domain for the first time.

Tab. 1 presents a comparative overview of MSITrack and several representative general-purpose and multispectral single object tracking datasets. The MUST dataset targets traffic-related objects from an aerial, top-down viewpoint. However, its limited range of object categories constrains its utility in broader visual tracking applications. HOT, one of the earliest general-purpose multispectral tracking datasets, has significantly contributed to the development of the field. Nevertheless, as shown in Fig. 2, HOT primarily includes relatively simple targets such as books, hands and cars, some of which are manually manipulated, and its scenes often overlap.

In contrast, MSITrack spans a significantly broader spectral range—from 395 nm to 950 nm—covering wavelengths from the visible to the near-infrared spectrum. It also marks a substantial increase in scale, with a greater variety of object categories (55 vs. 20) and more diverse, natural scenes (300 vs. 21), each of which is unique. Importantly, MSITrack emphasizes real-world complexity by including a substantial number of videos featuring similar-object interference (SOB) and low target-to-background separability in terms of color and texture (SC). These challenging attributes more precisely capture the complexities of real-world tracking scenarios involving heterogeneous object types and diverse environments, thereby facilitating further research in this domain.

3.3 Dataset Construction

Data Acquisition. MSITrack was developed using a snapshot-based multispectral camera capable of capturing images across eight spectral bands, covering both the visible and near-infrared ranges, with a spectral sensitivity from 395 nm to 950 nm. To ensure the dataset’s relevance to general-purpose object tracking tasks, data collection was conducted under diverse weather conditions, including sunny, overcast and foggy days.

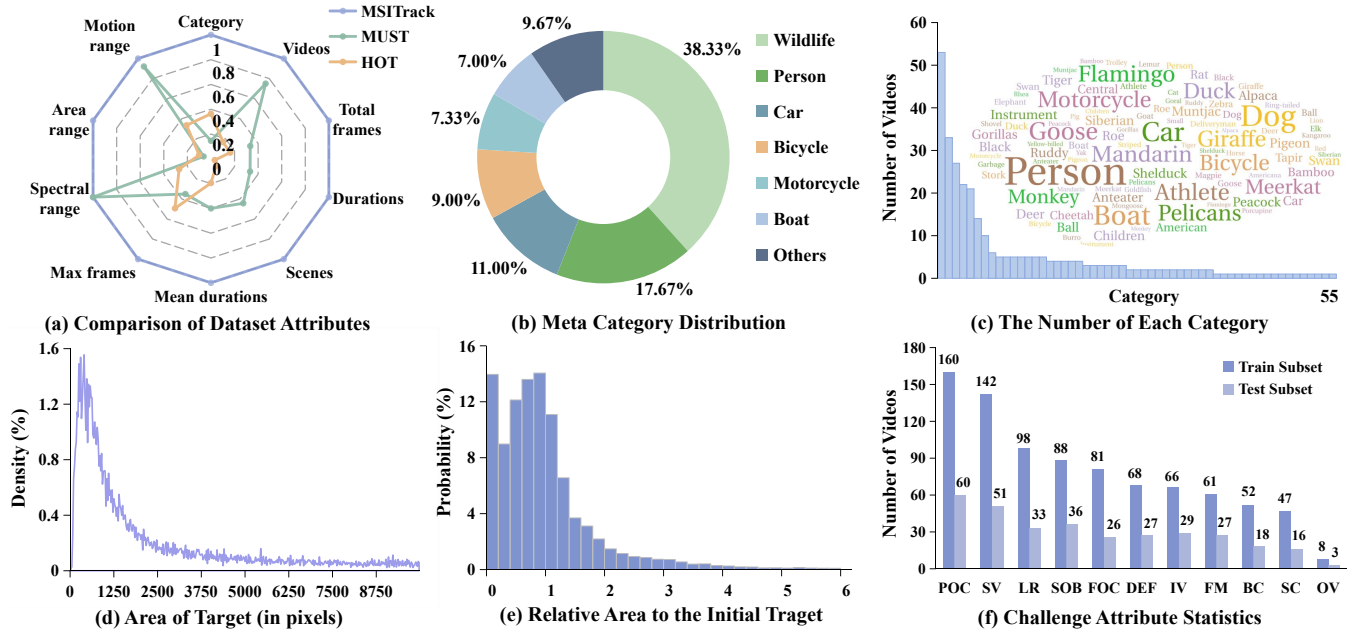


Figure 4: Comparison of datasets, category distribution, object size, relative area and dataset splitting in MSITrack.

Table 2: Descriptions of 11 different attributes in MSITrack

Attribute	Definition
POC	the target object is partially occluded in the sequence
SV	the ratio of bounding box is outside [0.5, 2]
LR	the area of target box is smaller than 1000 pixels
SOB	the target is surrounded by visually similar distractors
FOC	the target object is fully occluded in the sequence
DEF	the target object is deformable during tracking
IV	the illumination in the target region changes
FM	the motion of target object is larger than 50 pixels
BC	the background of the target has some clutter
SC	the background has a color similar to that of the target
OV	the target object completely leaves the video frame

The dataset spans a broad range of environments, such as urban streets, lakes and wetlands, forest parks, sports fields, playgrounds and wildlife reserves. Initially, over 290k raw frames were captured. This careful curation process resulted in a final dataset of 129k frames, forming 300 tracking sequences with an average length of 431 frames per sequence. All frames underwent precise alignment procedures, including both geometric and radiometric corrections, to ensure the generation of high-quality multispectral sequences optimized for general-purpose object tracking.

Annotation. MSITrack is a carefully curated dataset containing over 1300 human-hours of manual annotations. Bounding box annotations follow a rigorous set of guidelines[21]. If the target is fully occluded or out of the frame, no bounding box is provided. We established an annotation team, referred to as LYN, comprising a small group of experts and ten qualified annotators with experience

Table 3: Detailed statistics across datasets. Scenes: Non-repeating scenes; Time: Average duration; Frames: Maximum number of frames; Motion: Maximum moving pixels.

Dataset	Scences	Time	Frames	Motion
MUST	133	34s	790	453
HOT	21	17s	1111	16
MSITrack	300	86s	2251	492

in object tracking. The annotation process was organized into two stages—annotation and verification—implemented through a multi-step workflow that included manual labeling, visual inspection and cross-validation.

In the first stage, after an expert annotated the initial frame of each sequence, the remaining frames were manually annotated by the trained annotators using a standardized annotation tool. Upon completion, the process advanced to the verification stage, where three experienced annotators from the original team conducted cross-validation and refined the annotations to ensure accurate bounding box placement and sizing. Each bounding box was meticulously adjusted to tightly fit the target with minimal margin. Fig. 3 shows several annotation examples.

Dataset Splitting. MSITrack is divided into a training set and a test set. To ensure fairness and promote generalization, environmental factors—such as weather conditions—are evenly distributed across both subsets. The final split comprises 220 training sequences and 80 testing sequences. The training set contains 94k frames, with an average sequence length of 431 frames, while the test set includes 35k frames, with an average sequence length of 429 frames. As illustrated in Fig. 4(f), we carefully maintain an even distribution

of 11 key challenge attributes across both the training and testing sets, thereby ensuring comparable difficulty and diversity between the subsets.

3.4 Statistical Analysis

Challenging Attributes. As shown in Tab. 2, MSITrack incorporates annotations for 11 challenging tracking attributes across all sequences. As illustrated in Fig. 4(f), partial occlusion is the most frequently encountered challenge, present in 220 videos. More complex conditions—such as similar-object interference and target-background color similarity—occur in 124 sequences (41.3%) and 63 sequences (21.0%), respectively. The former evaluates a tracker’s ability to preserve target identity under object ambiguity, while the latter assesses sensitivity to subtle appearance differences and fine-grained target discrimination. As shown in Tab. 3, MSITrack offers a substantially greater diversity of unique scenes compared to existing datasets. These attribute-driven analyses highlight the complexity of the MSITrack, limiting the effectiveness of purely spatial feature approaches and motivating the development of more robust and adaptive tracking algorithms.

Category and Target Size Distribution. As shown in Fig. 4(b) and Fig. 4(c), MSITrack includes a diverse array of target categories, with the Wildlife meta-category comprising the largest proportion. Wildlife targets typically exhibit a wide range of fur textures, colors and body structures, thereby enriching the dataset’s spectral feature space and enhancing its relevance to real-world scenarios. In contrast to rigid objects such as vehicles, animal targets present substantial non-rigid deformations and irregular motion patterns, and are often situated in visually complex and dynamically changing environments. These attributes collectively support research in fine-grained target discrimination and robustness to deformation under practical conditions. The target area distribution depicted in Fig. 4(d) further substantiates this observation, with the primary peak occurring around medium-sized targets (about 1000 pixels), and a long right tail indicating the inclusion of large-scale targets. **Trajectory Duration and Inter-Frame Motion Analysis.** As summarized in Tab. 3, MSITrack features significantly longer average sequence durations—2.5 times those of MUST and 5.1 times those of HOT. Additionally, MSITrack surpasses existing multispectral datasets in both the maximum frame count per sequence and the maximum displacement of targets. The distribution of relative area to the initial target, shown in Fig. 4(e), reflects realistic and diverse scale variations, which are commonly encountered in practical tracking scenarios. These attributes introduce greater challenges for tracking algorithms, particularly in terms of responsiveness and robustness, thereby enabling a more comprehensive evaluation of their long-term stability and temporal consistency.

4 Experiments

4.1 Experimental Settings

We conducted algorithm evaluations on the MSITrack dataset using two types of input: RGB and MSI. For the RGB-based evaluation, we utilized images captured by the RGB camera of the multispectral system, synchronized and aligned to the same viewpoint. For the MSI-based evaluation, data from all eight spectral channels were employed, only changing the number of input channels to the

Table 4: Comparison with state-of-the-art trackers on MSI-Track. The top two results are highlighted in red and blue.

Method	AUC	SR _{0.5}	SR _{0.75}	Pre	Pre _N
SGLATrack[31]	35.9	43.7	27.6	48.5	43.9
HIPTrack[1]	36.2	43.1	29.0	47.8	43.0
OTrack ₂₅₆ [32]	42.0	50.9	34.2	55.7	50.5
OTrack ₃₈₄ [32]	43.1	52.5	36.1	56.2	52.4
LMTrack[30]	43.2	52.4	39.7	55.2	52.5
ZoomTrack[11]	44.3	54.7	35.1	59.0	54.4
NeighborTrack[2]	44.8	55.2	35.7	59.7	54.6
EVPTTrack[24]	46.2	55.8	43.9	58.3	56.0
AQATrack[28]	47.1	56.7	44.7	58.5	56.9
UNTrack[22]	47.3	58.4	40.5	61.8	57.8
GRM[6]	47.8	58.1	40.8	63.0	57.8

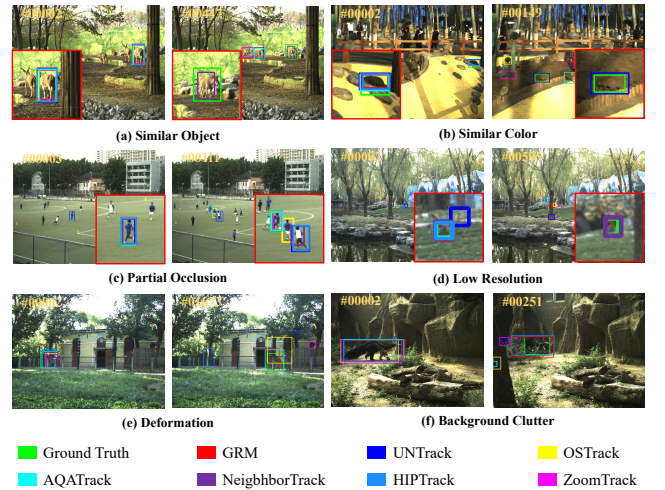


Figure 5: Comparison of some state-of-the-art trackers when dealing with various challenge attributes.

tracker from 3 to 8 during training and testing. Due to the lack of available pre-trained parameters for MSI, we instead initialize the network with ImageNet-trained weights[3, 7] and employ a simple yet effective parameter reconstruction strategy. The strategy aims to extend RGB-based pre-trained parameters to MSI-based vision tasks through replication.

Following[4, 20, 27], we use one-pass evaluation (OPE) and compare different trackers on MSITrack using five metrics: Area Under the Curve (AUC) of the success plot, SR_{0.5}, SR_{0.75}, precision (PRE) and normalized precision (NPRE).

4.2 Experimental Results and Analysis

Overall Performance. All models were evaluated under comprehensive and fair conditions, with detailed results presented in Tab. 4. Among the state-of-the-art trackers assessed, GRM achieved the best performance across all metrics, with scores of 0.478 AUC, 0.630 PRE and 0.578 NPPE. UNTrack ranked second overall, attaining scores of 0.473 AUC, 0.618 PRE and 0.578 NPPE. The third-ranked

Table 5: Comparison between RGB and MSI Input.

Method	Input	AUC	Pre	Pre _N
GRM[6]	RGB	42.5	55.2	51.8
	MSI	47.8 (↑5.3)	63.0	57.8
AQATrack[28]	RGB	43.4	56.2	52.5
	MSI	47.1 (↑3.7)	58.5	56.9
OSTrack ₂₅₆ [32]	RGB	38.7	51.0	46.8
	MSI	42.0 (↑3.3)	55.7	50.5

Table 6: A comparison of three state-of-the-art trackers in AUC across each challenge attribute.

Method	Input	LR	FOC	IV	SC	SV	POC	SOB
GRM	RGB	40.4	35.8	42.1	40.7	44.7	39.9	34.8
	MSI	48.5	43.5	50.3	46.9	50.3	43.4	39.0
AQATrack	RGB	40.3	36.7	43.4	42.5	43.9	40.8	33.8
	MSI	45.6	40.0	48.0	45.1	48.5	43.2	36.7
OSTrack ₂₅₆	RGB	35.3	28.1	40.5	35.8	40.0	33.4	31.3
	MSI	42.9	36.1	44.9	42.9	44.1	37.9	33.1

tracker, AQATrack, achieved scores of 0.471 AUC, 0.585 PRE and 0.569 NPPE. Notably, UNTrack is specifically designed for multi-spectral object tracking. It employs a spectral background removal module to aggregate useful spectral information while narrowing the search area, underscoring the significance of spectral information in deep tracking. The tracking results of each algorithm are visualized in Fig. 5.

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. 5, algorithms utilizing MSI input demonstrated significant improvements across all five performance metrics. Specifically, GRM’s AUC score increased by 5.3%, AQATrack’s AUC improved by 3.7%, and OSTrack₂₅₆’s AUC rose by 3.3%. These results underscore the effectiveness of spectral information in enhancing target representation under more challenging scenarios.

In addition, we compared the performance of trackers using RGB and MSI inputs across various challenging attributes, as shown in Tab. 6. Trackers with MSI input consistently outperformed their RGB counterparts in nearly all challenging scenarios. Spectral information demonstrated remarkable effectiveness in handling cases where visual appearance is severely compromised—such as low resolution (LR) and full occlusion (FOC)—achieving average AUC improvements of 7.0% and 6.3%, respectively, across the three trackers. Under illumination variation (IV), the average AUC gain reached 5.7%. In such conditions, the spatial appearance features relied upon by RGB inputs tend to degrade or become ambiguous, thereby reducing the distinguishability of the target. In contrast, the spectral features provided by MSI offer complementary cues that significantly enhance target separability. Notably, in more complex scenarios such as similar colors (SC) and similar object (SOB), the unique spectral signatures contributed to average AUC improvements of

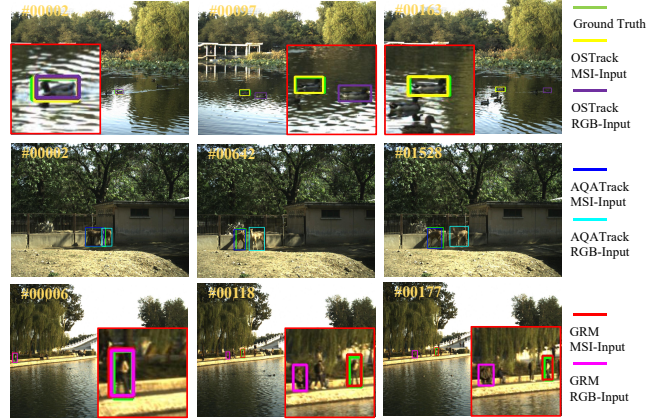


Figure 6: Comparison of three representative trackers (OSTrack, AQATrack and GRM) using RGB and MSI inputs.

5.3% and 2.9%, respectively. These findings highlight the critical role of spectral information in supporting robust object tracking under diverse and challenging conditions.

Qualitative Comparison of RGB and MSI Inputs. As shown in Fig. 6, multispectral input significantly enhances tracking reliability under visually challenging conditions. In the top row, the RGB-based OSTrack (purple box) fails to correctly follow the target mandarin duck due to interference from visually similar objects. In the second row, RGB-based AQATrack (cyan box) misidentifies the target elk in a scene affected by both illumination changes and similar-object interference. In the third row, RGB-based GRM (magenta box) struggles to effectively separate the pedestrian from the background in low-resolution with varying lighting conditions. In contrast, the MSI-based trackers demonstrate more accurate target identification and more stable tracking performance. These qualitative results are not isolated incidents but rather represent consistent patterns observed across the dataset. They highlight how multispectral input improves target discriminability, reduces misidentification and enhances overall tracking robustness in complex scenarios.

5 Conclusion

We present MSITrack, the largest and most comprehensive multi-spectral single-object tracking dataset to date, consisting of 300 videos and 129k multispectral images, spanning 55 object categories. MSITrack introduces greater tracking challenges by incorporating numerous scenes with visually similar distractors and low-saliency targets, thereby facilitating the development of models with enhanced robustness and background suppression capabilities in real-world environments. Extensive experiments on 11 state-of-the-art trackers consistently demonstrate that incorporating spectral inputs yields significant performance improvements, especially for small or densely clustered objects where spatial cues are limited. All data and code are publicly available to support further research in this domain.

Limitation. High-quality annotated tracking bounding boxes require substantial manual effort. Future work will explore scalable, unsupervised annotation methods, such as language supervision and unsupervised learning techniques.

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] Heng Fan, Liting Lin, Fan Yang, Peng Chu, Ge Deng, Sijia Yu, Hexin Bai, Yong Xu, Chunyuan Liao, and Haibin Ling. 2019. Lasot: A high-quality benchmark for large-scale single object tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 5374–5383.
- [5] Long Gao, Pan Liu, Yan Jiang, Weiying Xie, Jie Lei, Yunsong Li, and Qian Du. 2023. CBFF-Net: A new framework for efficient and accurate hyperspectral object tracking. *IEEE Transactions on Geoscience and Remote Sensing* 61 (2023), 1–14.
- [6] Shenyuan Gao, Chunlun 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.
- [7] 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.
- [8] Lianghua Huang, Xin Zhao, and Kaiqi Huang. 2019. Got-10k: A large high-diversity benchmark for generic object tracking in the wild. *IEEE transactions on pattern analysis and machine intelligence* 43, 5 (2019), 1562–1577.
- [9] Soonmin Hwang, Jaesik Park, Namil Kim, Yukyung Choi, and In So Kweon. 2015. Multispectral pedestrian detection: Benchmark dataset and baseline. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1037–1045.
- [10] Mohammad Aminul Islam, Wangzhi Xing, Jun Zhou, Yongsheng Gao, and Kuldip K Paliwal. 2024. Hy-Tracker: A Novel Framework for Enhancing Efficiency and Accuracy of Object Tracking in Hyperspectral Videos. *IEEE Transactions on Geoscience and Remote Sensing* (2024).
- [11] 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.
- [12] Matej Kristan, Jiri Matas, Aleš Leonardis, Tomáš Vojtíš, Roman Pflugfelder, Gustavo Fernandez, Georg Nebel, Fatih Porikli, and Luka Čehovin. 2016. A novel performance evaluation methodology for single-target trackers. *IEEE transactions on pattern analysis and machine intelligence* 38, 11 (2016), 2137–2155.
- [13] Xiangyuan Lan, Mang Ye, Shengping Zhang, Huiyu Zhou, and Pong C Yuen. 2020. Modality-correlation-aware sparse representation for RGB-infrared object tracking. *Pattern Recognition Letters* 130 (2020), 12–20.
- [14] Chenglong Li, Hui Cheng, Shiyi Hu, Xiaobai Liu, Jin Tang, and Liang Lin. 2016. Learning collaborative sparse representation for grayscale-thermal tracking. *IEEE Transactions on Image Processing* 25, 12 (2016), 5743–5756.
- [15] Chenglong Li, Xinyan Liang, Yijuan Lu, Nan Zhao, and Jin Tang. 2019. RGB-T object tracking: Benchmark and baseline. *Pattern Recognition* 96 (2019), 106977.
- [16] Chenglong Li, Wanlin Xue, Yaqing Jia, Zhichen Qu, Bin Luo, Jin Tang, and Dengdi Sun. 2021. LasHeR: A large-scale high-diversity benchmark for RGB-T tracking. *IEEE Transactions on Image Processing* 31 (2021), 392–404.
- [17] Chenglong Li, Nan Zhao, Yijuan Lu, Chengli Zhu, and Jin Tang. 2017. Weighted sparse representation regularized graph learning for RGB-T object tracking. In *Proceedings of the 25th ACM international conference on Multimedia*. 1856–1864.
- [18] Jie Liang, Jun Zhou, Lei Tong, Xiao Bai, and Bin Wang. 2018. Material based salient object detection from hyperspectral images. *Pattern Recognition* 76 (2018), 476–490.
- [19] Pengpeng Liang, Erik Blasch, and Haibin Ling. 2015. Encoding color information for visual tracking: Algorithms and benchmark. *IEEE transactions on image processing* 24, 12 (2015), 5630–5644.
- [20] Matthias Muller, Adel Bibi, Silvio Giancola, Salman Alsubaihi, and Bernard Ghanem. 2018. Trackingnet: A large-scale dataset and benchmark for object tracking in the wild. In *Proceedings of the European conference on computer vision (ECCV)*. 300–317.
- [21] Liang Peng, Junyuan Gao, Xinran Liu, Weihong Li, Shaohua Dong, Zhipeng Zhang, Heng Fan, and Libo Zhang. 2024. Vasttrack: Vast category visual object tracking. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.
- [22] 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).
- [23] Haolin Qin, Tingfa Xu, Peifu Liu, Jingxuan Xu, and Jianan Li. 2024. Dmssn: Distilled mixed spectral-spatial network for hyperspectral salient object detection. *IEEE Transactions on Geoscience and Remote Sensing* (2024).
- [24] 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.
- [25] Muhammad Uzair, Arif Mahmood, and Ajmal Mian. 2015. Hyperspectral face recognition with spatio-spectral information fusion and PLS regression. *IEEE Transactions on Image Processing* 24, 3 (2015), 1127–1137.
- [26] 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.
- [27] Yi Wu, Jongwoo Lim, and Ming-Hsuan Yang. 2013. Online object tracking: A benchmark. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2411–2418.
- [28] 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.
- [29] Fengchao Xiong, Jun Zhou, and Yuntao Qian. 2020. Material based object tracking in hyperspectral videos. *IEEE Transactions on Image Processing* 29 (2020), 3719–3733.
- [30] 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).
- [31] 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).
- [32] 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.
- [33] Chunyan Yu, Rui Han, Meiping Song, Caiyu Liu, and Chein-I Chang. 2021. Feedback attention-based dense CNN for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing* 60 (2021), 1–16.
- [34] Lichao Zhang, Martin Danelljan, Abel Gonzalez-Garcia, Joost Van De Weijer, and Fahad Shahbaz Khan. 2019. Multi-modal fusion for end-to-end RGB-T tracking. In *Proceedings of the IEEE/CVF International conference on computer vision workshops*. 0–0.