MUST: The First Dataset and Unified Framework for Multispectral UAV Single Object Tracking

Supplementary Material

1. Dataset Introduction

1.1. Collection

The rapid advancements in UAV and sensor technologies have enabled the collection of multispectral video sequences from UAV platforms. In this study, we employed a multispectral camera, which captures 8 spectral bands in the wavelength range of 390–950 nm, as illustrated in Tab. 1. Additionally, the camera is capable of recording video at 5 frames per second, with a spatial resolution of 1280×960 pixels per frame. To ensure a diverse dataset, we performed data collection across various scenarios and time periods, as depicted in Fig. 1 (a) and (b). Moreover, the UAV captured multi-angle views of several categories of common targets of interest in different flight postures, as shown in Fig. 1 (c). During the data collection process, the UAV maintained an altitude ranging from 20 to 250 meters above ground level, allowing for targets of varying sizes.

1.2. Processing

The multispectral camera adopts array distributed sensors to capture data across different spectral bands. To ensure accurate alignment of the multi-band data, we first perform geometric correction on each individual channel, mitigating parallax errors caused by variations in sensor positioning. Given that geometric correction can alter the resolution of each channel, we crop and retain the central 1200×900 pixels region of each frame. Additionally, the substantial distance between the UAV and the ground target exacerbates the effects of solar radiation and atmospheric scattering on sensors. To address these issues, we apply radiometric correction to calibrate the sensor responses and obtain accurate spectral curves. After these corrections, each multispectral image frame is represented as $\boldsymbol{I} \in \mathbb{R}^{1200 \times 900 \times 8}$.

We engaged domain experts to to annotate the dataset using the DarkLabel 2.3 toolbox. For each frame, we provided the target's bounding box and status flags as ground truth. The bounding box is represented as $[x_{min}, y_{min}, w, h]$, where (x_{min}, y_{min}) denote the coordinates of the upper-left corner, and (w, h) indicates target shape. If the target is fully occluded or temporarily out of view, the status flags are set to 1, with the bounding box marked as $[0, 0, 0, 0]$. After thorough screening, we retained 250 video sequences, totaling 42,671 frames and approximately 8,534 seconds. Detailed dataset information is provided in Table 2.

We divide the train and test subsets primarily based on the principle that the data across subsets is unbiased and consistently distributed. Initially, we randomly divide the data into two subsets. Then we adjust them to ensure consistency in terms of average frames across sequences, target size distribution, and proportion of challenge attributes.

Table 1. The distribution of spectral bands collected by the used multispectral camera.

Bands	Start (nm)	End (nm)	Center (nm)	Width (nm)
1	395	450	422.5	55
2	455	520	487.5	65
3	525	575	550	50
4	580	625	602.5	45
5	630	690	660	60
6	705	745	725	40
7	750	820	785	70
8	825	950	887.5	125

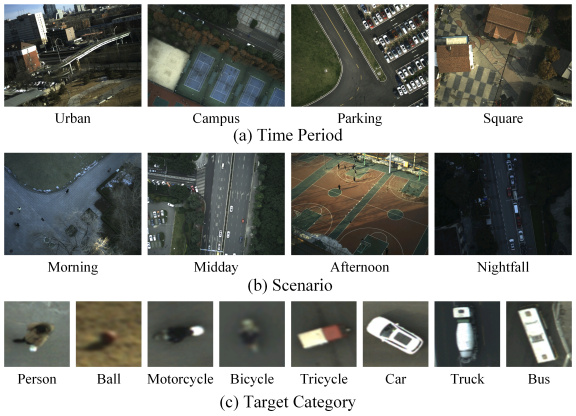

Figure 1. Eight target categories collected across various scenarios and time periods.

1.3. Challenge Attributes

To comprehensively evaluate state-of-the-art trackers, our proposed MUST dataset have defined 12 key challenge attributes that reflect the unique characteristics of multispectral UAV tracking tasks. These attributes are: Partial Occlusion (POC), Background Clutter (BC), Low Resolution (LR), Similar Object (SOB), Scale Variation: (SV), Motion Blur (MB), Fast Motion (FM), Similar Color (SC), Out of view (OV), Illumination Variation (IV), Full Occlusion (FOC) and Camera Motion (CM), as summarized in Tab. 3. Each video sequence in the dataset is annotated with multiple challenge attributes, and Fig. 2 exhibits typical examples

Table 2. Properties of the Multispectral UAV Single Object Tracking (MUST) Dataset.

Name	Video	Duration	Frames				Resolution		
			Total	Minimal	Maximal	Mean	Spatial Pixels	Spectral Range	Band Numbers
MUST	250	8534 s	42671	42	790	171	1200 × 900	390 - 950 nm	8

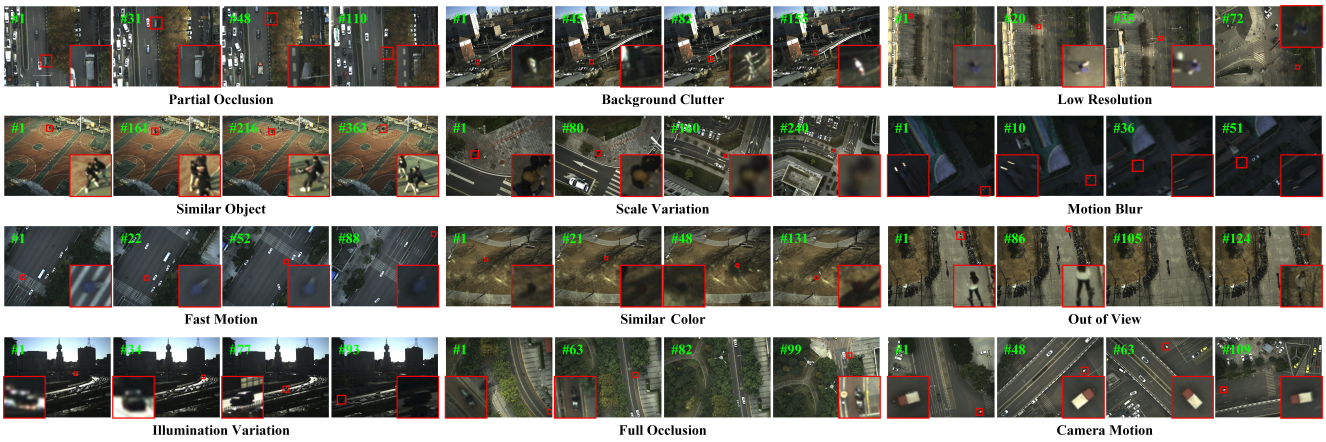

Figure 2. Typical examples of each challenge attribute in the MUST dataset, with the zoomed-in target areas.

of each attribute.

These challenge attributes present significant difficulties for traditional RGB-based trackers, whereas multispectral-based trackers offer more effective solutions. For instance, in the cases of Background Clutter and Similar Color, the target often shares similar appearance characteristics with surrounding objects and backgrounds, making them challenging to distinguish using RGB data alone. Moreover, in scenarios involving occlusion or small target sizes, the limited spatial features available in RGB images are insufficient to maintain robust tracking, leading to tracking drift. In contrast, multispectral data leverages spectral information that is inherently linked to the target's material composition, providing more stable and distinctive features throughout the tracking process. These attributes highlight the potential of multispectral data for improving UAV tracking performance in challenging scenarios.

2. Asymmetric Attention

Our proposed UNTrack employs an asymmetric attention structure to enable unified spectral-spatial-temporal feature extraction. The attention calculation process is illustrated in Fig. 3. Specifically, we model the relationship between embedded query $\boldsymbol{Q} = [\boldsymbol{Q}_{\mathrm{P}}; \boldsymbol{Q}_{\mathrm{T}}; \boldsymbol{Q}_{\mathrm{S}}]$, key $\boldsymbol{K} = [\boldsymbol{K}_{\mathrm{P}}; \boldsymbol{K}_{\mathrm{T}}; \boldsymbol{K}_{\mathrm{S}}]$, and value $\boldsymbol{V} = [\boldsymbol{V}_{\mathrm{P}}; \boldsymbol{V}_{\mathrm{T}}; \boldsymbol{V}_{\mathrm{S}}]$, generating the corresponding attention maps. These maps are divided into nine distinct blocks, each representing the interaction between different tokens:

(1) Self-attention on Prompt;
(2) Cross-attention on Prompt with Template;
(3) Cross-attention on Prompt with Search;
(4) Cross-attention on Template with Prompt;
(5) Self-attention on Template;
(6) Cross-attention on Template with Search;
(7) Cross-attention on Search with Prompt;
(8) Cross-attention on Search with Template;
(9) Self-attention on Search.

Recent research has shown that each block plays an inconsistent role in the tracking process [1, 3]. For example, blocks 1, 5, and 9 correspond to the self-attention results of the prompt, template, and search, respectively. These blocks capture deep semantic information that significantly enhances tracking performance. Moreover, the core of visual object tracking lies in accurately locating the target within search regions, a process that relies on effective cross-information aggregation in block 8. Given that the prompt encodes historical spectral information of the target, block 7, representing the interaction between the search and prompt, plays a crucial role in improving target discrimination, particularly in complex backgrounds.

In contrast, blocks 2 and 3 correspond to interactions between the prompt and other tokens, which tend to introduce noise and pollute the spectral information, thereby negatively affecting tracking accuracy. Blocks 4 and 6, which focus on the template, irrelevant to the generation of tracking results and only add unnecessary computational overhead. These blocks not only fail to aid in tracking, but also increase the computational cost without improving perfor-

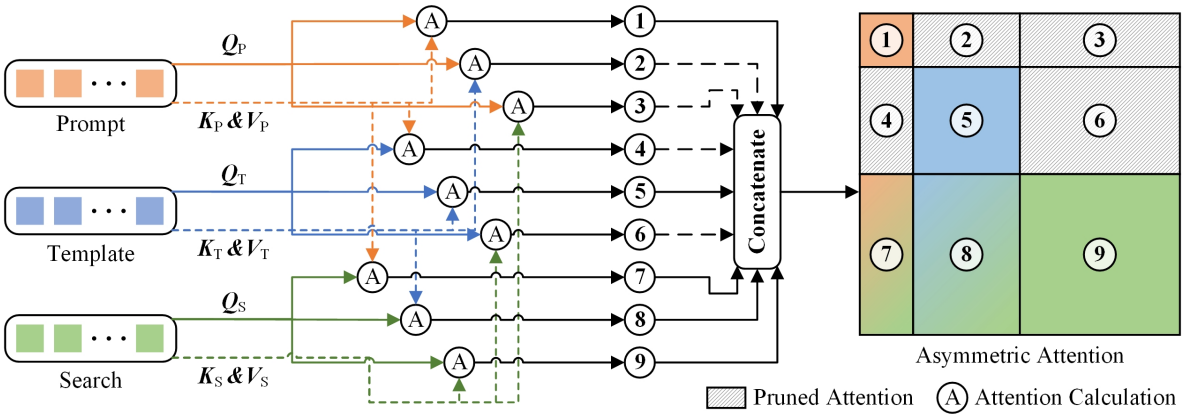

Figure 3. Illustration of asymmetric attention calculation.

Table 3. Detailed description of the challenge attributes.

Attribute	Description
POC	**Partial Occlusion**: The target is partially occluded by objects.
BC	**Background Clutter**: The background of the target is cluttered.
LR	**Low Resolution**: The target area is less than 100 pixels.
SOB	**Similar Object**: Similar objects exist around the target.
SV	**Scale Variation**: The target changes significantly in size or shape.
MB	**Motion Blur**: The movement of the target creates blur.
FM	**Fast Motion**: The distance between targets in two adjacent frames exceeds 20 pixels.
SC	**Similar Color**: The target and background are visually similar in color.
OV	**Out of View**: The target moves out of view and returns to view after a while.
IV	**Illumination Variation**: The illumination state of the environment around the target changes.
FOC	**Full Occlusion**: The target is completely occluded by other objects.
CM	**Camera Motion**: The UAV platform appears to be shaking or rotating.

mance. Consequently, we prune blocks 2, 3, 4, and 6 in the asymmetric attention structure, resulting in a more compact attention map that enhances tracking precision while reducing computational burden.

Table 4. Detailed description of the challenge attributes.

Infrared bands	AUC (%)	Pre (%)
Training from scratch	55.9	74.1
Replicating	59.7	79.2

3. Experiment Details

3.1. Parameter Reconstruction

In previous multispectral images processing, a prevalent practice adopts the input layer weights of pre-trained RGB-based networks to initialize those of multispectral-based networks [2, 6]. However, this approach ignores the differences between spectral bands, without considering the prior information from the spectral camera. In this work, we propose a simple yet effective parameter reconstruction strategy that enables the use of RGB-based pre-trained parameters for multispectral vision tasks.

Specifically, we assign a spectral band to each channel of the RGB-based parameters, following the definitions provided by the CIE standard [4], which correspond to the wavelengths: $\mathrm{Red} = 700.0\mathrm{nm}$, $\mathrm{Green} = 546.1\mathrm{nm}$, $\mathrm{Blue} = 435.8\mathrm{nm}$. Following previous works, we interpolate pre-trained RGB-based weights to initialize the input layer weights corresponding to visible bands. For infrared bands, we replicate the pre-trained RGB-based red channel weights, which are then assigned as the initial weights to each infrared band. The reconstruction process can be formalized as follows:

$$\boldsymbol{W}_{\mathrm{M}_i} = \begin{cases} \dfrac{\boldsymbol{W}_{\mathrm{B}}(\mathrm{G}-\mathrm{M}_i)+\boldsymbol{W}_{\mathrm{G}}(\mathrm{M}_i-\mathrm{B})}{\mathrm{G}-\mathrm{B}}, if \ \ \mathrm{M}_i < \mathrm{G} \\ \dfrac{\boldsymbol{W}_{\mathrm{G}}(\mathrm{R}-\mathrm{M}_i)+\boldsymbol{W}_{\mathrm{R}}(\mathrm{M}_i-\mathrm{G})}{\mathrm{R}-\mathrm{G}}, if \ \ \mathrm{G} < \mathrm{M}_i \leq \mathrm{R} \\ \boldsymbol{W}_{\mathrm{R}}, if \ \ \mathrm{R} < \mathrm{M}_i \end{cases} \tag{1}$$

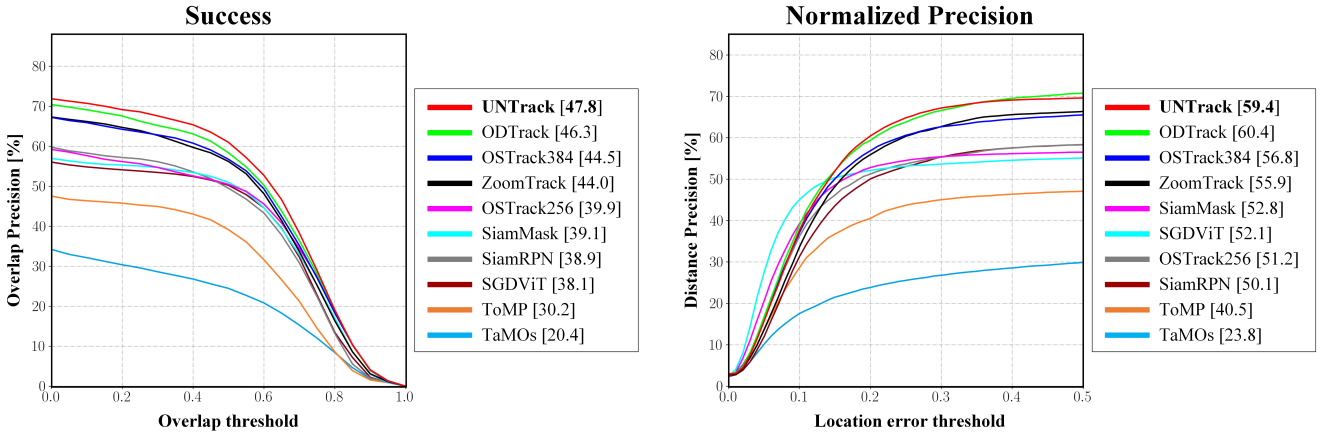

Figure 4. Success and Normalized Precision plots for state-of-the-art trackers on the MUST dataset.

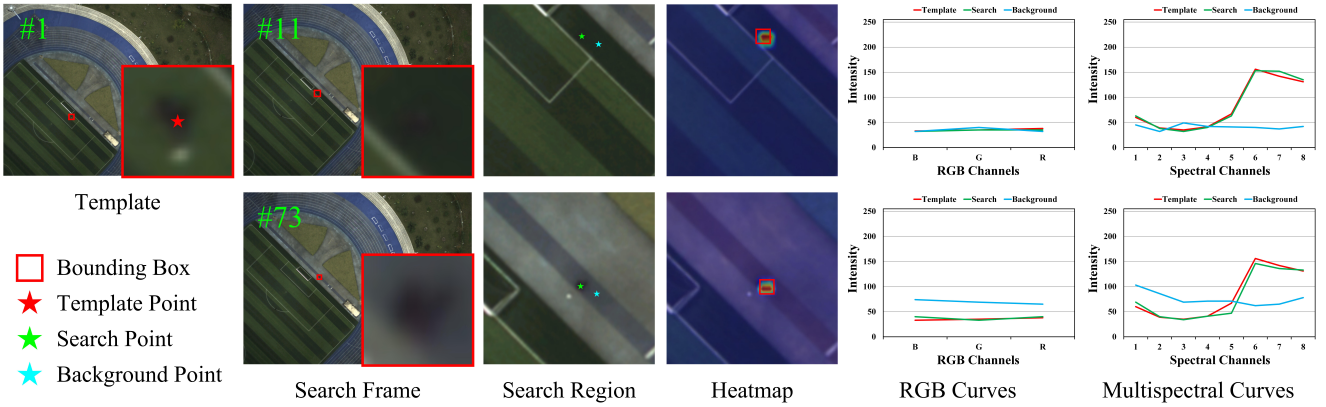

Figure 5. Performance of the proposed UNTrack in scenarios with similar target and background colors. We sample points for template, search, and background during tracking and visualize their spectral curves.

where $\mathrm{R}, \mathrm{G}, \mathrm{B}$ and $\boldsymbol{W}_{\mathrm{R}}, \boldsymbol{W}_{\mathrm{G}}, \boldsymbol{W}_{\mathrm{B}}$ represent the spectral bands and weights corresponding to the red, green, and blue channels in the RGB space. M_i and $\boldsymbol{W}_{\mathrm{M}_i}$ denote the i-th spectral band and the corresponding channel weights in MUST, where $i = 1, \ldots, 8$.

Notably, we compared the performance of different infrared bands initialization approaches. As shown in Sec. 2, we evaluate initializing infrared bands by training from scratch instead of replicating. Without utilizing pre-trained RGB-based weights as prior information, training infrared bands from scratch reduces AUC by 3.8%.

Our proposed initialization strategy is universally applicable, for example, it improves AUC of OSTrack_{256} by 12.4%, UNTrack by 11.9% and ZoomTrack by 9.3%. The exception is ODTrack, Which embeds video-clip input as the template but neglects spectral redundancy, resulting in limited improvement. ODTrack's inherent flaw leads to a shift in suboptimal methods, from ODTrack to OSTrack_{384}, before and after initialization. Consequently, UNTrack has performance discrepancy when compared with two suboptimal methods before and after initialization.

3.2. Comparative Analysis

To further evaluate the performance of our UNTrack against state-of-the-art trackers, we present the Success Plot and Normalized Precision Plot on the MUST dataset in Fig. 4. As shown, UNTrack outperforms the runner-up, OSTrack [5], with a performance gain of 1.5% in success rate, and demonstrates superior performance compared to other trackers. This improvement is attributed to UNTrack's comprehensive utilization of spectral information, which enables it to effectively address complex tracking challenges.

3.3. Visualization Analysis

To highlight the potential of multispectral-based tracking, we visualize UNTrack's performance on a challenging scenario characterized by the Similar Colors challenge attribute. As shown in Fig. 5, when the target and background exhibit similar colors, traditional RGB-based data fail to

distinguish them due to the lack of intensity differences across the RGB channels, leading to poor performance. In contrast, the spectral curves of the target and background exhibit distinct differences, with the target's spectral curve remaining stable throughout the tracking process. Leveraging this characteristic, UNTrack maintains accurate tracking by focusing on the target, even in situations where color similarity complicates tracking in the RGB space.

References

[1] Yutao Cui, Cheng Jiang, Limin Wang, and Gangshan Wu. Mixformer: End-to-end tracking with iterative mixed attention. In *CVPR*, pages 13608–13618, 2022. 2

[2] Zhenqi Liu, Xinyu Wang, Yanfei Zhong, Meng Shu, and Chen Sun. Siamhyper: Learning a hyperspectral object tracker from an rgb-based tracker. *IEEE TIP*, 31:7116–7129, 2022. 3

[3] Zikai Song, Run Luo, Junqing Yu, Yi-Ping Phoebe Chen, and Wei Yang. Compact transformer tracker with correlative masked modeling. In *AAAI*, pages 2321–2329, 2023. 2

[4] Andrew Stockman. Cone fundamentals and cie standards. *Current Opinion in Behavioral Sciences*, 30:87–93, 2019. 3

[5] Botao Ye, Hong Chang, Bingpeng Ma, Shiguang Shan, and Xilin Chen. Joint feature learning and relation modeling for tracking: A one-stream framework. In *ECCV*, pages 341–357. Springer, 2022. 4

[6] Jia-ju Ying, Shi-qu Yin, Wei-yong Yang, Hui Liu, and Xu Li. Hyperspectral image target recognition based on yolo model. In *Sixth Conference on Frontiers in Optical Imaging and Technology: Imaging Detection and Target Recognition*, pages 221–227. SPIE, 2024. 3