

# Tracking Tiny Drones against Clutter: Large-Scale Infrared Benchmark with Motion-Centric Adaptive Algorithm

Jiahao Zhang<sup>1,2</sup>, Zongli Jiang<sup>1</sup>, Jinli Zhang<sup>1\*</sup>, Yixin Wei<sup>1</sup>, Liang Li<sup>2</sup>, Yizheng Wang<sup>2</sup>, Gang Wang<sup>2\*</sup>

<sup>1</sup>College of Computer Science, Beijing University of Technology, Beijing, China

<sup>2</sup>NAIVE Lab, Brain Research Center, Beijing Institute of Basic Medical Sciences, Beijing, China

## Abstract

*Tracking flying drones in infrared videos is a crucial yet challenging task. Existing drone trackers and datasets have limitations in dealing with and characterizing tiny targets ( $\leq 20 \times 20$  pixels) against highly complex backgrounds. To tackle this issue, we have developed a large-scale benchmark for tiny drone tracking in infrared videos (TDTIV), which comprises 290k frames and 280k manually annotated bounding boxes. Unlike traditional trackers that primarily rely on appearance matching, we introduce a novel method called Motion-Centric Adaptive Tracking (MCATrack), which initially employs a magnocell-inspired motion response to enhance the local signal-to-noise ratio of tiny target regions while suppressing complex clutter. Moreover, we design a Dynamic Cross-Guided module that integrates both initial and updated target features to address pose variations in long-term tracking. This module captures the latest target information to generate highly relevant candidate regions and refines them through precise optimization to achieve more accurate tracking results. Extensive experiments performed on the TDTIV and the well-recognized Anti-UAV 410 datasets have demonstrated the superiority of MCATrack over state-of-the-art competing trackers. Code and dataset are available at <https://github.com/zhangjiahao02/MCATrack>.*

## 1. Introduction

While drones (*a.k.a.* UAVs) have gained increasing use in a wide range of fields, concerns have been raised about safety and privacy. Drones may intrude into sensitive areas, compromising personal privacy. Particularly in cases of inadequate regulation, drones could even be exploited for illicit surveillance, communication network attacks, and airport invasions. Monitoring drone activities, including their location and trajectory, has therefore become extremely important, especially for air administration and public secu-

rity sections. Infrared cameras have been widely used in air traffic management, as they can detect the infrared spectral characteristics of flying objects. However, since civil drones are typically small in size and often fly in dynamic environments, tracking these tiny drones in infrared videos remains a highly complex and challenging task.

Existing drone tracking datasets [22, 23] are constrained by their focus on large target sizes and simple backgrounds. To address this, we introduce TDTIV, a dataset that leverages infrared cameras to capture drones at long distances, tracking their movements across diverse and complex environments. As shown in Figure 1 (a), the left side compares TDTIV with Anti-UAV410: while Anti-UAV410 uses simple backgrounds, TDTIV presents diverse environments, such as forests, buildings, and hills, for enhanced tracking difficulty. The right side displays UAV area distribution: TDTIV's red curve peaks at lower ratios, indicating smaller UAVs and higher detection difficulty, whereas Anti-UAV410 shows a broader distribution covering larger UAV instances. Furthermore, to enhance dataset realism, we filtered out videos with overly large targets or simplistic backgrounds, ensuring that TDTIV serves as a more challenging and representative benchmark for UAV tracking.

In video-based tracking, motion and temporal information are crucial for accurate target localization. Motion captures dynamic properties such as displacement, velocity, and direction changes between frames, while temporal information reflects the target's continuity and behavioral patterns. Integrating both aspects improves tracking accuracy by enhancing motion trajectory prediction and state estimation. However, many existing trackers rely heavily on template matching, focusing only on spatial features within the search region while overlooking motion and temporal cues. As illustrated in Figure 1 (b), in frame 131, the target stands out against the background, allowing SiamDT [19] to track it reliably using spatial features alone. By frame 176, the target blends into the background, causing SiamDT to mistakenly track building lights. In contrast, our method leverages motion cues to enhance the target contrast against the background, ensuring stable tracking even in complex

\*Corresponding authors: g\_wang@foxmail.com, lz73798@gmail.com

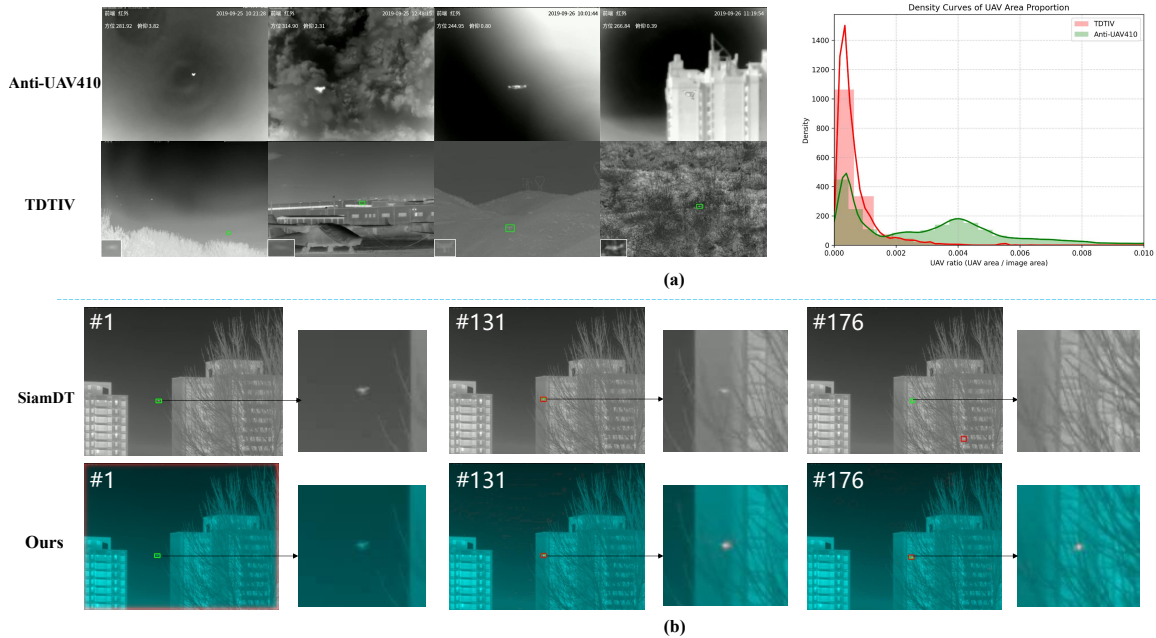


Figure 1. (a) Left: Visual comparison of background characteristics between our dataset and Anti-UAV410 (showing representative images). Right: UAV area proportion density curve relative to total image area. (b) Comparison of tracking performance between our method and SiamDT under varying contrast conditions. Zoomed-in regions: 50-pixel window centered on the ground truth box. Red indicates the predicted bounding box, and green indicates the ground truth.

environments. This over-reliance on template matching and neglect of motion and temporal cues reduces their effectiveness in dynamic scenarios, such as fast movements, occlusions, or complex backgrounds, making them less suitable for real-world applications.

To address this issue, we propose a Motion-Centric Adaptive Tracking (MCATrack) method for tiny drone tracking. Firstly, we introduce the Bio-inspired Magnomotion Enhancement module, inspired by the biological magnocellular pathway, which utilizes motion responses to enhance the local signal-to-noise ratio in small target regions while suppressing complex backgrounds. To tackle challenges such as significant pose variations caused by target flipping, drastic lighting changes, and dynamic occlusions during tracking, we further propose the Dynamic Cross-Guided module. This module operates in three stages: (1) The Dynamic Target Cross-Guidance Module uses a Siamese-branching Region Proposal Network (RPN) to model the search region by exploiting the relationship between the updated and initial targets, generating high-quality candidate proposals. (2) The Enhanced Feature Confidence Generation module refines the predicted proposals using enhanced target information, aiming to improve confidence accuracy. Unlike traditional Siamese trackers [38, 56], which rely exclusively on the initial template to generate and refine candidate boxes, our approach

integrates updated target information with the initial template. This integration facilitates more precise candidate box generation and refinement, thereby enhancing confidence accuracy. (3) The bounding box with the highest confidence score is selected as the final tracking result. Additionally, to ensure the tracking model adapts in real time to dynamic target changes, we employ a carefully designed strategy to update the dynamic template. This strategy ensures robust and accurate tracking of the target in complex and variable scenarios.

By integrating the Magnomotion module and the Dynamic Cross-Guided module, MCATrack effectively addresses the challenges of long-term tracking for small targets in complex backgrounds. Extensive experiments were conducted to evaluate its performance and demonstrate its superiority. This work makes the following contributions:

- We introduce the TDTIV, a large-scale tracking dataset contains tiny drone targets against various complex backgrounds to advance the research.
- We propose the MCATrack, a motion-centric adaptive tracker that leverages the Magnomotion module to enhance the tiny targets, while integrating the Dynamic Cross-Guided module for dynamic global target search.
- The MCATrack outperforms state-of-the-art competing trackers on the large-scale TDTIV and Anti-UAV410 datasets.

## 2. Related Work

### 2.1. Target Tracking

Recent advances in visual object tracking have been driven by benchmark datasets [12, 13, 20, 24, 32, 42] and sophisticated algorithms [7, 9, 10, 21, 25, 26, 35, 43, 53]. Early Siamese trackers [1, 2, 25] employed CNN-based feature extraction [17], followed by lightweight relational networks for template-search interactions. This approach, however, could only support simple unidirectional information exchange from the template to the search region, limiting cross-region interaction. The advent of transformers [37] replaced these lightweight networks in frameworks such as [7, 50], enabling bidirectional interactions through self-/cross-attention layers. However, these frameworks often suffered from speed bottlenecks due to the computational complexity of attention mechanisms. Subsequent works [49, 53] unified feature extraction and relation modeling in a pipeline, improving tracking with lower computational costs. However, existing methods continue to prioritize template-search matching, which remains inadequate in addressing the unique challenges of drone tracking, such as small targets, motion blur/occlusion, and high background similarity. Our Magno-motion module overcomes these limitations by employing short-term memory recursion, thereby enhancing spatio-temporal feature coherence.

### 2.2. Drone Tracking

Anti-UAV tracking aims to monitor and mitigate security threats posed by drones. Existing tracking frameworks often rely on Siamese networks. For example, Jiang et al. [23] introduced the Dual-Flow Semantic Consistency Tracker (DFSC), which improves feature discrimination by modulating semantic flow, while Zhang et al. [54] developed SiamFusion to combine multimodal data for better target perception. Huang et al. [19] proposed SiamDT, a multi-branch framework that integrates dual-semantic extraction and background interference suppression to handle dynamic backgrounds. Despite these advances, existing algorithms struggle to track small targets over extended periods, especially with changes in posture. Although methods like SiamDT improve tracking with dynamic background updates, they still face challenges in capturing long-term target information. To address this, we propose a dynamically updated target representation approach to enhance the tracking of the target's latest state.

### 2.3. Drone Tracking Dataset

In recent years, the development of anti-UAV tracking has been supported by several datasets. Hui et al. [22] introduced a dataset for fixed-wing drone tracking with 16,177 frames to address data scarcity. Jiang et al. [23] proposed a multimodal dataset with over 300 pairs of visible light

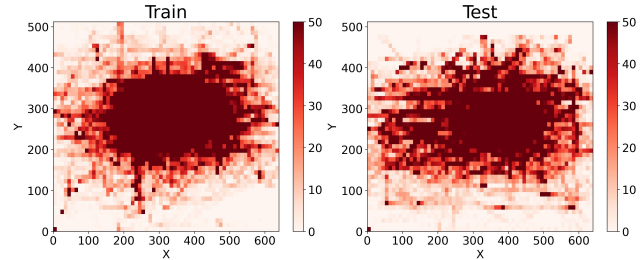


Figure 2. Position distribution in the TDTIV dataset.

and infrared videos covering various environments, significantly advancing anti-UAV tasks. The Anti-UAV410 dataset [19] expanded on this by adding diverse, complex scenarios and focusing on tiny drones to enhance data support. However, most existing datasets primarily feature large-scale targets with oversimplified backgrounds, which do not fully represent real-world conditions. Despite the inclusion of tiny drones in Anti-UAV410, the coverage remains insufficient for practical applications. To address this gap, we propose the TDTIV dataset, offering a comprehensive benchmark for training and evaluating anti-UAV tracking algorithms.

## 3. TDTIV Dataset

As aforementioned, existing datasets for tracking drones have limitations in characterizing tiny flying objects against diverse complex backgrounds. To address this issue, we collected tiny drone videos by observing the targets at a remote distance using an advanced infrared camera. The field-of-view of the infrared camera ranges from -15 degrees to 90 degrees in the vertical direction, while the target slope distance spans from 200 meters to 2000 meters. We selected diverse scenarios including woods, buildings, clouds, and mountains. Following the data paradigm of Anit-UAV (v1) and Anti-UAV410 (v2), we set the video frame resolution as either 640×480 pixels or 640×512 pixels. There is little appearance difference between the targets in videos with the two kinds of resolution. Besides, the video frame rate was maintained at 25 frames per second (FPS). Among the preliminary raw video data (900k frames), we removed the videos containing either non-tiny-size drones or non-complex backgrounds, thereby obtaining the dataset consisting of 260 video sequences (approximately 290k frames).

During the target annotation procedure, we used bounding boxes to manually annotate the 260 infrared video frames. The bounding box is marked in the form of upper left corner and lower right corner coordinates  $(x_1, y_1, x_2, y_2)$ . Each image corresponds to an annotation file in XML format, containing the target category and coordinates. Consequently, there are 280k manually annotated

target bounding boxes in the dataset. For each video sequence, the bounding box annotation of the first frame is initially given.

To make TDTIV dataset easy to deploy with benchmark, we built a website to introduce the details of the research topic backgrounds and the datasets, providing the community with standardized training/test data with examples and demos.

Note that the established TDTIV dataset has superiorities compared with the existing anti-drone or anti-UAV datasets. Our TDTIV dataset contains drone targets of various sizes, most ranging from 5×5 to 30×30 pixels. Statistical analysis suggests that most of the targets in TDTIV are smaller than 20×20 pixels, inevitably increasing the difficulty of visual tracking. Table 1 compares TDTIV with benchmark UAV datasets, showing that our dataset achieves the smallest dimensions (28×16 px) and lowest area ratio of 0.21%. In addition, we elaborately divided the full video data into the training set and test data based on different recording locations and times, ensuring that nearly 80% of the scenes in the test set do not overlap with those in the training set. Figure 2 shows UAV position distribution in TDTIV, revealing significant trajectory diversity across the training and test sets. Targets are uniformly dispersed without dominant clusters, achieving omnidirectional motion scenario coverage. Therefore, the evaluation results in terms of performance robustness and model generalization will be reliable.

Dataset	Avg. UAV Size	Avg. UAV Area Ratio
Anti-UAV	IR:52×30	IR: 0.48%
Anti-UAV410	39×24	0.35%
TDTIV	28×16	0.21%

Table 1. Size comparison of the TDTIV dataset with related UAV detection and tracking benchmarks.

## 4. Methodology

### 4.1. Overview

To address the tiny drone tracking against clutter, we propose the MCATrack, a novel motion-centric global tracking framework that systematically integrates motion perception and dynamic feature guidance. The architecture figures two innovative components: (1) A Bio-inspired Magno-motion Enhancement module that suppresses static background interference through spatiotemporal motion response modeling, effectively amplifying target-background contrast by simulating biological vision mechanisms; (2) A Dynamic Cross-Guided module that adaptively updates target representations using temporal context, specifically designed to resolve pose variation and occlusion challenges through

a cross-attention mechanism. The overall architecture of MCATrack is illustrated in Figure 3. Subsequent sections will provide comprehensive technical details including implementation specifics and theoretical analysis.

### 4.2. Bio-inspired Magno-motion Enhancement

Visual motion processing plays a fundamental role in both biological and artificial vision systems. In biological systems, most species have evolved dedicated neural pathways for motion analysis, while in computer vision, motion features have also been widely adopted for tasks like object detection [41, 52] and activity recognition [40]. Classical computational approaches for extracting such motion features include optical flow estimation [36, 48] and frame differencing. However, these methods face limitations: Optical flow algorithms often require intensive computation and exhibit sensitivity to illumination changes and noise, whereas frame differencing struggles to capture precise motion patterns under dynamic backgrounds or camera movements.

The biological visual pathway offers insights for addressing these challenges. In the mammalian retina, a multi-layered cellular architecture that comprises photoreceptors, bipolar cells, amacrine cells, and ganglion cells enables efficient motion processing [39, 44]. Signals pass through photoreceptors, which enhance dark contrast and balance downstream processing before transmitting to the outer plexiform layer: here, horizontal cells suppress spatio-temporal noise, while bipolar cells sharpen edges, with these processed signals then relayed to ganglion cells. Ganglion cells collaborate with bipolar and amacrine cells to process visual information, ultimately projecting coded signals to the lateral geniculate nucleus or primary visual cortex.

Notably, retinal processing bifurcates into two parallel pathways: the magnocellular pathway specializes in transient motion detection, while the parvocellular pathway analyzes static spatial details [34]. Such bio-inspired mechanisms have motivated novel neural network architectures, though the simulation accuracy remains unsolved compared to biological systems. Specifically, in the magnocellular pathway of the inner plexiform, the amacrine cell has one end connected with the bipolar cell, and the other end connected with the ganglion cell. Amacrine cells act as a temporal high-pass filter bank that enhances the temporal and spatial variation. In discrete digital video processing, this high-pass filter bank can be considered as a difference equation. The transfer function in terms of  $\mathcal{Z}$ -transform is:

$$\begin{aligned} \text{Ama}(z) &= \omega \cdot \frac{1 - z^{-1}}{1 - \omega \cdot z^{-1}} \\ \omega &= e^{-\Delta t/\tau} \end{aligned} \quad (1)$$

where  $\Delta t = 1$  stands for the discrete time step, while  $\tau$

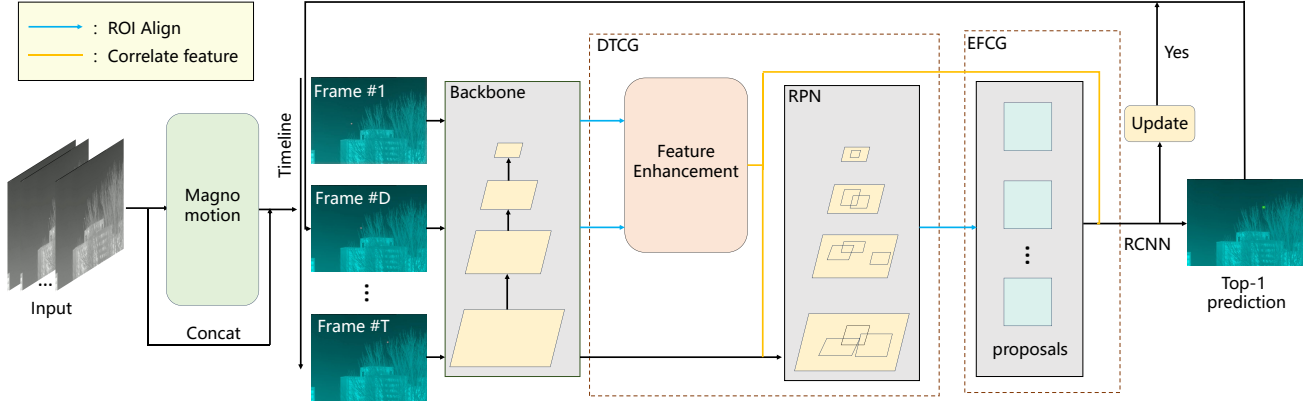


Figure 3. The overall architecture of MCATrack consists of two modules: a Bio-inspired Magno-motion Enhancement module, which enhances target contrast against the background, and a Dynamic Cross-Guided module, which performs a global search and generates the final prediction. In the Magno-motion module, we combine the extracted motion information with the original image to reduce noise interference. During tracking, we use dynamic object features to enhance the initial target features, leading to more accurate predictions.

denotes a time constant of the computation model that is empirically set to be  $\tau = 2$ .

Based on the elaborations above, the yielded responses of the bio-inspired magno-motion module depends on both the current input and the prior output, akin to the iterative nature of difference equation solutions. To mitigate motion artifacts induced by camera displacement, spatial-temporal alignment of adjacent-frame cuboids through inter-frame registration are empirically essential. This preprocessing step ensures feature consistency across sequential frames. It is worth noting that the Magno-motion enhancement module leverages temporal coherence by deriving motion information from multi-frame observations. This scheme can inherently suppress random noise through statistical averaging across consecutive frames, while preserving salient motion patterns.

### 4.3. Dynamic Cross-Guided (DCG)

**Dynamic Target Cross-Guidance (DTCG).** Existing Siamese trackers [21] typically rely on an initial template to guide the search region for generating candidate proposals. However, when the target undergoes flipping or rapid movement, small targets may exhibit significant pose variations, complicating precise tracking. While template updating is critical for maintaining tracking accuracy, improper updates can result in error accumulation. To address this issue, we propose a Target Feature Enhancement Module that aggregates the latest target information from the dynamic template and fuses it with the initial features, thereby enhancing the representation of the target region.

Specifically, let  $z \in \mathbb{R}^{k \times k \times c}$  denote the Region Of Interest (ROI) features of the initial template image,  $d \in \mathbb{R}^{k \times k \times c}$  denote the ROI features of the dynamic template image, and  $x \in \mathbb{R}^{h \times w \times c}$  denote the search image features. We employ

cross-attention [37] to aggregate features:

$$\begin{aligned} F &= \text{Softmax} \left( \frac{QK^T}{\sqrt{C}} \right) \cdot V \\ &= \text{Softmax} \left( \frac{(\text{Conv}_q(z))(\text{Conv}_k(d))^T}{\sqrt{C}} \right) \cdot (\text{Conv}_v(d)) \end{aligned} \quad (2)$$

where  $\text{Conv}_q$ ,  $\text{Conv}_k$ ,  $\text{Conv}_v$  denote the convolution layers with a  $1 \times 1$  convolution kernel size, which are used to generate the query, key and value, respectively.  $F$  represents the fusion features of the ROI region of the initial and dynamic template. Then, we use it to update the initial target feature:

$$z_r = \gamma F + z \quad (3)$$

$\gamma$  is a learnable scalar parameter. By integrating the fusion features and the initial ROI region features of the template, we obtain enhanced features  $z_r$  containing the latest target information.

Next, we use the enhanced target features to guide the search region features, encoding the correlation between  $z_r$  and  $x$  as  $\varphi(x, z_r) \in \mathbb{R}^{h \times w \times c}$ :

$$\varphi(x, z_r) = \text{Conv}_{\text{out}}(\text{Conv}_x(x) \otimes \text{Conv}_z(z_r)) \quad (4)$$

Here,  $\otimes$  denotes the depth-wise convolution operator. We define a  $k \times k$  convolutional layer named  $\text{Conv}_z$  with zero padding, which transforms  $z_r$  into a  $1 \times 1$  convolutional kernel.  $\text{Conv}_x$  represents the application of a  $3 \times 3$  convolution layer with 1-pixel padding to the feature  $x$ .  $\text{Conv}_{\text{out}}$  refers to a  $1 \times 1$  convolution layer, which aims to adjust the number of channels back to  $c$ .

Since  $\varphi(x, z_r)$  retains the size of  $x$ , we can reuse directly the RPN module to perform subsequent operations to generate proposals. For training the Dynamic Target

Cross-Guidance module, we employ the same losses as those of RPN [16], where the classification and regression losses  $L_{\text{cls}}$  and  $L_{\text{reg}}$  are binary cross - entropy and smooth L1, respectively. The total loss for Dynamic Target Cross-Guidance is:

$$L_{\text{dtc}} = L_{\text{rpn}}(\varphi(x, z_r)) = \frac{1}{N_{\text{cls}}} \sum_i L_{\text{cls}}(p_i, p_i^*) + \lambda \frac{1}{N_{\text{reg}}} \sum_i p_i^* L_{\text{reg}}(b_i, b_i^*) \quad (5)$$

where  $p_i^*$  and  $p_i$  denote the ground truth label and predicted score of the  $i$ th proposal,  $b_i^*$  and  $b_i$  the coordinates of the  $i$ th ground truth box and the  $i$ th predicted box. The weight  $\lambda = 1$  is used to balance classification and regression losses.

**Enhanced Feature Confidence Generation (EFCG).** Similar to Dynamic Target Cross-Guidance networks, we utilize the enhanced target region feature  $z_r \in \mathbb{R}^{k \times k \times c}$ . Specifically, let  $x_i \in \mathbb{R}^{k \times k \times c}$  denote the  $i$ th proposal feature. The  $i$ th correlation feature  $\varphi_i(x_i, z_r)$  encoded by the enhanced target region feature is calculated as follows:

$$\varphi_i(x_i, z_r) = \text{Conv}'_{\text{out}}(\text{Conv}'_{\text{x}}(x_i) \odot \text{Conv}'_{\text{z}}(z_r)) \quad (6)$$

where  $\odot$  denotes the Hadamard product,  $\text{Conv}'_{\text{x}}$  and  $\text{Conv}'_{\text{z}}$  represent feature projections for  $x_i$  and  $z_r$ ,  $\text{Conv}'_{\text{out}}$  is a  $1 \times 1$  convolution layer used to adjust the number of output channels to  $c$ .

Since the correlation features of the  $i$ -th proposal have the same size as the original features, we follow the conventional RCNN procedures, performing classification and regression on the proposals to derive the final predictions. In optimizing our model, we use binary cross-entropy and smooth L1 as classification and regression losses, similar to the Dynamic Target Cross-Guidance approach. The total loss for the Enhanced Feature Confidence Generation module is:

$$L_{\text{efc}} = \frac{1}{N_{\text{prop}}} \sum_i L_{\text{rcnn}}(\varphi_i(x_i, z_r)) = \frac{1}{N_{\text{prop}}} \sum_i (L_{\text{cls}}(p_i, p_i^*) + \lambda p_i^* L_{\text{reg}}(b_i, b_i^*)) \quad (7)$$

where  $N_{\text{prop}}$  represents the number of proposals.  $p_i^*$  and  $p_i$  denote the ground truth label and the prediction confidence scores.  $b_i^*$  and  $b_i$  represent the ground truth and the predicted bounding boxes, respectively. The weight  $\lambda = 1$  is used to balance classification and regression losses.

## 5. Experiments

To validate the effectiveness of our proposed method for tracking tiny drones, we compared MCATrack with the latest state-of-the-art tracking models on large-scale datasets, including TDTIV and Anti-UAV410.

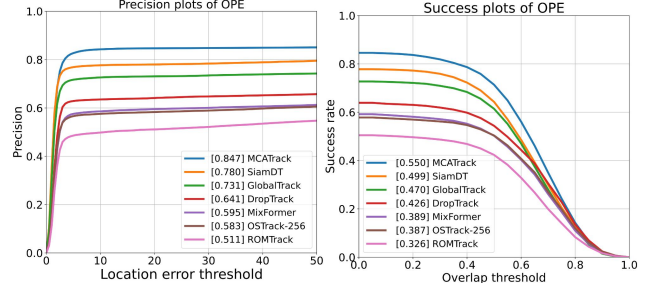


Figure 4. Precision and success plots of MCATrack and state-of-the-art trackers obtained on the TDTIV dataset.

### 5.1. Experimental Settings

The experiment was carried out on a platform based on the Ubuntu OS, with the algorithm implemented on a NVIDIA GeForce RTX 4090. We employ ResNet50 as the backbone, with the channel dimension of the characteristic network set at  $c = 256$ . We set the IoU thresholds for positive and negative samples at 0.7 and 0.3, respectively. To extract ROI features, we adopt ROI Align [18] with an output feature size of  $k=7$ . In the inference phase, the target image is updated when the tracking confidence in consecutive frames is  $\geq 0.95$  and the center displacement is  $\leq 20$  pixels.

### 5.2. Evaluation Metrics

Three evaluation metrics are used to assess trackers through One-Pass Evaluation (OPE). The precision plot calculates the percentage of frames where the center distance between the predicted and ground truth bounding boxes is below varying pixel thresholds, with a 20-pixel threshold used for ranking. The success plot measures the percentage of frames where the IoU between the predicted and ground truth bounding boxes exceeds a threshold (ranging from 0 to 1), with the AUC used for ranking. State Accuracy (SA) [23] reflects the average overlap ratio between the predicted bounding boxes and the ground truth across all sequences. This metric requires accurate prediction of target position and scale, assessment of presence/absence in the current frame, and detection of complete disappearance events.

### 5.3. Experiments on TDTIV

To demonstrate the superior performance of our proposed method in anti-UAV tracking, we selected 20 state-of-the-art tracking models recently introduced for comparative analysis. We evaluated these models using the established TDTIV dataset to obtain their performance results.

According to the results presented in Table 2, the MCATrack method achieves the highest SA score of 54.97. Compared to the latest 2024 drone tracking method, SiamDT, MCATrack shows a significant performance improvement, surpassing it by more than 5%. It also outperforms recent

Methods	Source	SA
GlobalTrack[21]	AAAI20	46.98
Stark-ST50[50]	ICCV21	41.16
Stark-ST101[50]	ICCV21	37.98
TOMP50[30]	CVPR22	37.19
TOMP101[30]	CVPR22	39.49
SimTrack[6]	ECCV22	32.53
OTrack-256[53]	ECCV22	38.81
OTrack-384[53]	ECCV22	44.62
MixFormer[9]	CVPR22	38.97
GRM[15]	CVPR23	37.99
DropTrack[46]	CVPR23	42.70
ROMTrack[4]	ICCV23	32.66
ARTrack[45]	CVPR23	36.29
ODTrack[55]	AAAI24	35.37
EVPTTrack[33]	AAAI24	39.58
TaMOs-ResNet50[31]	WACV24	32.28
TaMOs-Swin[31]	WACV24	36.70
AQATrack[47]	CVPR24	41.59
AVTrack[27]	ICML24	36.60
SiamDT[19]	TPAMI24	49.89
<b>MCATrack</b>		<b>54.97</b>

Table 2. Comparison of State Accuracy (SA) for MCATrack and state-of-the-art trackers on the TDTIV dataset.

transformer-based tracking algorithms, such as OTrack and DropTrack, with improvements exceeding 10%, making it stand out in the field of tracking algorithms. To further illustrate the enhancements, Figure 4 visualizes the precision and success curves on the TDTIV dataset. MCATrack achieves absolute gains of 6.7% in precision and 5.1% in success rate compared to the previous best tracker, SiamDT, which confirms its effectiveness for tracking tiny drones.

#### 5.4. Ablation Studies

To validate the effectiveness of our MCATrack algorithm in different components, we conducted extensive ablation experiments and performed detailed analyses. In these experiments, we paid particular attention to ensuring the consistency of hyperparameters during the training process to a fair comparative analysis. All experiments were evaluated in the TDTIV test set, and we used State Accuracy (SA) as the primary metric to assess tracking performance.

**Effect of Magno-motion.** In this section, we explore the effectiveness and contributions of Magno-motion. The results are shown in Table 3. Specifically, incorporating motion information from consecutive frames, we observed a significant 7.13% improvement in performance. This motion module extracts target motion information over multiple frames and combines it with the original image, enhancing the contrast between the target and the background. The

Motion information	Feature Enhancement	SA
-	-	46.98
✓	-	54.11
-	✓	47.60
✓	✓	54.97

Table 3. An analysis of the effectiveness of motion information and feature enhancement in dynamic scenes.

Methods	Source	SA
GlobalTrack[21]	AAAI20	66.45
PrDiMP50[11]	CVPR20	54.69
ROAM[51]	CVPR20	43.03
Siam R-CNN[38]	CVPR20	63.00
SiamBAN[8]	CVPR20	47.32
KYS[3]	ECCV20	44.90
KeepTrack[29]	ICCV21	56.80
Stark-ST101[50]	ICCV21	57.15
AiATrack[14]	ECCV22	59.56
OTrack-256[53]	ECCV22	49.56
OTrack-384[53]	ECCV22	60.08
ToMP50[30]	CVPR22	55.09
ToMP101[30]	CVPR22	55.10
TCTrack[5]	CVPR22	41.64
SwinTrack-Tiny[28]	NIPS22	53.15
SwinTrack-Base[28]	NIPS22	55.74
DropTrack[46]	CVPR23	60.15
ROMTrack[4]	ICCV23	46.81
MixFormerV2[10]	NIPS24	59.65
SiamDT[19]	TPAMI24	68.19
<b>MCATrack</b>		<b>69.18</b>

Table 4. Performance comparison of MCATrack and state-of-the-art methods on the Anti-UAV410 dataset.

analysis suggests that, due to the small size of drone targets and their susceptibility to occlusion in complex backgrounds, this motion-based enhancement proves effective in boosting UAV tracking performance. This improvement is particularly beneficial for tracking targets in complex scenarios, especially tiny drone targets.

**Effect of Feature Enhancement.** We conducted experiments to validate the impact of the Feature Enhancement module on tracking performance, as shown in Table 3. By incorporating the Feature Enhancement module, our method achieved a 0.62% improvement in the TDTIV dataset. This improvement underscores the crucial role of using the most recent target information, particularly in scenarios where pose variations, complex backgrounds, and changing lighting conditions cause the target to deviate substantially from the initial template. By continuously updat-

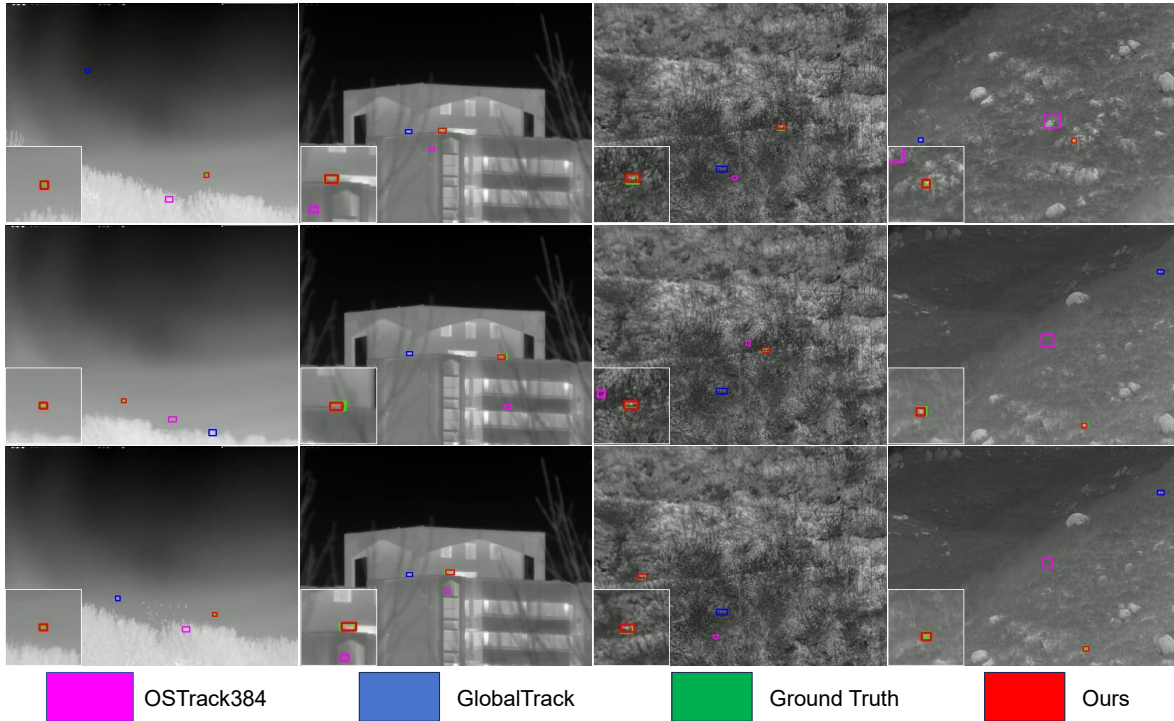


Figure 5. Qualitative comparisons with other trackers on the TDTIV dataset show our approach outperforms in tracking tiny drones in complex backgrounds. Each image’s lower-left region shows a 100-pixel range centered on the ground-truth box.

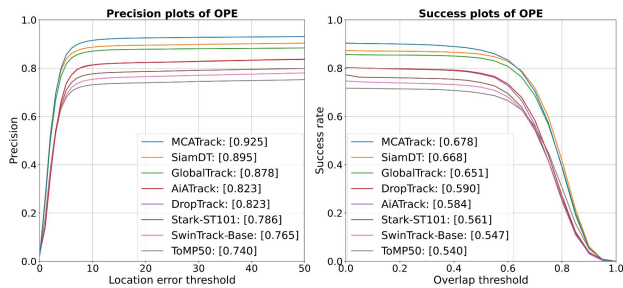


Figure 6. Precision and success plots of MCATrack and state-of-the-art trackers on the test set of Anti-UAV410.

ing and integrating the latest target information, the Feature Enhancement module enables more accurate position estimation, compensating for discrepancies caused by such dynamic factors and enhancing overall tracking performance.

### 5.5. Experiments on Anti-UAV410

To further validate the generalization ability of the MCA-Track method in small UAV target tracking, we evaluated its tracking performance on the Anti-UAV410 dataset using State Accuracy (SA).

As shown in Table 4, our method ranks first in the Anti-UAV410 dataset, achieving an improvement in accuracy of

more than 10% over most competing algorithms. Compared to the previous best tracker, SiamDT [19], our method demonstrates an improvement of 1%. The precision and success rate plots for the Anti-UAV410 dataset are shown in Figure 6. Compared to SiamDT, our method achieves improvements of 3% in precision and 1% in success. These improvements further demonstrate the effectiveness of our approach for UAV target tracking.

## 6. Conclusion

To address the challenge of tracking tiny drones in infrared videos against complex backgrounds, we have constructed a large-scale TDTIV dataset to advance research in this field. The dataset encompasses diverse real-world scenarios, providing researchers with a benchmark to evaluate the performance of anti-UAV tracking algorithms. Given the limitations of existing trackers in this task, we propose MCATrack, a two-stage tracking framework designed to effectively handle complex backgrounds and long-term target variations. Experimental results demonstrate that MCATrack outperforms existing state-of-the-art trackers on both the TDTIV dataset and the widely used Anti-UAV410 dataset. In future work, we aim to explore more effective ways to integrate motion and temporal information, further enhancing long-term tracking performance.

## Acknowledgements

The authors would like to thank Dajun Xing (BNU), Weiming Hu (CASIA), Tao Zhang (IP, CAS), Jiayi Zhang (Fudan), and Zhihua Wu (SHU) for their constructive suggestions. This work is sponsored by Beijing Nova Program (20220484097, 20240484703).

## References

- [1] Luca Bertinetto, Jack Valmadre, Joao F Henriques, Andrea Vedaldi, and Philip HS Torr. Fully-convolutional siamese networks for object tracking. In *Computer Vision—ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8–10 and 15–16, 2016, Proceedings, Part II 14*, pages 850–865. Springer, 2016. 3
- [2] Goutam Bhat, Martin Danelljan, Luc Van Gool, and Radu Timofte. Learning discriminative model prediction for tracking. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6182–6191, 2019. 3
- [3] Goutam Bhat, Martin Danelljan, Luc Van Gool, and Radu Timofte. Know your surroundings: Exploiting scene information for object tracking. In *European conference on computer vision*, pages 205–221. Springer, 2020. 7
- [4] Yidong Cai, Jie Liu, Jie Tang, and Gangshan Wu. Robust object modeling for visual tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9589–9600, 2023. 7
- [5] Ziang Cao, Ziyuan Huang, Liang Pan, Shiwei Zhang, Ziwei Liu, and Changhong Fu. Tctrack: Temporal contexts for aerial tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14798–14808, 2022. 7
- [6] Boyu Chen, Peixia Li, Lei Bai, Lei Qiao, Qihong Shen, Bo Li, Weihao Gan, Wei Wu, and Wanli Ouyang. Backbone is all your need: A simplified architecture for visual object tracking. In *European Conference on Computer Vision*, pages 375–392. Springer, 2022. 7
- [7] Xin Chen, Bin Yan, Jiawen Zhu, Dong Wang, Xiaoyun Yang, and Huchuan Lu. Transformer tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8126–8135, 2021. 3
- [8] Zedu Chen, Bineng Zhong, Guorong Li, Shengping Zhang, and Rongrong Ji. Siamese box adaptive network for visual tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6668–6677, 2020. 7
- [9] Yutao Cui, Cheng Jiang, Limin Wang, and Gangshan Wu. Mixformer: End-to-end tracking with iterative mixed attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13608–13618, 2022. 3, 7
- [10] Yutao Cui, Tianhui Song, Gangshan Wu, and Limin Wang. Mixformerv2: Efficient fully transformer tracking. *Advances in Neural Information Processing Systems*, 36, 2024. 3, 7
- [11] Martin Danelljan, Luc Van Gool, and Radu Timofte. Probabilistic regression for visual tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7183–7192, 2020. 7
- [12] Heng Fan, Liting Lin, Fan Yang, Peng Chu, Ge Deng, Sijia Yu, Hexin Bai, Yong Xu, Chunyuan Liao, and Haibin Ling. Lasot: A high-quality benchmark for large-scale single object tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5374–5383, 2019. 3
- [13] Heng Fan, Hexin Bai, Liting Lin, Fan Yang, Peng Chu, Ge Deng, Sijia Yu, Harshit, Mingzhen Huang, Juehuan Liu, et al. Lasot: A high-quality large-scale single object tracking benchmark. *International Journal of Computer Vision*, 129: 439–461, 2021. 3
- [14] Shenyuan Gao, Chunluan Zhou, Chao Ma, Xinggang Wang, and Junsong Yuan. Aiatrack: Attention in attention for transformer visual tracking. In *European Conference on Computer Vision*, pages 146–164. Springer, 2022. 7
- [15] Shenyuan Gao, Chunluan Zhou, and Jun Zhang. Generalized relation modeling for transformer tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18686–18695, 2023. 7
- [16] R Girshick. Fast r-cnn. *arXiv preprint arXiv:1504.08083*, 2015. 6
- [17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 3
- [18] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 6
- [19] Bo Huang, Jianan Li, Junjie Chen, Gang Wang, Jian Zhao, and Tingfa Xu. Anti-uav410: A thermal infrared benchmark and customized scheme for tracking drones in the wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 1, 3, 7, 8
- [20] Lianghua Huang, Xin Zhao, and Kaiqi Huang. Got-10k: A large high-diversity benchmark for generic object tracking in the wild. *IEEE transactions on pattern analysis and machine intelligence*, 43(5):1562–1577, 2019. 3
- [21] Lianghua Huang, Xin Zhao, and Kaiqi Huang. Globaltrack: A simple and strong baseline for long-term tracking. In *Proceedings of the AAAI conference on artificial intelligence*, pages 11037–11044, 2020. 3, 5, 7
- [22] Bingwei Hui, Zhiyong Song, Hongqi Fan, P Zhong, W Hu, X Zhang, J Ling, H Su, W Jin, Y Zhang, et al. A dataset for infrared detection and tracking of dim-small aircraft targets under ground/air background. *China Sci. Data*, 5(3):291–302, 2020. 1, 3
- [23] Nan Jiang, Kuiran Wang, Xiaoke Peng, Xuehui Yu, Qiang Wang, Junliang Xing, Guorong Li, Jian Zhao, Guodong Guo, and Zhenjun Han. Anti-uav: A large multi-modal benchmark for uav tracking. *arXiv preprint arXiv:2101.08466*, 2021. 1, 3, 6
- [24] Hamed Kiani Galoogahi, Ashton Fagg, Chen Huang, Deva Ramanan, and Simon Lucey. Need for speed: A benchmark for higher frame rate object tracking. In *Proceedings of the IEEE international conference on computer vision*, pages 1125–1134, 2017. 3

- [25] Bo Li, Junjie Yan, Wei Wu, Zheng Zhu, and Xiaolin Hu. High performance visual tracking with siamese region proposal network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8971–8980, 2018. 3
- [26] Xin Li, Yuqing Huang, Zhenyu He, Yaowei Wang, Huchuan Lu, and Ming-Hsuan Yang. Citetracker: Correlating image and text for visual tracking. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9974–9983, 2023. 3
- [27] Yongxin Li, Mengyuan Liu, You Wu, Xucheng Wang, Xi-angyang Yang, and Shuiwang Li. Learning adaptive and view-invariant vision transformer for real-time uav tracking. In *Forty-first International Conference on Machine Learning*, 2024. 7
- [28] Liting Lin, Heng Fan, Zhipeng Zhang, Yong Xu, and Haibin Ling. Swintrack: A simple and strong baseline for transformer tracking. *Advances in Neural Information Processing Systems*, 35:16743–16754, 2022. 7
- [29] Christoph Mayer, Martin Danelljan, Danda Pani Paudel, and Luc Van Gool. Learning target candidate association to keep track of what not to track. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 13444–13454, 2021. 7
- [30] 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. 7
- [31] Christoph Mayer, Martin Danelljan, Ming-Hsuan Yang, Vittorio Ferrari, Luc Van Gool, and Alina Kuznetsova. Beyond sot: Tracking multiple generic objects at once. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6826–6836, 2024. 7
- [32] Matthias Muller, Adel Bibi, Silvio Giancola, Salman Alsubaihi, and Bernard Ghanem. Trackingnet: A large-scale dataset and benchmark for object tracking in the wild. In *Proceedings of the European conference on computer vision (ECCV)*, pages 300–317, 2018. 3
- [33] Liangtao Shi, Bineng Zhong, Qihua Liang, Ning Li, Shengping Zhang, and Xianxian Li. Explicit visual prompts for visual object tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4838–4846, 2024. 7
- [34] Stelios M Smirnakis, Michael J Berry, David K Warland, William Bialek, and Markus Meister. Adaptation of retinal processing to image contrast and spatial scale. *Nature*, 386(6620):69–73, 1997. 4
- [35] Zikai Song, Junqing Yu, Yi-Ping Phoebe Chen, and Wei Yang. Transformer tracking with cyclic shifting window attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8791–8800, 2022. 3
- [36] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 402–419. Springer, 2020. 4
- [37] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017. 3, 5
- [38] Paul Voigtlaender, Jonathon Luiten, Philip HS Torr, and Bastian Leibe. Siam r-cnn: Visual tracking by re-detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6578–6588, 2020. 2, 7
- [39] Gang Wang, Xin Yang, Liang Li, Kai Gao, Jin Gao, Jia-yi Zhang, Da-jun Xing, and Yi-zheng Wang. Tiny drone object detection in videos guided by the bio-inspired magnocellular computation model. *Applied Soft Computing*, 163:111892, 2024. 4
- [40] Qianqian Wang, Yen-Yu Chang, Ruojin Cai, Zhengqi Li, Bharath Hariharan, Aleksander Holynski, and Noah Snavely. Tracking everything everywhere all at once. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19795–19806, 2023. 4
- [41] Rui Wang, Filiz Bunyak, Guna Seetharaman, and Kannappan Palaniappan. Static and moving object detection using flux tensor with split gaussian models. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 414–418, 2014. 4
- [42] Xiao Wang, Xiujun Shu, Zhipeng Zhang, Bo Jiang, Yaowei Wang, Yonghong Tian, and Feng Wu. Towards more flexible and accurate object tracking with natural language: Algorithms and benchmark. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13763–13773, 2021. 3
- [43] Yifan Wang, Jian Zhao, Zhaoxin Fan, Xin Zhang, Xuecheng Wu, Yudian Zhang, Lei Jin, Xinyue Li, Gang Wang, Mengxi Jia, et al. Jtd-uav: Mllm-enhanced joint tracking and description framework for anti-uav systems. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1633–1644, 2025. 3
- [44] Wei Wei. Neural mechanisms of motion processing in the mammalian retina. *Annual Review of Vision Science*, 4:165–192, 2018. 4
- [45] Xing Wei, Yifan Bai, Yongchao Zheng, Dahu Shi, and Yihong Gong. Autoregressive visual tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9697–9706, 2023. 7
- [46] 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. 7
- [47] Jinxia Xie, Bineng Zhong, Zhiyi Mo, Shengping Zhang, Liangtao Shi, Shuxiang Song, and Rongrong Ji. Autoregressive queries for adaptive tracking with spatio-temporal transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19300–19309, 2024. 7
- [48] Haofei Xu, Jing Zhang, Jianfei Cai, Hamid Rezafofighi, and Dacheng Tao. Gmflow: Learning optical flow via global matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8121–8130, 2022. 4
- [49] Zizheng Xun, Shangzhe Di, Yulu Gao, Zongheng Tang, Gang Wang, Si Liu, and Bo Li. Linker: Learning long short-

- term associations for robust visual tracking. *IEEE Transactions on Multimedia*, 2024. [3](#)
- [50] Bin Yan, Houwen Peng, Jianlong Fu, Dong Wang, and Huchuan Lu. Learning spatio-temporal transformer for visual tracking. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10448–10457, 2021. [3](#), [7](#)
- [51] Tianyu Yang, Pengfei Xu, Runbo Hu, Hua Chai, and Antoni B Chan. Roam: Recurrently optimizing tracking model. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6718–6727, 2020. [7](#)
- [52] Xin Yang, Gang Wang, Weiming Hu, Jin Gao, Shubo Lin, Liang Li, Kai Gao, and Yizheng Wang. Video tiny-object detection guided by the spatial-temporal motion information. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3054–3063, 2023. [4](#)
- [53] 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. [3](#), [7](#)
- [54] Zhihao Zhang, Lei Jin, Shengjie Li, JianQiang Xia, Jun Wang, Zun Li, Zheng Zhu, Wenhan Yang, PengFei Zhang, Jian Zhao, et al. Modality meets long-term tracker: A siamese dual fusion framework for tracking uav. In *2023 IEEE International Conference on Image Processing (ICIP)*, pages 1975–1979. IEEE, 2023. [3](#)
- [55] Yaozong Zheng, Bineng Zhong, Qihua Liang, Zhiyi Mo, Shengping Zhang, and Xianxian Li. Odtrack: Online dense temporal token learning for visual tracking. In *Proceedings of the AAAI conference on artificial intelligence*, pages 7588–7596, 2024. [7](#)
- [56] Zheng Zhu, Qiang Wang, Bo Li, Wei Wu, Junjie Yan, and Weiming Hu. Distractor-aware siamese networks for visual object tracking. In *Proceedings of the European conference on computer vision (ECCV)*, pages 101–117, 2018. [2](#)