Electronic Research Archive

https://www.aimspress.com/journal/era

ERA, 34(8): 5580–5611.
DOI: 10.3934/era.2026249
Received: 22 April 2026
Revised: 15 June 2026
Accepted: 17 June 2026
Published: 30 June 2026

Research article

Correlation filter learning with dual-modal motion fusion and total variation for UAV tracking

Yu-Feng Yu[1], **Ziwei Wang**[1], **Xiaoying Tan**[1], **Yang Zhang**[2,*] **and Ke-Kun Huang**[3]

[1] Department of Statistics, Guangzhou University, Guangzhou, China
[2] Department of Computer and Information Science, University of Macau, Macau 999078, China
[3] School of Mathematics, Jiaying University, Meizhou 514015, China

* **Correspondence:** Email: yc47958@um.edu.mo.

Abstract: Unmanned aerial vehicle (UAV) visual tracking is an important task in low-altitude perception and aerial monitoring, where trackers must handle fast camera motion, scale variation, occlusion, background clutter, and limited onboard computational resources. These challenges can easily lead to tracking drift or failure. To address these issues, we propose a correlation filter tracker named correlation filter learning with dual-modal motion fusion and total variation for UAV tracking (CFLDT). First, the search region is updated using the dual-modal motion fusion (DMMF), and features are then extracted from the calibrated area. This process integrates local motion cues and correlation response information to improve target localization under complex UAV motion. In addition, a multi-response spatial penalty matrix (MRSP) is introduced to dynamically constrain the filter, while the environment-aware regularization term is optimized through the total variation constraint module (TVCM). Finally, both the MRSP and the environment-aware regularization term are incorporated into the filter training process, yielding a more discriminative model for distinguishing the target from the background. Extensive comparative experiments on multiple UAV tracking datasets demonstrate that CFLDT achieves competitive tracking accuracy, robustness, and real-time performance.

Keywords: UAV target tracking; total variation; correlation filtering

1. Introduction

Unmanned aerial vehicle (UAV) visual tracking has become an important task in low-altitude perception because UAV platforms provide flexible viewpoints, wide-area coverage, and rapid deployment in traffic monitoring, public safety, and mobile robotics. Public UAV tracking benchmarks, including DTB70, UAV123, and UAVDT, have further promoted systematic evaluation of trackers under UAV-specific challenges such as abrupt camera motion, small target size, severe scale variation, partial occlusion, motion blur, and background clutter [1–3].

UAV target tracking also provides technical support for intelligent transportation applications, such as real-time traffic mobility analysis and UAV-enabled vehicle monitoring [4, 5]. However, in complex and dynamic real-world environments, UAV tracking still faces multiple challenges, including occlusion, fast motion, illumination variation, and distractors with similar appearances. These factors impose stringent requirements on the robustness and real-time performance of tracking algorithms.

In recent years, the rapid development of deep learning has advanced drone object tracking technology. Deep learning-based trackers are generally categorized into convolutional neural network (CNN)-based trackers and vision transformer (ViT)-based trackers. CNN-based trackers enhance the robustness of tracking algorithms by optimizing feature representation [6, 7]. Meanwhile, transformer-based trackers leverage spatiotemporal information [8, 9] to mitigate interference from complex environments. Representative studies further show that window-level attention and cyclic-shift strategies can preserve target integrity and improve sample diversity, while frequency-domain reformulation can effectively reduce the computational overhead introduced by shifted-window matching [10]. Although CNN-based and ViT-based methods achieve higher accuracy, they suffer from large model sizes and high computational overhead. In contrast, discriminative correlation filter (DCF)-based methods exhibit clear advantages in memory usage and computational resource requirements. Given that drone onboard platforms are typically resource-constrained, DCF-based trackers often demonstrate greater practical competitiveness due to their lightweight and high-efficiency characteristics.

In UAV object tracking tasks, the high-speed and stochastic motion of targets poses significant challenges to traditional DCF-based trackers. This is particularly evident as the target approaches the image boundary. In this scenario, truncated background information severely hampers the filter's ability to distinguish the target from the background. To address this issue, spatial-temporal regularized correlation filters (STRCF) [11] introduces spatial regularization to penalize background regions, thereby learning a more discriminative filter. By optimizing the spatial regularization term, we suppress the filter's attention to irrelevant background features. This substantially mitigates interference from complex backgrounds.

Adaptive spatial-temporal surrounding-aware correlation filter tracker via ensemble learning (ASTSAELT) [12] further proposes a selective spatial regularization strategy to preserve valuable information within the target bounding box. Meanwhile, dynamic aberrance-repressed temporal regularized correlation filter (DARTCF) [13] adopts a mechanism for dynamically adjusting regularization strength to adapt to varying scenarios. Recent correlation-filter research also introduces joint spatial-channel attention with structured sparsity (e.g., fiber/slice group sparsity) to suppress redundant features and improve anti-occlusion robustness through reliability-aware update strategies [14]. However, existing methods still underutilize response information during optimization. To overcome this limitation, we innovatively integrate the target's motion response and correlation response. This integration enables the construction of a more comprehensive penalty matrix, which optimizes the spatial regularization process.

In UAV video sequences, abrupt environmental changes often lead to target tracking drift. Environmental mutation-insensitive correlation filters (EMCF) [15] improves target localization accuracy to some extent by incorporating environmental change information as contextual cues into the tracker training process. In addition, saliency-aware DCF methods have shown that jointly modeling spatial saliency, spatio-temporal constraints, and reliability-aware template updates can effectively alleviate filter degradation in cluttered and occluded UAV scenarios [16]. However, existing approaches still do not fully utilize inter-frame information, resulting in notable limitations when dealing with sudden environmental variations. To address this issue, we propose to integrate a total variation (TV) modulation module and a residual regularization term. This integration yields a novel environment-aware regularization term.

Furthermore, robust feature can significantly enhance both the precision and success rate of trackers. To improve the quality of feature, this paper proposes a dual-modal motion fusion module (DMMF) that achieves feature pre-optimization by calibrating the search region. Unlike conventional DCF-trackers, this module innovatively integrates dual-modal information. It combines both motion response and correlation response to enable dynamic adjustment of the search area. It provides accurate motion compensation to effectively address the challenge of locating search regions in complex motion scenarios.

This paper proposes a correlation filter learning with dual-modal motion fusion and total variation for UAV tracking (CFLDT) to address the challenges in UAV tracking. The detailed workflow of the proposed method is illustrated in Figure 1. CFLDT consists of two main stages: feature processing and filter learning. In the feature processing, DMMF is employed to update the search region. By integrating motion response and correlation response, CFLDT is able to perform object tracking within a more precise search area. During the model training phase, we introduce a total variation constraint model (TVCM) and a multi-response spatial penalty matrix (MRSP). The TVCM-based environment-aware regularization term improves the tracker's sensitivity to environmental changes. This enhancement allows it to effectively handle disturbances caused by sudden scene variations. Meanwhile, the multi-response fusion adaptively adjusts the weights of the spatial regularization term. This effectively suppresses background distractions and highlights the target region. By incorporating motion response information, CFLDT achieves enhanced robustness against target occlusion.

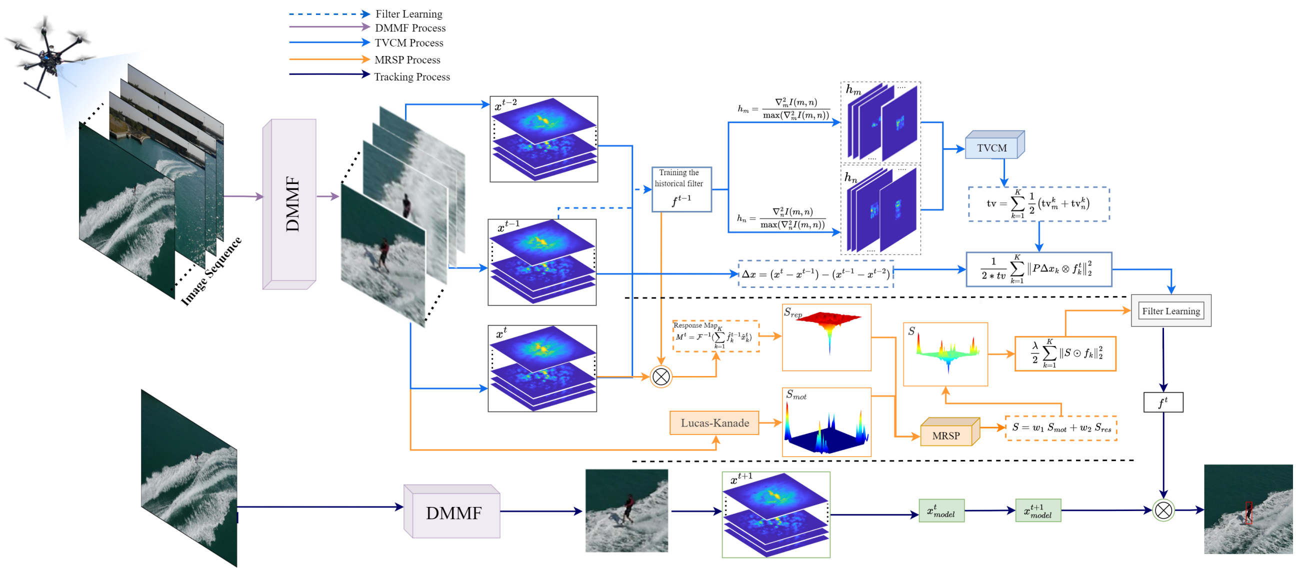


Figure 1. The framework of the proposed tracker. The CFLDT enhances the target features via DMMF. Subsequently, it jointly trains a new filter by integrating TVCM and MRSP, leading to high-precision object tracking.

We summarize the contributions of this paper as follows.

- The proposed DMMF tackles the tracking of high-speed vehicles through a dynamic search region calibration mechanism. This is achieved by fusing motion and correlation response maps to generate a compensation vector. The module breaks through the limitations of traditional unimodal motion prediction. It enables intelligent adjustment of the search window, thereby effectively mitigating issues like target truncation and background interference. Features extracted from the calibrated region exhibit enhanced spatial consistency and discriminability. The optimized features enhance tracking stability in complex traffic scenarios. This improvement contributes to the technical foundation of intelligent traffic management systems.
- This paper introduces the TVCM, designed to innovatively quantify the smoothness characteristics of the filter. This is achieved by analyzing the local gradient features of the filter response map along with its TV L_2 norm. The tracker integrates this module with an adaptive environment-aware regularization term. This integration enables it to dynamically perceive changes in the environment. This design significantly enhances the algorithm's tracking performance in complex scenarios, directly contributing to the stability and accuracy of traffic monitoring and management.
- This paper proposes the MRSP. By leveraging the complementary information from both motion and correlation response maps, a spatially and temporally consistent penalty matrix is constructed. This matrix captures the spatio-temporal evolution characteristics of the target and background in a more comprehensive manner. Integrated into a spatial regularization framework, this penalty mechanism enables precise suppression of distracting environments. The proposed fusion strategy enhances the robustness of the algorithm in complex backgrounds.
- This paper proposes a unified and robust visual tracking framework, whose core innovation lies in the synergistic integration of the DMMF, the MRSP, and the TVCM. This holistic design significantly enhances the tracker's environmental awareness and resistance to interference in complex scenarios.

The rest of this paper is organized as follows. Section 2 reviews the related work. Section 3 details the proposed CFLDT tracker model. In Section 4, we present an extensive experimental analysis. Finally, Section 5 summarizes the paper.

2. Related work

2.1. Tracking with enhanced features

A robust feature representation should maximize the distinctiveness of the target, thereby effectively distinguishing it from the background and similar objects. Deep learning-based methods typically optimize feature extraction to obtain more robust representations. For instance, the query-guided redetection tracker (QRDT) [17] tracker dynamically updates the target appearance representation through its the query update (QU) branch, significantly enhancing discriminative capability during target retrieval. In contrast, conventional DCF-based trackers often rely on hand-crafted features such as histogram of oriented gradients (HOG) and color names (CN). To further improve tracking accuracy, The aberrance repressed correlation filter (ARCF) [18] integrates HOG and CN features in a multi-channel manner.

This fusion strategy enables a more comprehensive feature representation, thereby effectively enhancing the stability and robustness of the tracker.

Furthermore, several studies have achieved effective feature enhancement through preprocessing techniques. Image denoising serves as a widely adopted strategy for this purpose. The learning dynamic-sensitivity enhanced correlation filter (LDECF) [19] employs a non-local means algorithm to denoise images prior to feature extraction. The robust correlation filter learning (RCFL) [20] applies Gaussian denoising directly to the features. Both methods significantly improve the discriminative capability of target contours, resulting in more robust feature representations. On the other hand, the enhanced robust spatial feature selection and correlation filter learning (EFSCF) [21] enhances features specifically within the target region, thereby increasing the contrast between the target and the background and strengthening the discriminative power of the tracker. However, these approaches largely overlook the impact of search positioning on feature representation. To enable more accurate target feature extraction, this paper introduces the use of the DMMF method for a dynamic update of the search location.

2.2. *Trackers with response information*

With the rapid advancement of object tracking technology, DCF-based trackers have undergone significant optimization in various aspects. In recent years, the effective utilization of response information has garnered increasing attention in the research community. Mutation sensitive correlation filter (MSCF) [22] constructs more adaptive labels based on response information, thereby enhancing the tracker's capability to handle abrupt target changes. Meanwhile, spatial reliability enhanced correlation filter (SRECF) [23] optimizes the labeling process by integrating spatial reliability maps and response pooling, which significantly improves the distinguishability between the target and the background. Such label refinement strategies effectively mitigate target drift.

On the other hand, the ReCF tracker [24], which exploits response reasoning for correlation filters, leverages historical information to mitigate tracking drift. In addition, theoretical analyses of correlation-filter modeling reveal the intrinsic equivalence among correlation, convolution, and circulant-matrix formulations, and explain why frequency-domain computation can simultaneously improve efficiency and numerical consistency [25]. For response modeling in transformer trackers, frequency-domain reformulation has also been used to accelerate cyclic-shift attention while preserving matching performance [10]. Given the critical influence of regularization on tracker performance, this paper further proposes a penalty matrix derived from response and motion information. This matrix imposes more comprehensive and adaptive constraints on the regularization term, enabling the tracker to focus more precisely on the target region. The proposed mechanism helps alleviate boundary effects, thereby substantially improving the robustness and accuracy of the tracking system.

2.3. *Trackers with background information*

Background information holds significant value in object tracking tasks; however, its potential remains underexploited. To leverage such information more effectively, researchers have proposed various integration strategies. For instance, background-aware correlation filters (BACF) [26] extracts negative sample patches from background regions to participate in model training, thereby enhancing adaptability to complex backgrounds. Meanwhile, object saliency-aware dual regularized correlation filter (DRCF) [27] employs a dual regularization strategy to impose stricter constraints on irrelevant background areas, significantly improving the discriminative performance of the filter.

Environmental variation itself can be regarded as a form of dynamic background information. EMCF [15] enhances the representation and utilization of background information through a residual modeling mechanism. Similarly, bidirectional incongruity-aware correlation filter (BiCF) [28] introduces a bidirectional inconsistency error term to efficiently learn appearance variations. Nevertheless, these methods still fail to adequately account for the spatiotemporal inconsistencies of environmental changes. To address this challenge, we incorporate an environmental mutation-aware residual term and propose TVCM, which dynamically perceives environmental variations. This design significantly improves the tracker's adaptability and robustness in complex scenarios.

3. Proposed tracking methods

3.1. Dual-modal motion fusion module

A fundamental premise for achieving robust object tracking is the acquisition of high-quality feature representations. The efficacy of feature extraction is contingent upon the selection of the search region. Crucially, the location of the search region is determined by the predicted center point of the target. Thus, refining the accuracy of center point prediction is a critical objective.

In DCF-based trackers, the target's center point in the next frame is typically predicted by utilizing a motion compensation mechanism based on its historical displacement. The search region is centered on this predicted point. This strategy effectively minimizes the risk of the target leaving the search window. However, the target may experience rapid motion or partial occlusion. In these cases, relying solely on historical displacement for compensation can fail to capture its true motion state. Such limitations can lead to prediction bias and even cause target drift.

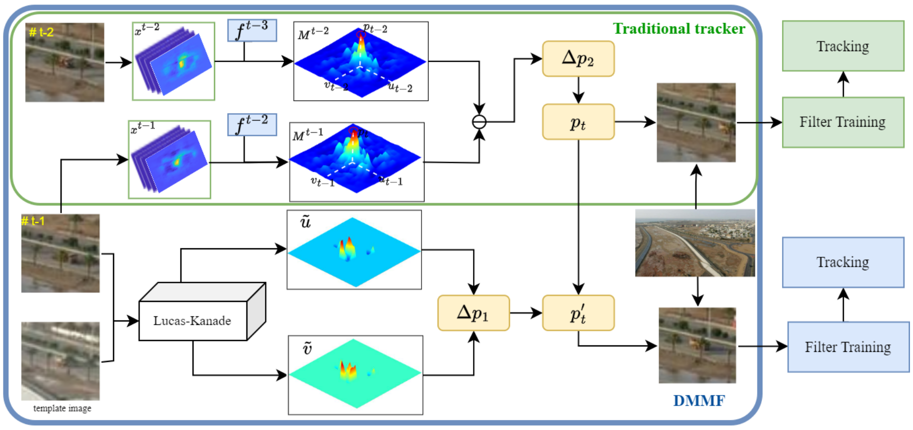


Figure 2. The framework of the DMMF. The local compensation Δp_1 is estimated by the Lucas-Kanade (LK) optical flow and provides pixel-level motion cues. The global compensation Δp_2 is derived from response-guided historical displacement and captures the target's global motion trend. Their weighted fusion refines the target center and calibrates the search region.

To address this issue, we propose the DMMF. Figure 2 provides a detailed illustration of the DMMF.

When dealing with rapidly moving targets, traditional methods often cause the target to drift toward the boundary of the search region. In contrast, the DMMF maintains the target consistently near the center of the image. This comparative advantage is clearly demonstrated in Figure 2. The employed centering strategy mitigates boundary effects and improves tracking stability and robustness.

Unlike conventional single-source methods, our module generates a weighted motion compensation vector $\Delta p = (\bar{u}, \bar{v})$. This is achieved by fusing a local compensation vector ($\Delta p_1 = (\bar{u}_1, \bar{v}_1)$) from optical flow with a global compensation vector ($\Delta p_2 = (\bar{u}_2, \bar{v}_2)$) derived from historical displacement. The fusion module is particularly effective in challenging scenarios, such as rapid motion or partial occlusion. It achieves synergistic complementarity from multi-source information, which enhances the tracker's robustness.

It should be emphasized that the proposed DMMF is not a direct reuse of LK optical flow as an independent motion predictor. The novelty lies in coupling local optical-flow displacement with the global displacement inferred from the correlation-filter response and historical trajectory. The fused compensation is used to recenter the search region before feature extraction, reducing boundary truncation and background contamination. Moreover, the same motion and response cues are further reused in the MRSP construction, which links search-region calibration and spatial regularization within a unified correlation-filter framework.

The refined search window center is denoted as $p'_t = (u'_t, v'_t)$. It is computed by adding the motion compensation vector Δp to the previous target center $p_{t-1} = (u_{t-1}, v_{t-1})$. The formulation of u'_t and v'_t is as follows:

$$\begin{cases} u'_t = & u_{t-1} + \bar{u}, \\ v'_t = & v_{t-1} + \bar{v}, \end{cases} \tag{3.1}$$

where $\bar{u} \in \mathbb{R}$ denotes the horizontal displacement component, and $\bar{v} \in \mathbb{R}$ represents the vertical displacement component. $\bar{u}$ and $\bar{v}$ are derived as:

$$\begin{cases} \bar{u} = & \alpha * \bar{u}_1 + (1 - \alpha) * \bar{u}_2, \\ \bar{v} = & \alpha * \bar{v}_1 + (1 - \alpha) * \bar{v}_2, \end{cases} \tag{3.2}$$

where $\bar{u}_1$ and $\bar{v}_1$ represent the horizontal and vertical displacement components obtained by optical flow estimation. $\bar{u}_2$ and $\bar{v}_2$ denote the corresponding displacement components derived from historical motion trajectory analysis. The weighting parameter $\alpha \in [0.1, 1)$ dynamically balances the contributions of optical flow and displacement.

Local compensation vector Δp_1: The LK optical flow method enhances tracker robustness by estimating a dense motion field, which provides effective motion cues. In this work, the LK algorithm is used to compute the optical flow field between the $t-1$-th frame and a template image to achieve local motion compensation. Considering that the initial frame contains complete appearance information for the target, the first frame is selected as the template.

Based on the brightness constancy assumption, the grayscale information of the image is utilized for computation. Let $G(m, n, t-1)$ denote the grayscale value at position (m, n) in the $t-1$-th frame. For each pixel (m, n) in the image, a fundamental optical flow constraint equation can be established. This equation describes the relationship between the spatial and temporal variations of the intensity of the image and the motion vector $(\tilde{u}_{mn}, \tilde{v}_{mn})$. The optical flow equation for $G(m, n, t-1)$ is formulated as follows:

$$0 = \frac{\partial G(m,n,t-1)}{\partial m} \cdot \tilde{u}_{mn} + \frac{\partial G(m,n,t-1)}{\partial n} \cdot \tilde{v}_{mn} + \frac{\partial G(m,n,t-1)}{\partial (t-1)}. \tag{3.3}$$

The terms $\frac{\partial G(m,n,t-1)}{\partial m}$ and $\frac{\partial G(m,n,t-1)}{\partial n}$ denote the spatial gradients of the grayscale image G^{t-1} along the horizontal (m-axis) and vertical (n-axis) directions, respectively. $\frac{\partial G(m,n,t-1)}{\partial (t-1)}$ represents the temporal gradient at position (m, n), characterizing the intensity change over time. The specific calculation process is as follows:

$$\begin{cases} \frac{\partial G(m,n,t-1)}{\partial m} = & G(m+1,n,t-1) - G(m,n,t-1), \\ \frac{\partial G(m,n,t-1)}{\partial n} = & G(m,n+1,t-1) - G(m,n,t-1), \\ \frac{\partial G(m,n,t-1)}{\partial (t-1)} = & G(m,n,t-1) - G(m,n,1), \end{cases} \tag{3.4}$$

Equation (3.3) is underconstrained for a single pixel. The LK method assumes that the motion vector is constant within a small local neighborhood (e.g., 5×5 window) around each pixel (m, n). This allows us to form an over-constrained system of equations for each neighborhood. This system is then solved via least-squares estimation to obtain the motion vector $(\tilde{u}_{mn}, \tilde{v}_{mn})$. Then, a weighted average of all motion vectors is computed to obtain the local motion compensation $\Delta p_1 = (\bar{u}_1, \bar{v}_1)$. $\bar{u}_1$ and $\bar{v}_1$ are derived as.

$$\begin{cases} \bar{u}_1 = & \frac{1}{MN} \sum_{m=1}^{M} \sum_{n=1}^{N} \tilde{u}_{mn}, \\ \bar{v}_1 = & \frac{1}{MN} \sum_{m=1}^{M} \sum_{n=1}^{N} \tilde{v}_{mn}, \end{cases} \tag{3.5}$$

Global compensation vector Δp_2: Historical displacement information preserves global motion patterns while suppressing noise interference. Therefore, this work constructs global motion compensation $\Delta p_2 = (\bar{u}_2, \bar{v}_2)$ based on historical displacement. The target positions in the $t-1$-th frame ($p_{t-1} = (u_{t-1}, v_{t-1})$) and the $t-2$-th frame ($p_{t-2} = (u_{t-2}, v_{t-2})$) are determined by locating the peak points in their respective response maps. The response map for the $t-1$-th frame M^{t-1} is presented below.

$$M^{t-1} = \mathcal{F}^{-1}(\sum_{k=1}^{K} \hat{f}_k^{t-2} \hat{x}_k^{t-1}). \tag{3.6}$$

Based on these positional correspondences between consecutive frames, the global motion compensation $\Delta p_2 = (\bar{u}_2, \bar{v}_2)$ is derived. The displacement components $\bar{u}_2$ and $\bar{v}_2$ are calculated as follows.

$$\begin{cases} \bar{u}_2 = & u_{t-1} - u_{t-2}, \\ \bar{v}_2 = & v_{t-1} - v_{t-2}. \end{cases} \tag{3.7}$$

Based on the center point coordinates $p'_t = (u'_t, v'_t)$ computed by the DMMF, we dynamically adjust and update the search region. Within this refined and more precise search area, we subsequently extract the feature representation of the target. For clarity and consistency in the subsequent exposition, we uniformly employ the notation x_k^t. It denotes the target features extracted from the search region updated by the DMMF.

3.2. Correlation filter learning with dual-modal motion fusion and total variation

We employ BACF as the baseline tracker. To better elucidate the proposed method, we first provide a concise introduction to BACF. The K-channel features are denoted as $x_k \in \mathbb{R}^{M\times N}$. The predefined class label determined using the Gaussian formula is $y \in \mathbb{R}^{M\times N}$. The BACF objective function is then defined as:

$$\varphi(f) = \frac{1}{2}\left\| y - \sum_{k=1}^{K} P x_k^t \otimes f_k^t \right\|_2^2 + \sum_{k=1}^{K} \|f_k\|_2^2, \tag{3.8}$$

where t denotes the t-th frame of the video sequence. $f = [f_1, f_2, \cdots, f_K] \in \mathbb{R}^{M\times N\times K}$ represents filters for each feature channel, where $f_k \in \mathbb{R}^{M\times N}$ denotes the k-channel filter. $x_k^t \in \mathbb{R}^{D\times N}$ is further refined by using a binary matrix P to crop the central region of the feature x_k^t, where the correlation operator is denoted by $\otimes$.

Leveraging the temporal continuity property of video sequences, our proposed CFLDT incorporates two key components: the environment-aware residual term $ER^t = \frac{1}{2*tv}\sum_{k=1}^{K}\left\|P\Delta x_k \otimes f_k^t\right\|_2^2$ and the spatially sensitive weighting matrix S. Through the TV regulation coefficient tv, CFLDT significantly enhances the filter's adaptability to environmental changes between consecutive frames. The CFLDT objective function is defined as.

$$\begin{aligned}\varphi(f) =& \frac{1}{2}\left\| y - \sum_{k=1}^{K} P x_k^t \otimes f_k^t \right\|_2^2 + \frac{\lambda}{2}\sum_{k=1}^{K} \|S \odot f_k\|_2^2 \\ &+ \frac{1}{2*tv}\sum_{k=1}^{K}\left\|P\Delta x_k \otimes f_k^t\right\|_2^2,\end{aligned} \tag{3.9}$$

$\Delta x_k = (x_k^t - x_k^{t-1}) - (x_k^{t-1} - x_k^{t-2})$. $\odot$ represents element-by-element multiplication.

Total variation regulation coefficient ***tv*****:** The objective of TV regularization is to suppress noise while preserving edge and structural information in images or signals by constraining gradient variations. For a given image, the TV regularization term can be mathematically expressed as.

$$\mathrm{TV} = \sum_{k=1}^{K}\left(\sum_{m=1}^{M-1}\sum_{n=1}^{N-1}\Big((\nabla_m I(m,n))^2 + (\nabla_n I(m,n))^2\Big)\right). \tag{3.10}$$

$I(m,n)$ denotes the pixel intensity at position (m,n) in the image. $\nabla_m I(m,n)$ and $\nabla_n I(m,n)$ represent the horizontal and vertical gradient components, respectively. K indicates the image channel. The fundamental principle of this formulation lies in quantifying the image's smoothness through computation of directional gradient variations.

In visual object tracking, the filter from the previous frame can be regarded as a feature map. This allows its TV regularization term to be computed in a similar manner. For each feature channel k, the gradient information is employed to characterize the magnitude of variation in the image or filter, while the L_2-norm quantifies its smoothness.

Calculate the gradients in the horizontal $\nabla_m I(m,n)$ and vertical $\nabla_n I(m,n)$ directions.

$$\begin{cases} \nabla_m I(m,n) = & I(m+1,n) - I(m,n), \\ \nabla_n I(m,n) = & I(m,n+1) - I(m,n), \end{cases} \tag{3.11}$$

Then calculate the square of the gradient and perform gradient normalization.

$$\begin{cases} h_m = & \frac{\nabla_m^2 I(m,n)}{\max(\nabla_m^2 I(m,n))}, \\ h_n = & \frac{\nabla_n^2 I(m,n)}{\max(\nabla_n^2 I(m,n))}, \end{cases} \tag{3.12}$$

Compute the sum for each channel individually.

$$\begin{cases} \mathrm{tv}_m = & \sum_{m=1}^{M-1} \sum_{n=1}^{N} h_m(m,n), \\ \mathrm{tv}_n = & \sum_{m=1}^{M} \sum_{n=1}^{N-1} h_n(m,n), \end{cases} \tag{3.13}$$

The final TV value is the L_2 norm of the horizontal and vertical gradients.

$$\mathrm{tv} = \sum_{k=1}^{K} \sqrt{\frac{1}{2}\left(\mathrm{tv}_m^k + \mathrm{tv}_n^k\right)}. \tag{3.14}$$

The learned correlation filter can be viewed as a structured spatial response template. Although the filter is optimized in the Fourier domain for computational efficiency, its spatial structure still determines the stability and discriminative quality of the response map. When the training sample contains background clutter, partial occlusion, or abrupt environmental changes, the filter may learn high-frequency oscillatory coefficients that fit background details rather than the target support, producing unstable side peaks in the response map. The TV coefficient measures the local spatial variation of the filter structure and is used to normalize the environment-aware residual term in Eq (3.9). In this way, the TV-modulated residual suppresses unstable background-induced responses while preserving coherent target-support structures.

Multi-response spatial penalty matrix S: Applying uniform spatial weights to filters fails to account for importance variations across target regions. To address this limitation, we propose an adaptive spatial regularization weighting method. This method constructs a dynamic weight matrix by integrating the motion and appearance information. This approach enables the complementary fusion of multi-modal information. It thereby effectively mitigates the limited representational capacity inherent in single-modal methods. The detailed construction process of S is described as follows.

Initially, the motion saliency map S_{mot} is calculated using optical flow estimation. For each local region, we compute the corresponding motion vector by solving Eq (3.3) using the LK algorithm with least-squares fitting. The motion response map $s_{mot} = (\mathcal{L}_{mn}(s_{mot})) \in \mathbb{R}^{M\times N}$ of the target is then generated by calculating the magnitude of these motion vectors. $\mathcal{L}_{mn}(s_{mot})$ is described as follows.

$$\mathcal{L}_{mn}(s_{mot}) = \sqrt{(u_{mn})^2 + (v_{mn})^2}. \tag{3.15}$$

The motion saliency map $S_{mot} = (\mathcal{L}_{mn}(S_{mot})) \in \mathbb{R}^{M\times N}$ is obtained through normalization processing. $\mathcal{L}_{mn}(S_{mot})$ is described as follows.

$$\mathcal{L}_{mn}(S_{mot}) = 1 - \frac{\mathcal{L}_{mn}(s_{mot}) - \min(\mathcal{L}_{mn}(s_{mot}))}{\max(\mathcal{L}_{mn}(s_{mot}) - \min(\mathcal{L}_{mn}(s_{mot}))}. \tag{3.16}$$

Subsequently, the appearance saliency map S_{res} is obtained by normalizing the response map M^t computed from the t-1-th frame's filter f_k^{t-1} and the features x_k^t. S_{res} is described as follows.

$$S_{res} = \frac{\max(M^t) - M^t}{\max(M^t) - \min(M^t)}. \tag{3.17}$$

Finally, the MRSP matrix S is constructed by fusing the motion saliency map S_{mot} and the appearance saliency map S_{res} with appropriate weighting coefficients:

$$S = w_1 S_{mot} + w_2 S_{res}, \tag{3.18}$$

where $w_1 + w_2 = 1$.

3.3. ADMM optimization

In this section, we perform model derivation by selecting the k-th channel. To optimize the target filter f_k, an auxiliary variable $g_k^t = P^{\mathrm{T}} f_k^t \in \mathbb{R}^{M\times N}$ is introduced. With this variable, Eq (3.9) can be reformulated as Eq (3.19).

$$\begin{aligned} \varphi_k(g_k, f_k) = &\frac{1}{2}\left\|y - x_k^t \otimes g_k^t\right\|_2^2 + \frac{\lambda}{2}\left\|S \odot f_k^t\right\|_2^2 \\ &+ \frac{1}{2*tv}\left\|\Delta x_k \otimes g_k^t\right\|_2^2. \\ \text{s.t.} \quad &g_k^t = P^{\mathrm{T}} f_k^t,\ k = 1, 2, \ldots, K \end{aligned} \tag{3.19}$$

According to Parseval's theorem, we transform Eq (3.19) into the Fourier domain representation to improve computational efficiency. The specific formulation is presented below.

$$\begin{aligned} \varphi_k(\hat{g}_k, f_k) = &\frac{1}{2}\left\|\hat{y} - \hat{x}_k^t \odot \hat{g}_k^t\right\|_2^2 + \frac{\lambda}{2}\left\|S \odot f_k^t\right\|_2^2 \\ &+ \frac{1}{2*tv}\left\|\Delta\hat{x}_k \odot \hat{g}_k^t\right\|_2^2, \\ \text{s.t.} \quad &\hat{g}_k^t = \sqrt{N} F P^{\mathrm{T}} f_k^t,\ k = 1, 2, \ldots, K \end{aligned} \tag{3.20}$$

where $\hat{(\cdot)}$ denotes the discrete Fourier transform (DFT) of the corresponding variable, e.g., $\hat{x}_k^t = \mathcal{F}(x_k^t)$ and $\Delta\hat{x}_k = \mathcal{F}(\Delta x_k)$. The superscript t is the frame index, and $\overline{(\cdot)}$ used below denotes complex conjugation in the frequency domain. The DFT matrix $F \in \mathbb{C}^{M\times M}$ transforms the spatial-domain variable a into its frequency-domain representation $\hat{a} = \sqrt{N} F a$. For the projected filter variable, this gives $\hat{g}_k^t = \sqrt{N} F g_k^t = \sqrt{N} F P^{\mathrm{T}} f_k^t$.

By constructing the augmented Lagrangian, Eq (3.20) can be rewritten as

$$\begin{aligned} \mathcal{L}(\hat{g}_k, f_k, \hat{\varepsilon}_k) = &\frac{1}{2}\left\|\hat{y} - \hat{x}_k^t \odot \hat{g}_k^t\right\|_2^2 + \frac{1}{2*\mathrm{tv}}\left\|\Delta\hat{x}_k \odot \hat{g}_k^t\right\|_2^2 + \frac{\lambda}{2}\left\|S \odot f_k^t\right\|_2^2 \\ &+ \frac{\mu}{2}\left\|\hat{g}_k^t - \sqrt{N} F P^{\mathrm{T}} f_k^t + \frac{1}{\mu}\hat{\varepsilon}_k\right\|_2^2, \end{aligned} \tag{3.21}$$

where $\hat{\varepsilon}_k$ denotes the Lagrangian multiplier for the k-th channel, and μ is the penalty parameter. We then decompose the original optimization problem into two primal subproblems and one multiplier update:

$$\begin{cases} f_k^{(i+1)} = \arg\min\limits_{f_k} \frac{\lambda}{2}\|S \odot f_k^t\|_2^2 + \frac{\mu}{2}\left\|\hat{g}_k^{(i)} - \sqrt{N} F P^{\mathrm{T}} f_k^t + \frac{1}{\mu}\hat{\varepsilon}_k^{(i)}\right\|_2^2, \\ \hat{g}_k^{(i+1)} = \arg\min\limits_{\hat{g}_k} \frac{1}{2}\left\|\hat{y} - \hat{x}_k^t \odot \hat{g}_k^t\right\|_2^2 + \frac{1}{2\mathrm{tv}}\|\Delta\hat{x}_k \odot \hat{g}_k^t\|_2^2 \\ \qquad\qquad + \frac{\mu}{2}\left\|\hat{g}_k^t - \hat{f}_k^{(i+1)} + \frac{1}{\mu}\hat{\varepsilon}_k^{(i)}\right\|_2^2, \\ \hat{\varepsilon}_k^{(i+1)} = \hat{\varepsilon}_k^{(i)} + \mu\left(\hat{g}_k^{(i+1)} - \hat{f}_k^{(i+1)}\right). \end{cases} \tag{3.22}$$

i denotes the current iteration index. The detailed iterative solutions for each subproblem are presented as follows.

Subproblem f_k : By taking the derivative of Eq (3.22) with respect to f_k and setting it to zero, we obtain the closed-form solution of $f_k^{(i+1)}$ as follows.

$$\left(\frac{\lambda}{N} S \odot S + \mu\right) f_k^{(i+1)} = \mathcal{F}^{-1}\left(\mu \hat{g}_k^{(i)} + \hat{\varepsilon}_k^{(i)}\right). \tag{3.23}$$

Therefore,

$$f_k^{(i+1)} = \frac{\mathcal{F}^{-1}\left(\mu \hat{g}_k^{(i)} + \hat{\varepsilon}_k^{(i)}\right)}{\frac{\lambda S \odot S}{N} + \mu}. \tag{3.24}$$

The division operator denotes element-wise division.

Subproblem g_k: Fixing $f_k^{(i+1)}$ and $\hat{\varepsilon}_k^{(i)}$, the optimal solution for $\hat{g}_k^{(i+1)}$ is obtained by solving the following first-order condition:

$$\left(\overline{\hat{x}_k^t} \odot \hat{x}_k^t + \frac{1}{\text{tv}} \overline{\Delta \hat{x}_k} \odot \Delta \hat{x}_k + \mu\right) \hat{g}_k^{(i+1)} = \overline{\hat{x}_k^t} \odot \hat{y} + \mu \hat{f}_k^{(i+1)} - \hat{\varepsilon}_k^{(i)}. \tag{3.25}$$

The complex conjugate terms $\overline{\hat{x}_k^t}$ and $\overline{\Delta \hat{x}_k}$ arise from the normal equation of the complex-valued least-squares problem in the Fourier domain, where the Hermitian transpose of the diagonal operators $\text{diag}(\hat{x}_k^t)$ and $\text{diag}(\Delta \hat{x}_k)$ reduces to element-wise complex conjugation.

Thus, the closed-form solution is

$$\hat{g}_k^{(i+1)} = \frac{\overline{\hat{x}_k^t} \odot \hat{y} + \mu \hat{f}_k^{(i+1)} - \hat{\varepsilon}_k^{(i)}}{\overline{\hat{x}_k^t} \odot \hat{x}_k^t + \frac{1}{\text{tv}} \overline{\Delta \hat{x}_k} \odot \Delta \hat{x}_k + \mu}. \tag{3.26}$$

Update $\hat{\varepsilon}_{\mathbf{k}}$: The Lagrange multipliers are updated as follows.

$$\hat{\varepsilon}_k^{(i+1)} = \hat{\varepsilon}_k^{(i)} + \mu\left(\hat{g}_k^{(i+1)} - \hat{f}_k^{(i+1)}\right), \tag{3.27}$$

where $\hat{f}_k^{(i+1)} = \sqrt{N} F P^{\mathrm{T}} f_k^{(i+1)}$. In the i-th iteration of alternating direction method of multipliers (ADMM), the update of parameter μ is as follows:

$$\mu^{(i+1)} = \min(\mu_{\max}, \beta \mu^{(i)}). \tag{3.28}$$

Update $\hat{x}_{k,\text{model}}$:

$$\hat{x}_{k,\text{model}} = (1-\eta) \hat{x}_{k,\text{model}}^{t-1} + \eta \hat{x}_k. \tag{3.29}$$

where $\hat{x}_{k,\text{model}}^{t-1}$ and $\hat{x}_k$ represent the appearance models of the t-th frame and the $t-1$-th frame, respectively, and η is the online adaptive rate.

The detailed process of CFLDT optimization in the d channel of the t-th frame is summarized in Algorithm 1.

Algorithm 1: CFLDT Tracker

Input: Image: I^t,
Previous filter: $\hat{f}^{t-1}$,
Historical features: x^{t-2}, x^{t-1},
Grayscale feature: G^{t-1};
Output: Correlation filters: f_k^t;
Initialize: Initialize $\hat{g}_k^t(0), \hat{f}_k^t(0), \hat{\varepsilon}_k(0)$;
Set max ADMM iteration D;
Computation:
1: Update $p'_t = (u'_t, v'_t)$ via Eqs (3.1)–(3.7);
2: Update the search area based on $p'_t = (u'_t, v'_t)$;
3: Extract the t-th feature x^t in the search area;
4: Update tv via Eqs (3.11)–(3.14);
5: Update S via Eqs (3.15)–(3.18);
6: for $i = 1$ to D
7: Update $f_k^{(i+1)}$ via Eq (3.24);
8: Update $\hat{g}_k^{(i+1)}$ via Eq (3.26);
9: Update $\mu^{(i+1)}$ via Eq (3.28);
10: Update $\hat{x}_{k,\text{model}}^t$ via Eq (3.29).

4. Quantitative analysis of experimental results

4.1. Dataset introduction

To evaluate the performance of CFLDT in UAV object tracking, we selected three representative datasets: DTB70 [1], UAV123@10FPS [2], and UAVDT [3]. On these benchmarks, we compared CFLDT against various state-of-the-art trackers. These datasets address key challenges in urban traffic surveillance, such as aspect ratio variation, occlusion, and background clutter. This allows for a rigorous evaluation of the CFLDT's performance in realistic scenarios.

Comprising 70 UAV-captured video sequences, the DTB70 dataset features diverse targets (e.g., pedestrians, vehicles) primarily in outdoor settings. Figure 3 displays representative examples from the dataset. These video sequences feature highly complex and dynamic traffic monitoring scenes. Challenges such as vehicle occlusion and abrupt illumination changes within them present significant difficulties for visual tracking.

The UAV123@10fps dataset comprises 123 video sequences captured by low-altitude UAVs. This dataset employs raw video downsampling to simulate common UAV challenges. It specifically targets scenarios with limited computational resources or bandwidth. Furthermore, its diverse viewpoint variations allow for effective evaluation of tracking algorithms in complex observation conditions. Figure 4 illustrates representative challenging scenarios from a top-down perspective within this dataset.

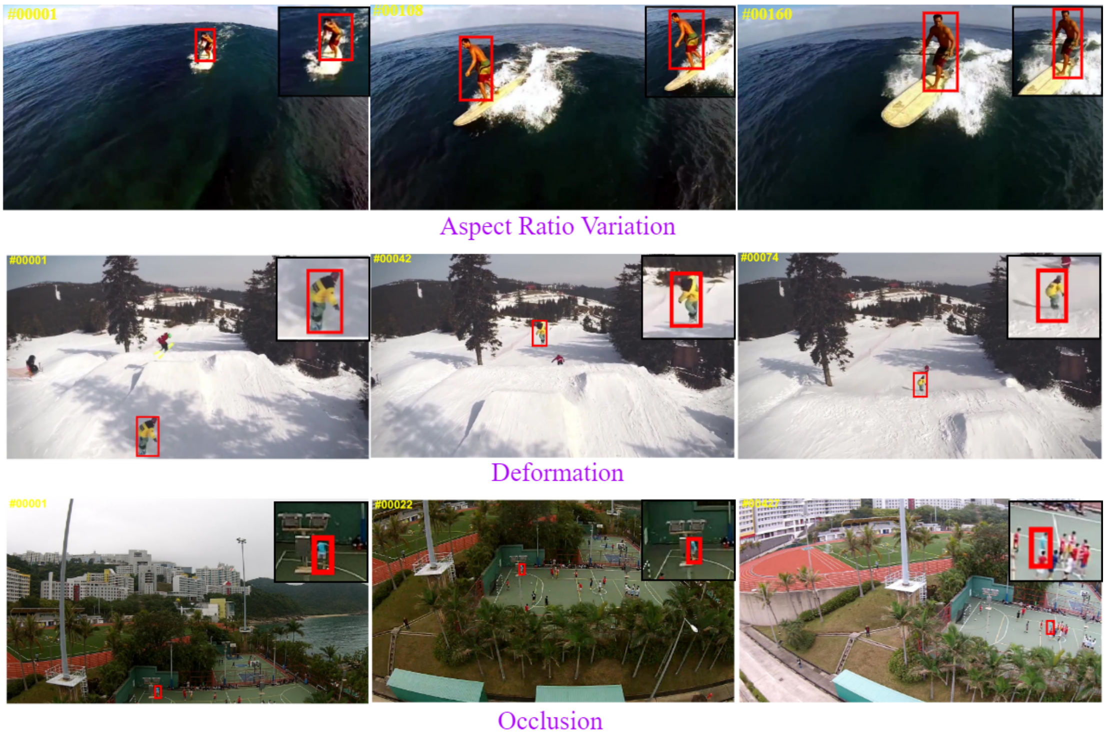


Figure 3. Some attribute changes on the DTB70 dataset.

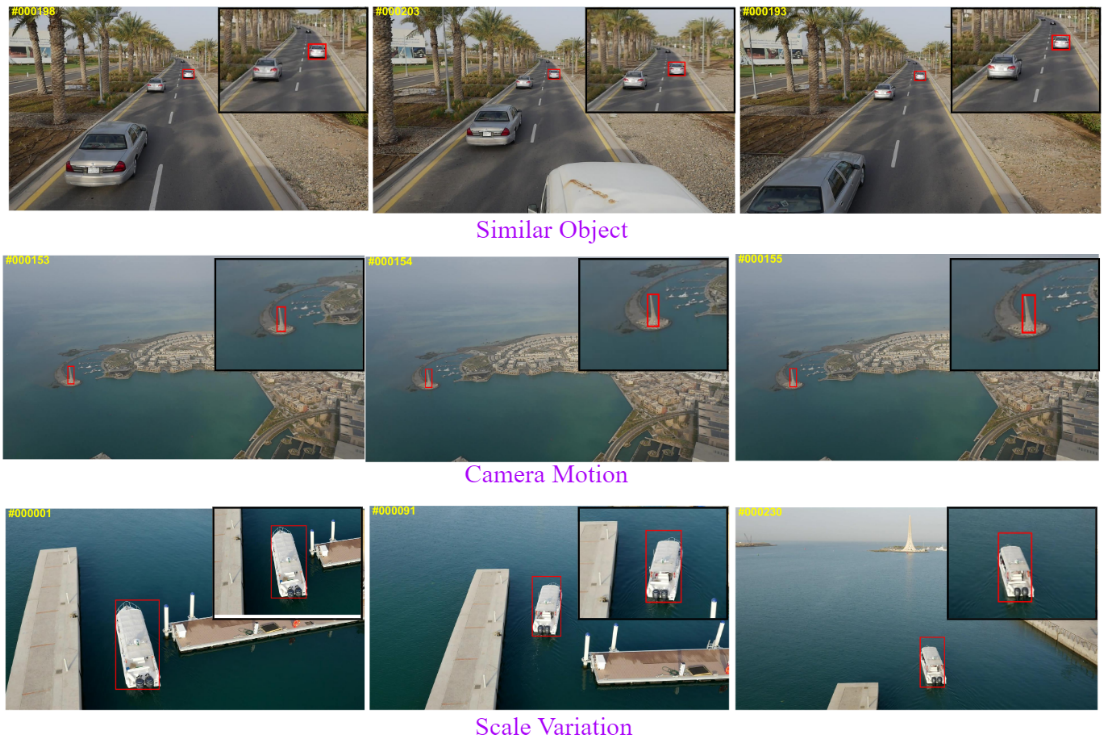


Figure 4. Some attribute changes on the UAV123@10fps dataset.

Figure 5. Some attribute changes on the UAVDT dataset.

The UAVDT dataset comprises 50 video sequences captured by UAVs in real-world scenarios such as urban roads and public squares. This dataset specifically focuses on challenges from a UAV perspective. These include background clutter, illumination variations, and motion blur. Figure 5 provides illustrative examples of some typical scenes.

4.2. Implementation detail

This paper employs the one-pass evaluation (OPE) criterion to assess CFLDT's performance. The evaluation includes two metrics: precision and success rate. Furthermore, frames per second (FPS) is adopted as the metric for tracking speed.

Parameter settings affect model performance. The specific parameter values used in the proposed CFLDT tracker are as follows: $\alpha = 0.6$, $\lambda = 66$, $\mu = 3$, $\beta = 42$, and $w_1 = 0.5$. The learning rate θ is set to 0.029.

All experiments were conducted in MATLAB R2021b on a workstation equipped with a 13th Gen Intel(R) Core(TM) i7-13620H CPU and an NVIDIA GeForce RTX 4060 Laptop GPU. To transparently evaluate the computational overhead of the proposed modules, we further measured the per-frame runtime breakdown on the Yacht2 sequence from the DTB70 dataset using the CPU platform. As reported in Table 1, the DMMF motion compensation module, which includes the LK optical-flow-based local compensation, requires 0.466 ms per frame. The TVCM and ADMM training stages require 2.992 and 1.994 ms per frame, respectively. The complete tracking pipeline achieves an average runtime of 11.841 ms per frame, corresponding to 84.81 FPS on Yacht2. This sequence-level test is used to analyze the latency contribution of each module, while the overall dataset-level speed comparison is provided in the subsequent UAVDT experiment, where CFLDT reaches 103 FPS on the CPU platform. These results jointly demonstrate that CFLDT maintains real-time performance both at the module level and in dataset-level evaluation.

Table 1. Per-frame runtime breakdown of CFLDT on the DTB70 Yacht2 sequence using the CPU platform.

Component	Runtime
DMMF motion compensation	0.466 ms/frame
TVCM computation	2.992 ms/frame
ADMM filter training	1.994 ms/frame
Total tracking pipeline	11.841 ms/frame
FPS	84.81

4.3. Experimental results on the DTB70

On the DTB70 dataset, we compare the proposed CFLDT tracker with 19 state-of-the-art trackers, including efficient convolution operators for tracking (ECO) [29], BACF [26], STRCF [11], ARCF [18], DRCF [27], disruptor-aware interval-based response inconsistency (IBRI) [30], AutoTrack [31], BiCF [28], ReCF [24], multi-regularized correlation filter (MRCF) [32], MSCF [22], EMCF [15], channel attentional correlation filters (CACF) [33], RCFL [20], response-weighted background residual and spatio-temporal regularization correlation filter (RBSCF) [34], EFSCF [21], multi-scale enhanced features correlation filters (MSEFCF) [35], SRECF [23], and LDECF [19]. The experimental results show that CFLDT achieves competitive tracking performance.

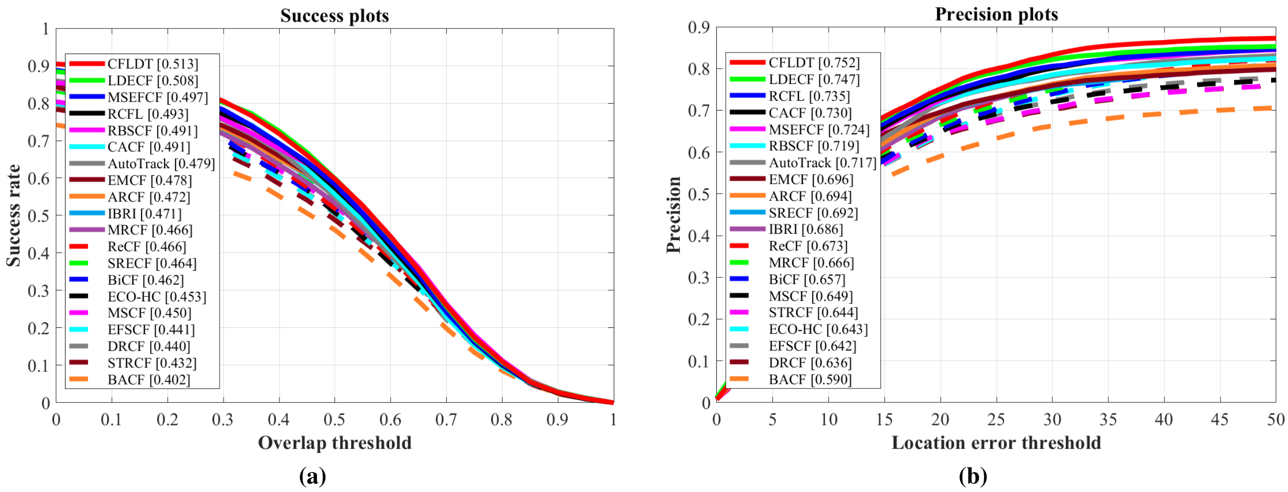


Figure 6. Success plots (a) and precision plots (b) of the proposed CFLDT and other trackers on the DTB70 database.

As illustrated in Figure 6, CFLDT ranks first in both precision and success rate. Specifically, it achieves a tracking precision of 75.2%, outperforming the second-ranked LDECF by 5% and surpassing BACF by 16.2%. In terms of success rate, CFLDT attains a score of 51.3%, which exceeds LDECF by 5% and leads BACF by 11.1%. These results demonstrate that the improved CFLDT exhibits a significant enhancement in tracking performance.

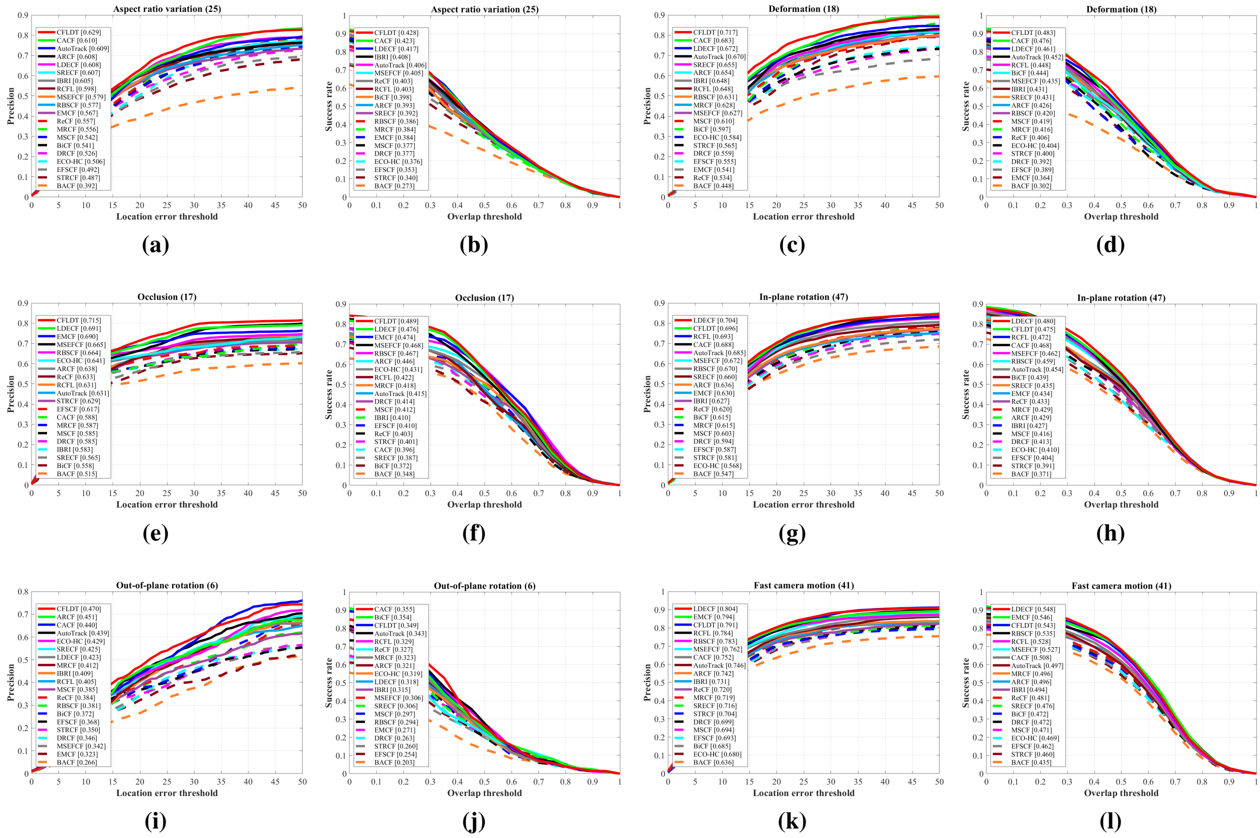
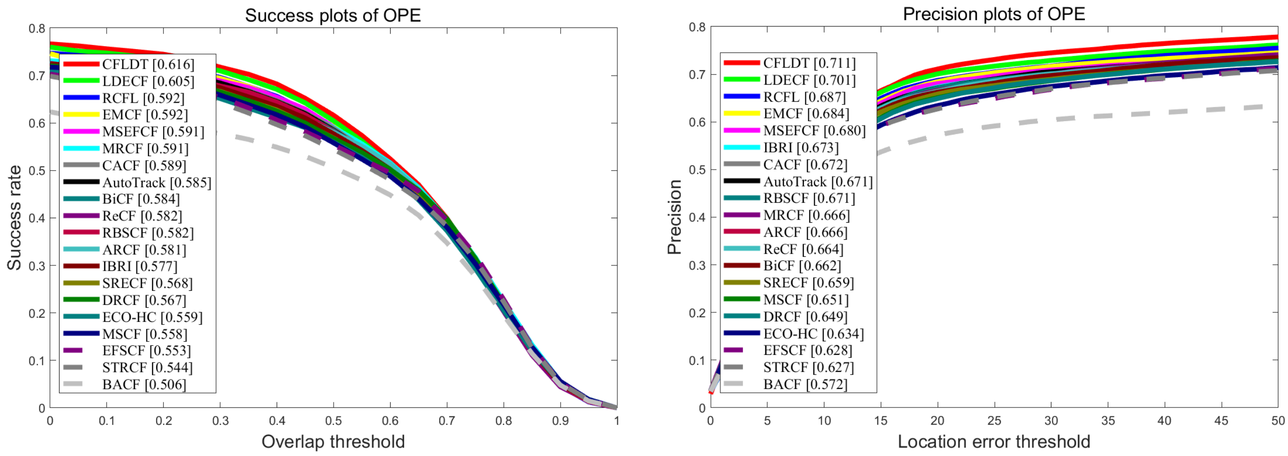


Figure 7. Success plots and precision plots of CFLDT and other trackers on partial attributes of the DTB70 dataset.

To comprehensively evaluate the performance of CFLDT, we conducted comparative experiments under various attributes. Figure 7 presents the experimental results on a subset of these attributes. As can be observed from the figure, CFLDT demonstrates superior performance across multiple attributes. Specifically, it achieves the highest scores in both success rate and precision on the aspect ratio variation (ARV), deformation (DEF), and occlusion (OCC) attributes. This improvement can be attributed to the introduced DMMF module, which optimizes feature representation. In addition, the TVCM enhances adaptability to target mutations through environment-aware residuals. On the out-of-plane rotation (OPR), CFLDT attains the highest precision. This improvement stems from the spatial penalty introduced via multi-response map fusion. This mechanism suppresses environmental interference and enhances focus on the target. Furthermore, CFLDT maintains rankings within the top three in both success rate and precision on the in-plane rotation (IPR) and fast camera motion (FCM) attributes. Overall, these results indicate that the enhanced tracker exhibits robust performance across diverse challenging scenarios.

4.4. Experimental results on the UAV123@10fps

On the UAV123@10fps dataset, we compare CFLDT against the same set of 19 trackers in a comprehensive evaluation. As illustrated in Figure 8, CFLDT achieves the highest performance, ranking first in both success rate and precision. Specifically, in success rate, CFLDT outperforms the second-best tracker (LDECF) by 1.5% and surpasses the baseline (BACF) by a significant margin of 9.9%. Regarding precision, CFLDT exceeds LDECF by 0.5% and demonstrates a substantial improvement of 13.4% over the baseline.



(a) **(b)**

Figure 8. Success plots (a) and precision plots (b) on the UAV123@10fps database.

Table 2. Success rate comparison of the proposed CFLDT and 19 trackers across various attributes on the UAV123@10fps dataset. The top three methods for each attribute are highlighted in red, green, and blue, respectively. An upward arrow ⇑ indicates superior performance compared to the baseline.

	SV	ARC	LR	FM	FO	PO	OV	BC	IV	VC	CM	SO
ECO	0.506	0.449	0.346	0.390	0.283	0.462	0.453	0.425	0.430	0.459	0.544	0.566
BACF	0.456	0.396	0.293	0.332	0.467	0.387	0.375	0.333	0.369	0.409	0.482	0.528
STRCF	0.494	0.413	0.334	0.377	0.261	0.454	0.450	0.374	0.406	0.439	0.526	0.555
ARCF	0.529	0.461	0.394	0.378	0.268	0.498	0.453	0.380	0.467	0.472	0.530	0.573
DRCF	0.520	0.442	0.354	0.386	0.291	0.500	0.469	0.384	0.442	0.464	0.531	0.620
IBRI	0.529	0.461	0.393	0.383	0.269	0.498	0.456	0.360	0.548	0.590	0.623	0.573
AutoTrack	0.535	0.476	0.372	0.407	0.291	0.496	0.490	0.415	0.467	0.480	0.564	0.569
BiCF	0.534	0.466	0.371	0.376	0.305	0.502	0.481	0.427	0.494	0.496	0.555	0.609
ReCF	0.529	0.465	0.359	0.406	0.301	0.499	0.466	0.432	0.501	0.491	0.562	0.598
MRCF	0.542	0.472	0.398	0.404	0.291	0.514	0.483	0.388	0.477	0.489	0.555	0.592
MSCF	0.513	0.446	0.343	0.401	0.263	0.477	0.483	0.316	0.424	0.469	0.524	0.553
EMCF	0.545	0.479	0.395	0.391	0.286	0.494	0.493	0.382	0.478	0.489	0.569	0.576
CACF	0.541	0.486	0.380	0.433	0.289	0.501	0.501	0.364	0.470	0.479	0.565	0.585
RCFL	0.546	0.489	0.390	0.438	0.284	0.507	0.498	0.412	0.466	0.485	0.570	0.593
RBSCF	0.535	0.462	0.393	0.386	0.273	0.483	0.482	0.369	0.452	0.474	0.552	0.572
EFSCF	0.505	0.433	0.351	0.379	0.253	0.466	0.455	0.373	0.403	0.441	0.526	0.557
MSEFCF	0.547	0.412	0.388	0.539	0.289	0.497	0.480	0.370	0.452	0.471	0.574	0.592
LDECF	0.563	0.502	0.419	0.438	0.289	0.512	0.481	0.388	0.489	0.492	0.587	0.604
CFLDT	0.571⇑	0.496⇑	0.431⇑	0.450⇑	0.304⇑	0.537⇑	0.499⇑	0.426⇑	0.500 ⇑	0.511⇑	0.602⇑	0.607⇑

To more accurately evaluate the performance of the improved CFLDT, we conducted comparative experiments under 12 highly challenging attributes. Table 2 shows the tracking success rates of various trackers under these different attributes. The results indicate that CFLDT outperforms other trackers on both scale variation (SV) and viewpoint change (VC). This improvement can be attributed to the introduced MRSP. This mechanism enhances the model's ability to distinguish between the target and the background. Meanwhile, CFLDT also achieves the best performance under low resolution (LR) and camera motion (CM). This is attributed to the TVCM module's capability to effectively preserve target features while suppressing background and irrelevant noise. Furthermore, CFLDT performs optimally on attributes such as fast motion (FM), and partial occlusion (PO). This is likely because the DMMF provides a more accurate search region. By dynamically adjusting the search window position, it effectively mitigates target drift caused by OCC and FM.

4.5. Experimental results on the VisDrone-SOT

To further evaluate the proposed CFLDT in scene-specific UAV-view tracking scenarios, we compare it with representative DCF-based trackers on the VisDrone-SOT dataset. The evaluated DCF-based trackers include BACF [26], deep spatial-temporal regularized correlation filters (DeepSTRCF) [11], ARCF [18], BiCF [28], AutoTrack [31], MSCF [22], IBRI [30], spatial regularization correlation filters with the hilbert-schmidt independence criterion (HSIC_SRCF) [36], DARTCF [13], efficient spatial-temporal UAV visual tracker with a temporal enhancement model update strategy (ETEU) [37], LDECF [19], and K-nearest neighbor correlation filters learning with p-Laplacian regularization (KNN-CFL) [38]. Unlike the UAVDT experiment, this comparison focuses only on correlation-filter-based methods, so that the effectiveness of the proposed improvements over recent DCF trackers can be examined more directly. The results of the compared trackers are collected from their original papers or publicly reported results.

Table 3. Comparisons of performance on the VisDrone-SOT dataset between CFLDT and other DCF-based trackers. The superscript § indicates that the tracker utilizes deep features.

Method	Venue	Prec.	Succ.
BACF	17'ICCV	0.774	0.567
DeepSTRCF§	18'CVPR	0.778	0.567
ARCF	19'ICCV	0.791	0.563
BiCF	20'ICRA	0.800	0.562
AutoTrack	20'CVPR	0.788	0.574
MSCF	21'ICRA	0.776	0.570
IBRI	21'TGRS	0.817	0.593
HSIC_SRCF	23'TIM	0.815	0.594
DARTCF	23'JSTARS	0.825	0.595
ETEU	25'SIVP	0.800	0.581
LDECF	25'TITS	0.758	0.550
KNN-CFL	25'TST	0.772	0.543
CFLDT	**Ours**	**0.763**	**0.667**

As shown in Table 3, CFLDT obtains a precision score of 0.763 and a success score of 0.667 on the VisDrone-SOT dataset. Compared with classical DCF trackers such as BACF, STRCF, ARCF, and AutoTrack, CFLDT achieves a higher success score, indicating better overlap-based tracking stability. Although several recent DCF trackers obtain higher precision scores, CFLDT achieves the best success score among the listed DCF-based methods. This suggests that the proposed DMMF and TVCM help the BACF-based framework maintain stable target localization under scale variation, background clutter, and rapid motion in practical UAV-view scenes.

4.6. Experimental results on the UAVDT

Table 4. Comparisons of performance on the UAVDT dataset between CFLDT and other trackers. The top three methods in precision (Prec.), success (Succ.), and FPS are highlighted in red, green, and blue, respectively. The superscript § indicates that the tracker utilizes deep features.

	Method	Venue	Prec.	Succ.	GPU	CPU
DCF-based	BACF	17'ICCV	0.706	0.442	-	51.9
	DeepSTRCF§	18'CVPR	0.667	0.437	6.8	-
	UDT+§	19'CVPR	0.674	0.442	73.3	-
	MN_ECO§	20'ACM	0.691	0.435	30.6	-
	fECO§	20'TIP	0.699	0.415	20.6	-
	CCF§	23'TCSVT	0.736	0.467	45	-
	BSTCF§	23'AI	0.685	0.441	19	-
	DARTCF	23'JSTARS	0.760	0.477	-	-
	MSEFCF	24'TIV	0.733	0.448	-	30
	SRECF	24'TMM	0.717	0.450	-	-
	ETEU	25'SIVP	0.750	0.478	-	-
	LDECF	25'TITS	0.744	0.448	-	44.9
	CFLDT	**Ours**	**0.749**	**0.523**	-	103
CNN-based	HiFT	21'ICCV	0.652	0.475	160.3	-
	F-SiamFC++	22'IJCNN	0.794	0.555	255.4	51.6
	TCTrack	22'CVPR	0.725	0.530	139.6	-
	ABDNet	23'RAL	0.755	0.553	130.2	-
	QRDT	24'TIM	0.606	0.477	-	-
ViT-based	STARK-ST101	21'ICCV	0.704	0.469	30.2	-
	HiT	23'ICCV	0.623	0.471	237.7	43.2
	Mixformerv2	23'NIPS	0.578	0.421	184.6	37.2
	BDTrack-ViT	24'arXiv	0.789	0.573	235.2	56.9
	TATTrack-ViT	24'TGRS	0.805	0.586	203.3	47.7
	SGLATrack-DeiT	25'CVPR	0.819	0.599	-	-
	ORTrack-DeiT	25'CVPR	0.834	0.601	-	-
	MMKLTrack	26'ESWA	0.791	0.573	-	-
	Aba-ViTrack++	26'TCSVT	0.847	0.614	-	-

To validate the effectiveness of the proposed CFLDT method, we conducted comparative experiments against three distinct types of trackers: DCF-based, CNN-based, and ViT-based trackers. The evaluated DCF-based trackers include BACF [26], DeepSTRCF [11], unsupervised deep tracking (UDT+) [39], meta-updater network with ECO (MN_ECO) [40], correlation filters based on complementary appearance model and reversibility reasoning (CCF) [41], background-aware and spatial-temporal regularized correlation filters (BSTCF) [42], DARTCF [13], MSEFCF [35], SRECF [23], ETEU [37], and LDECF [19]. The CNN-based trackers comprise hierarchical feature transformer (HiFT) [43], F-SiamFC++ [44], temporal contexts for aerial tracking (TCTrack) [45], adversarial blur-deblur network (ABDNet) [46], and QRDT [17]. The ViT-based trackers include spatio-temporal transformer tracker (STARK-ST101) [47], hierarchical vision transformer tracker (HiT) [48], MixFormerV2 [49], motion-blur-robust vision transformer tracker (BDTrack-ViT) [50], target-aware vision transformer tracker (TATTrack-ViT) [51], similarity-guided layer-adaptive vision transformer tracker (SGLATrack-DeiT) [52], occlusion-robust vision transformer tracker (ORTrack-DeiT) [53], adaptive and background-aware vision transformer tracker (Aba-ViTrack++) [54], and MMKLTrack [55].

As summarized in Table 4, the comparative results in terms of success rate, precision, and tracking speed on both GPU and CPU platforms are provided. Among DCF-based trackers, CFLDT achieves the best success score of 0.523, surpassing recent correlation-filter trackers such as DARTCF, MSEFCF, SRECF, ETEU, and LDECF. In terms of precision, CFLDT also achieves competitive performance among DCF-based methods, with a score of 0.749. Although several deep learning-based trackers obtain higher precision or success scores, they usually rely on deep features and GPU acceleration. In contrast, CFLDT maintains a high CPU speed of 103 FPS while achieving strong tracking accuracy, which confirms its practical advantage for resource-constrained UAV platforms.

4.7. Ablation study

This study improves the tracker from three main components: the DMMF for feature preprocessing, the MRSP for spatial penalty construction, and the TVCM-based environment-aware regularization term for model learning. To evaluate the contribution of each component to the overall performance, we design a fine-grained ablation study on the UAV123@10fps dataset. The *Baseline* denotes BACF. The symbols S, T, and DM denote the MRSP, the TVCM-based environment-aware regularization term, and the DMMF, respectively. Accordingly, eight variants are compared: $Baseline$, $Baseline+S$, $Baseline+T$, $Baseline+ST$, $Baseline+DM$, $Baseline+S+DM$, $Baseline+T+DM$, and $Baseline+ST+DM$. Table 5 reports the detailed precision and success results, while Figure 9 provides the branch-level comparison on the DTB70 dataset.

Table 5 presents the fine-grained ablation results on the UAV123@10fps dataset. Compared with the *Baseline*, the separated $Baseline+T$ variant improves the overall precision and success rate from 0.572 and 0.506 to 0.689 and 0.602, respectively. In addition, $Baseline+T$ improves all listed attributes compared with the *Baseline*, including full occlusion (FO), LR, PO, SV, and VC. These results provide direct experimental support for the TV-based constraint. By reducing unstable background-induced responses, the learned filter becomes more discriminative under occlusion, low resolution, and scale variation. When MRSP and TVCM are combined, $Baseline+ST$ obtains 0.690 in precision and 0.601 in success rate, confirming the effectiveness of the proposed regularization branch.

Table 5. Fine-grained ablation study on the UAV123@10fps dataset. S, T, and DM denote the MRSP, TVCM-based environment-aware regularization term, and DMMF, respectively. The best results are highlighted in bold.

Tracker	Prec.	Succ.	ARC	CM	FM	FO	IV	LR	PO	SO	SV	VC
Baseline(BACF)	0.572	0.506	0.478	0.532	0.407	0.336	0.430	0.431	0.467	0.605	0.525	0.491
$Baseline + S$	0.680	0.600	0.588	0.641	0.516	0.403	0.548	0.584	0.589	0.680	0.639	0.608
$Baseline + T$	0.689	0.602	0.603	0.650	0.517	0.452	0.580	0.589	0.600	0.661	0.650	0.594
$Baseline + ST$	0.690	0.601	0.604	0.651	0.517	0.450	0.581	0.589	0.601	0.661	0.650	0.595
$Baseline + DM$	0.698	0.609	0.631	0.676	0.559	0.462	0.573	0.610	0.610	0.700	0.668	0.606
$Baseline + S + DM$	0.693	0.605	0.622	0.674	0.553	0.462	0.573	0.610	0.610	0.699	0.663	0.597
$Baseline + T + DM$	0.707	0.615	0.631	0.680	0.570	**0.464**	0.604	0.608	0.623	**0.708**	0.669	0.624
$Baseline + ST + DM$(**CFLDT**)	**0.711**	**0.616**	**0.635**	**0.687**	**0.587**	**0.464**	**0.620**	**0.614**	**0.630**	0.707	**0.674**	**0.632**

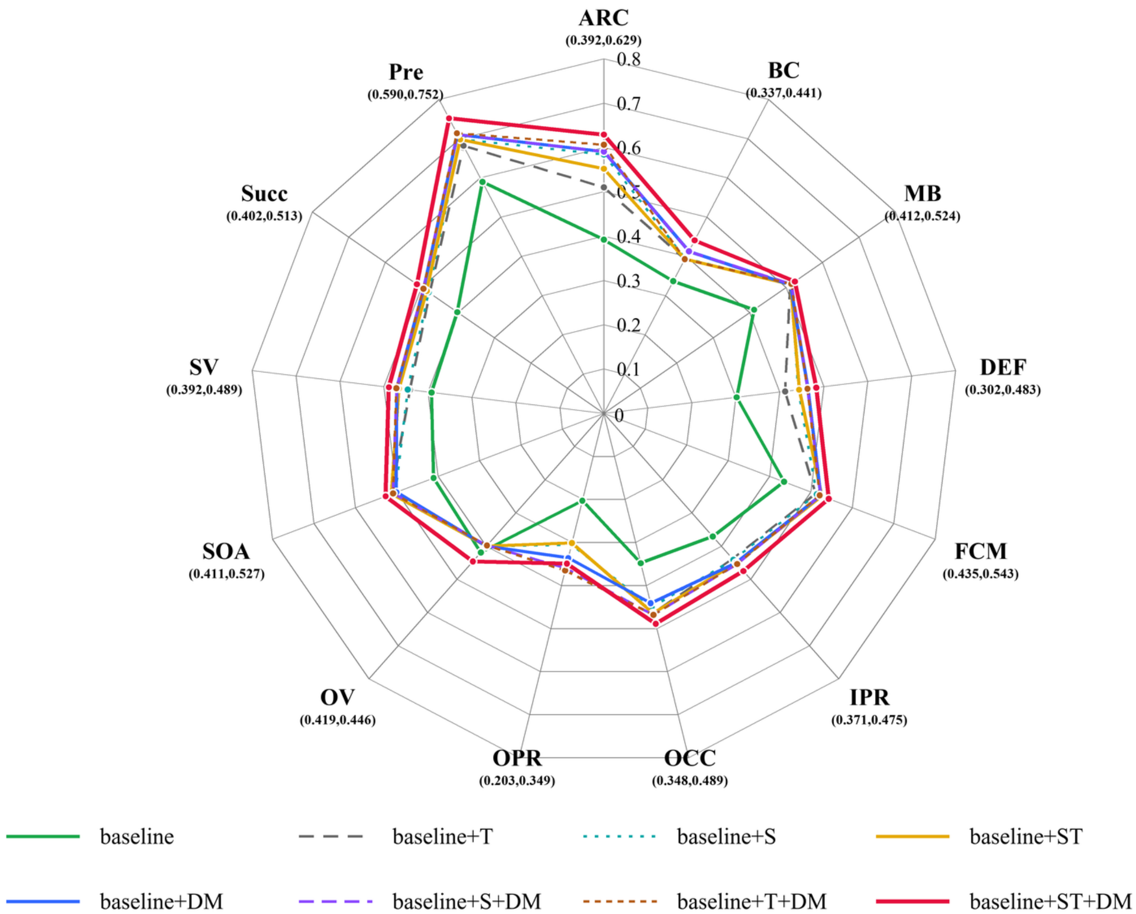


Figure 9. Ablation study on DTB70.

The DMMF also contributes clearly to tracking performance. *Baseline* + *DM* achieves 0.698 in precision and 0.609 in success rate, outperforming the *Baseline* by 12.6% and 10.3%, respectively. The complete model *Baseline* + ST + *DM* achieves the best overall performance, with a precision score of 0.711 and a success score of 0.616. Compared with *Baseline* + *DM*, the complete model further improves precision and success rate by 1.3% and 0.7%, respectively. In terms of attribute-based precision, *Baseline*+ST+*DM* achieves the best or tied-best results on most attributes, including ARC, CM, FM, FO, illumination variation (IV), LR, PO, SV, and VC. These results show that the DMMF, MRSP, and TVCM provide complementary contributions to the proposed CFLDT.

Experimental results on the DTB70 dataset further demonstrate the effectiveness of the main branches. As illustrated in Figure 9, the complete *Baseline* + ST + *DM* achieves a precision of 75.2% and a success rate of 51.3%, improving the *Baseline* by 16.2% and 11.1%, respectively.

The comparison among branch-level variants shows that both the DMMF branch and the regularization branch contribute to the tracking performance. Compared with the *Baseline*, the variants with DMMF generally obtain larger radar areas, indicating improved target localization under challenging UAV motion. Meanwhile, the variants with spatial penalty and TVCM-based regularization also show consistent improvements over the *Baseline* on most attributes.

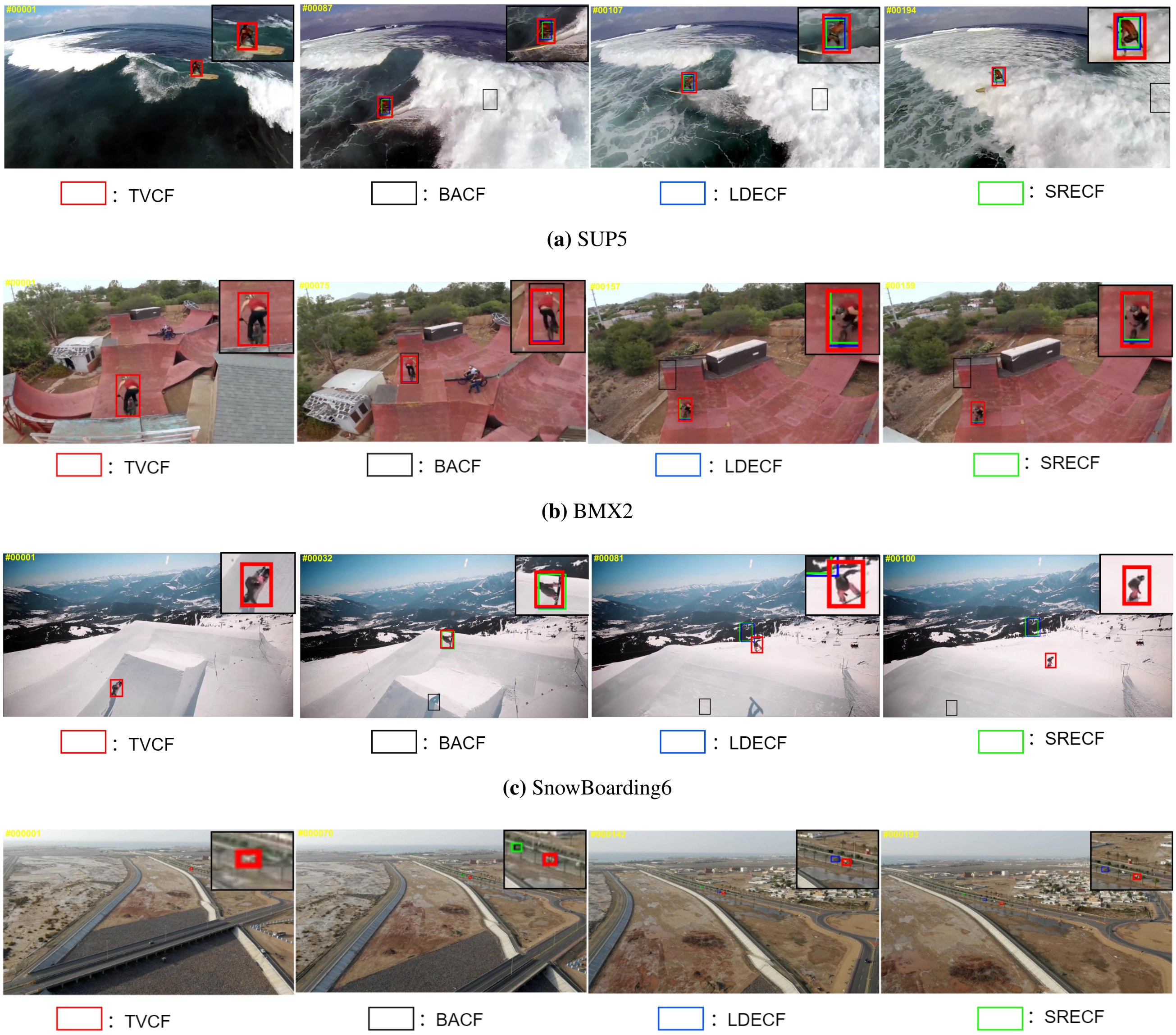
Furthermore, the complete *Baseline* + ST + *DM* achieves the best or near-best performance on most evaluated attributes. This indicates that motion fusion and model regularization are complementary rather than isolated improvements. Overall, the DTB70 ablation results verify that the proposed components jointly enhance the robustness of CFLDT in challenging UAV tracking scenarios.

4.8. Tracking analysis

To evaluate the real-time tracking performance of CFLDT, it was compared with three state-of-the-art trackers (BACF, LDECF, and SRECF) in a comparative experiment. To ensure the validity of the comparison, four challenging video sequences were selected. These sequences feature complex background environments and include various challenging factors such as deformation and occlusion. Conducting the comparison under such complex scenarios allows for a more comprehensive and convincing validation of CFLDT's real-time tracking capability. Figure 10 illustrates the tracking results on representative video sequences, while Figure 11 presents the real-time accuracy comparison.

Figure 10(a) shows the $SUP5$ video sequence of a surfer riding waves in the ocean. It presents dual challenges of large target deformation and dynamic background variation. Throughout the motion, the surfer undergoes continuous deformation. Simultaneously, the background alternates between dark seawater and bright waves, accompanied by significant luminance variations. These factors imposed increase the difficulty for the tracking task. Compared to other trackers, CFLDT achieves more accurate target localization. Furthermore, the precision curve in Figure 11(a) shows some fluctuations during tracking. Nevertheless, the proposed CFLDT method maintains high localization accuracy under these challenging conditions.

This paper selects the $BMX2$ video sequence in which the target and background exhibit high color similarity to evaluate the real-time tracking performance of CFLDT. As shown in Figure 10(b), the sequence depicts a rider in the red shirt performing freestyle movements in a predominantly red environment. As the rider performs various motions such as cycling and flipping, the target undergoes significant deformation and motion blur, further increasing the difficulty of tracking. Under these challenging conditions, the BACF fails to maintain tracking. However CFLDT consistently and stably locks onto the target. This result can likely be attributed to the TVCM, which enhances CFLDT's ability to discriminate between the target and background. Supported by the results shown in Figure 11(b), CFLDT demonstrates sustained and stable tracking performance throughout the entire sequence.



(a) SUP5

(b) BMX2

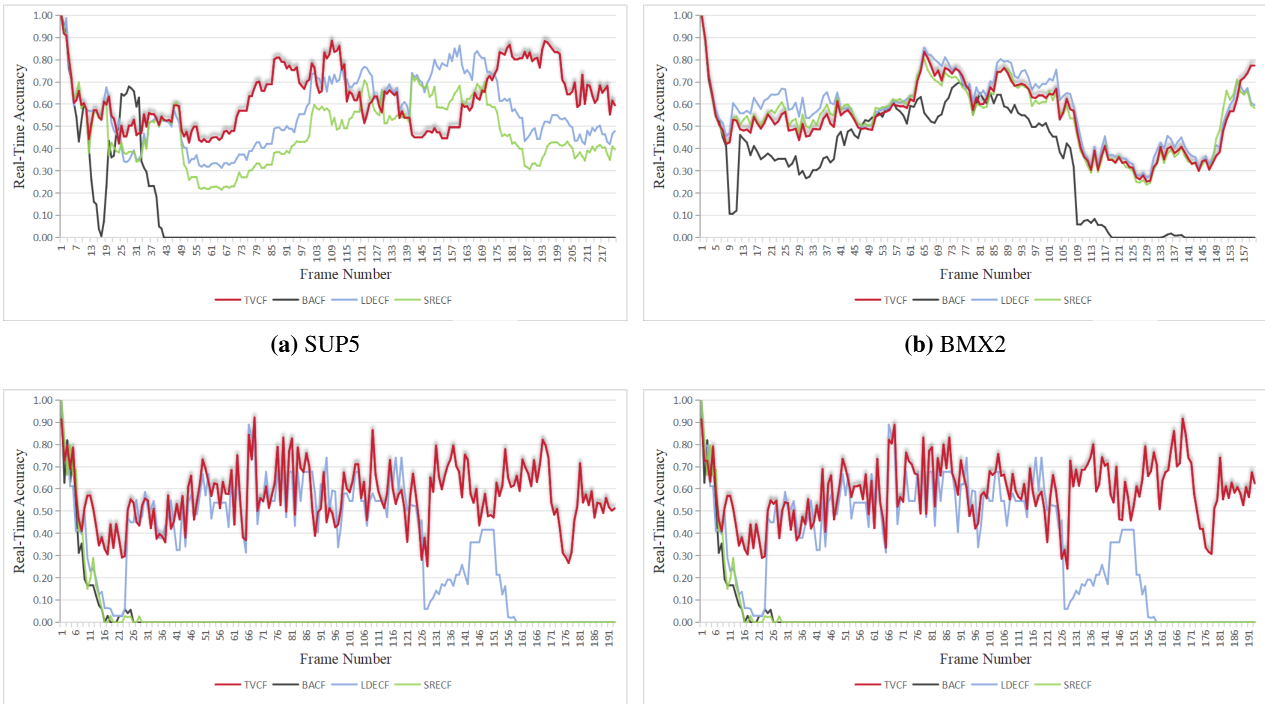
(c) SnowBoarding6

(d) truck4_1

Figure 10. Performance of different algorithms in real-time tracking.

In the tracking sequence *SnowBoarding6* depicted in Figure 10(c), a skier is moving at high speed while performing a series of challenging maneuvers such as flips. This results in significant rotation and deformation of the target. Under such complex conditions, the other three trackers quickly lose the target. In contrast, CFLDT maintains stable tracking. This robustness is likely due to the DMMF, which improves search region accuracy. Consequently, it effectively mitigates tracking drift caused by target deformation. As supported by the results in Figure 11(c), CFLDT demonstrates precise and stable localization performance across the entire video sequence.

In the tracking sequence *truck* 4_1 shown in Figure 10(d), the truck is moving rapidly and is frequently occluded by trees. The small size of the target and its high color similarity to the background pose significant challenges to the tracking task. Under these conditions, CFLDT optimizes the localization of the search region via the DMMF, effectively mitigating interference caused by occlusion. As illustrated by the real-time tracking accuracy in Figure 11(d), CFLDT maintains continuous and stable tracking of the target even in such a complex environment. A comparison of the four trackers shows clear performance differences. Among them, CFLDT demonstrates stronger robustness.



(a) SUP5 **(b)** BMX2

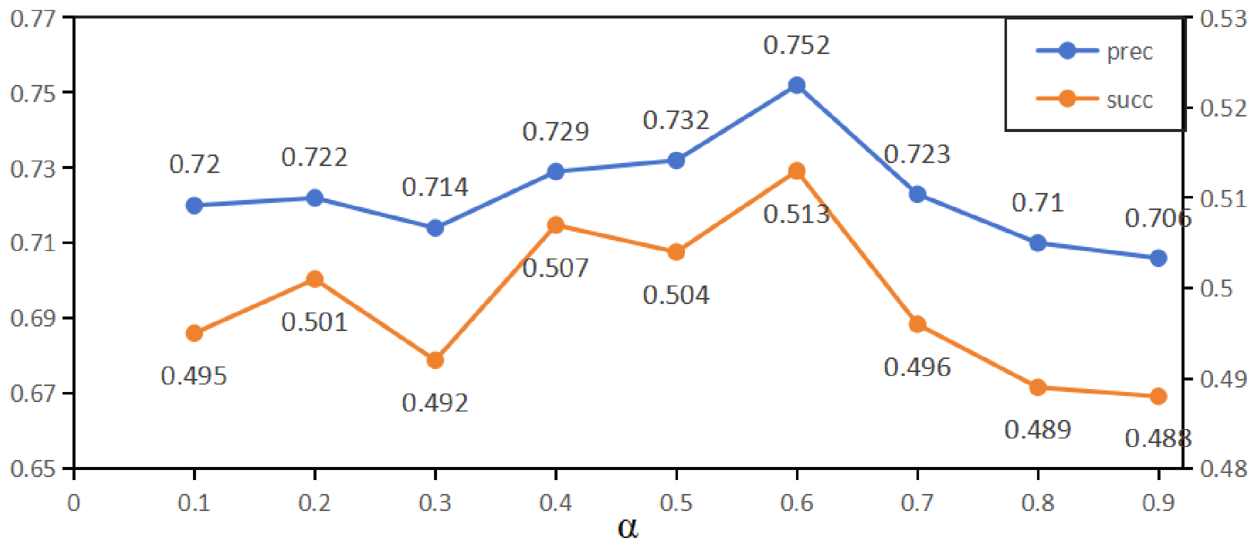
(c) SnowBoarding6 **(d)** truck4_1

Figure 11. The real-time accuracy of the tracking video sequence.

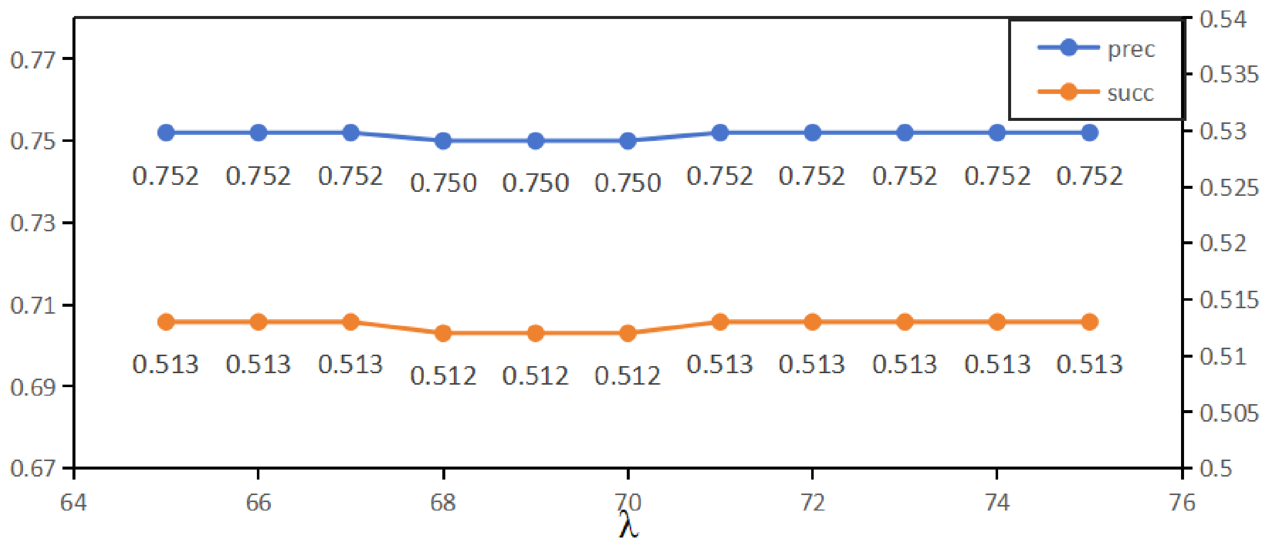
4.9. Parameter analysis

This paper integrates dual-modal cues in the DMMF through a weighted fusion strategy, wherein the weighting parameter α serves as a critical hyperparameter. To evaluate its impact, a sensitivity analysis of α was conducted. On the DTB70 dataset, the parameter α is evaluated across a range of [0.1, 1.0) with an incremental step size of 0.1. The results (as shown in Figure 12(a)) demonstrate that the value of α significantly influences both the precision and success rate of the CFLDT. As α increases, the tracker depends more on displacement information for rapid and precise localization. This reduces its reliance on potentially inaccurate motion priors. However, when α exceeds 0.6, the tracker over-relies on the appearance model's instantaneous response. This diminishes robustness to noise and occlusion, resulting in performance degradation. Based on these findings, α is set to 0.6 in this study.

This paper introduces the MRSP to impose an adaptive spatial penalty weight on the filter, where the parameter λ serves as a key hyperparameter controlling the penalty strength. To systematically evaluate its impact on tracking performance, we conducted a sensitivity analysis on the DTB70 dataset. Experiments were configured with λ increasing from 65 to 75 in steps of 1, with the results presented in Figure 12(b). The analysis indicates that the value of λ has a noticeable influence on both the precision and success rate of CFLDT. Although both metrics slightly decrease as λ increases, they remain largely stable overall, demonstrating a degree of robustness against parameter variations. Based on these findings, λ is set to 66 in this study.



(a) Key Parameters α.



(b) Key Parameters λ.

Figure 12. Analysis of key parameters.

5. Conclusions

Considering the impact of feature quality on tracking performance, this paper proposes a novel feature enhancement method. By introducing the DMMF, more accurate search region localization is achieved. Specifically, the module effectively predicts target motion trends by integrating motion response with correlation response information. This design mitigates the adverse effects of occlusion and significantly improves the tracker's capability to handle fast motion. These capabilities are critical for reliably monitoring high-speed vehicles in intelligent transportation systems. This enhancement effectively reduces the risk of target loss, thereby ensuring continuous tracking of vehicles even in complex urban traffic environments.

Furthermore, TVCM is employed to optimize the environment-aware regularization term. The incorporation of TVCM helps suppress noise interference and enables dynamic adaptation to environmental changes. This is crucial for handling challenges commonly encountered in traffic surveillance, such as varying weather conditions and lighting transitions. This capability allows for precise adjustment of the filter's sensitivity to background responses. Meanwhile,the MRSP is introduced. By imposing greater penalty weights on background regions, this matrix effectively enhances the discriminative capability of the tracker. This improvement enables more reliable vehicle identification amidst dense traffic flows. The proposed method demonstrates strong potential for smart city applications by consistently outperforming other algorithms across multiple UAV datasets. This proves its capability to enhance the accuracy and reliability of traffic monitoring.

In future work, we will continue to explore strategies for enhancing feature representation capabilities. For instance, we will investigate the deep integration of deep and hand-crafted features. We will also explore adaptive reliability-based fusion for the DMMF weight α to further improve robustness under unpredictable UAV motion. Progress in this direction is expected to broaden the application scope of UAV trackers, enabling their deployment in more complex and diverse real-world scenarios.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

This work was supported in part by the Guangdong Basic and Applied Basic Research Foundation under Grants 2024A1515012040, 2026A1515011166 and 2026A1515011429, and in part by the Science and Technology Planning Project of Guangzhou Under Grants 2024A03J0401.

Conflict of interest

The authors declare there is no conflict of interest.

References

1. S. Li, D. Yeung, Visual object tracking for unmanned aerial vehicles: A benchmark and new motion models, in *Proceedings of the AAAI Conference on Artificial Intelligence*, **31** (2017). https://doi.org/10.1609/aaai.v31i1.11205

2. M. Mueller, N. Smith, B. Ghanem, A benchmark and simulator for uav tracking, in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, **9905** (2016), 445–461. https://doi.org/10.1007/978-3-319-46448-0_27

3. D. Du, Y. Qi, H. Yu, Y. Yang, K. Duan, G. Li, et al., The unmanned aerial vehicle benchmark: Object detection and tracking, in *Proceedings of the European Conference on Computer Vision (ECCV)*, (2018), 370–386.

4. M. Bakirci, Internet of things-enabled unmanned aerial vehicles for real-time traffic mobility analysis in smart cities, *Comput. Electr. Eng.*, **123** (2025), 110313. https://doi.org/10.1016/j.compeleceng.2025.110313

5. W. Ouyang, J. Mu, X. Jing, Y. Wang, Efficient vehicle recognition and tracking for UAV-enabled intelligent transport systems: A multi-agent reinforcement learning method, *IEEE Trans. Intell. Transp. Syst.*, **26** (2025), 20930–20940. https://doi.org/10.1109/TITS.2025.3601740

6. J. Liu, H. Wang, C. Ma, Y. Su, X. Yang, Siamdmu: Siamese dual mask update network for visual object tracking, *IEEE Trans. Emerging Top. Comput. Intell.*, **8** (2024), 1656–1669. https://doi.org/10.1109/TETCI.2024.3353674

7. Z. Xin, J. Yu, X. He, Y. Song, H. Li, Siamraan: Siamese residual attentional aggregation network for visual object tracking, *Neural Process. Lett.*, **56** (2024), 98. https://doi.org/10.1007/s11063-024-11556-6

8. Z. Zhang, Z. Guo, L. Wang, Y. Li, Ctiftrack: Continuous temporal information fusion for object track, *Expert Syst. Appl.*, **262** (2025), 125654. https://doi.org/10.1016/j.eswa.2024.125654

9. K. Wang, Z. Wang, X. Zhang, M. Liu, Bstrack: Robust UAV tracking using feature extraction of bilateral filters and sparse attention mechanism, *Expert Syst. Appl.*, **267** (2025), 126202. https://doi.org/10.1016/j.eswa.2024.126202

10. H. Wu, Y. Chen, C. Wu, R. Zhang, K. Chen, A multi-scale cyclic-shift window transformer object tracker based on fast fourier transform, *Electron. Res. Arch.*, **33** (2025), 3638–3672. https://doi.org/10.3934/era.2025162

11. F. Li, C. Tian, W. Zuo, L. Zhang, M. Yang, Learning spatial-temporal regularized correlation filters for visual tracking, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2018), 4904–4913.

12. S. Moorthy, Y. H. Joo, Adaptive spatial-temporal surrounding-aware correlation filter tracking via ensemble learning, *Pattern Recogn.*, **139** (2023), 109457. https://doi.org/10.1016/j.patcog.2023.109457

13. H. Zhang, Y. Li, Y. Yang, Y. Feng, Y. Li, C. Deng, et al., UAV tracking based on correlation filters with dynamic aberrance-repressed temporal regularizations, *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, **16** (2023), 7749–7762. https://doi.org/10.1109/JSTARS.2023.3306273

14. K. Chen, Y. Chen, R. Zhang, Y. Chen, H. Han, Y. He, et al., Correlation filter object tracking algorithm based on spatial and channel attention mechanism, *Electron. Res. Arch.*, **33** (2025), 4857–4892. https://doi.org/10.3934/era.2025219

15. F. Zhang, S. Ma, Y. Zhang, Z. Qiu, Perceiving temporal environment for correlation filters in real-time UAV tracking, *IEEE Signal Process. Lett.*, **29** (2021), 6–10. https://doi.org/10.1109/LSP.2021.3120943

16. C. Wu, J. Shen, K. Chen, Y. Chen, Y. Liao, UAV object tracking algorithm based on spatial saliency-aware correlation filter, *Electron. Res. Arch.*, **33** (2025), 1446. https://doi.org/10.3934/era.2025068

17. Y. Xue, T. Shen, G. Jin, L. Tan, N. Wang, L. Wang, et al., Handling occlusion in UAV visual tracking with query-guided redetection, *IEEE Trans. Instrum. Meas.*, **73** (2024), 1–17. https://doi.org/10.1109/TIM.2024.3440378

18. Z. Huang, C. Fu, Y. Li, F. Lin, P. Lu, Learning aberrance repressed correlation filters for real-time UAV tracking, in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, (2019), 2891–2900.

19. Y. F. Yu, Z. Chen, Y. Zhang, C. Zhang, W. Ding, Learning dynamic-sensitivity enhanced correlation filter with adaptive second-order difference spatial regularization for UAV tracking, *IEEE Trans. Intell. Transp. Syst.*, **26** (2025), 7211–7230. https://doi.org/10.1109/TITS.2025.3533953

20. Y. Zhang, Y. F. Yu, L. Chen, W. Ding, Robust correlation filter learning with continuously weighted dynamic response for UAV visual tracking, *IEEE Trans. Geosci. Remote Sens.*, **61** (2023), 1–14. https://doi.org/10.1109/TGRS.2023.3325337

21. J. Wen, H. Chu, Z. Lai, T. Xu, L. Shen, Enhanced robust spatial feature selection and correlation filter learning for UAV tracking, *Neural Netw.*, **161** (2023), 39–54. https://doi.org/10.1016/j.neunet.2023.01.003

22. G. Zheng, C. Fu, J. Ye, F. Lin, F. Ding, Mutation sensitive correlation filter for real-time UAV tracking with adaptive hybrid label, in *2021 IEEE International Conference on Robotics and Automation (ICRA)*, (2021), 503–509. https://doi.org/10.1109/ICRA48506.2021.9561931

23. C. Fu, J. Jin, F. Ding, Y. Li, G. Lu, Spatial reliability enhanced correlation filter: An efficient approach for real-time UAV tracking, *IEEE Trans. Multimedia*, **26** (2024), 4123–4137. https://doi.org/10.1109/TMM.2021.3118891

24. F. Lin, C. Fu, Y. He, W. Xiong, F. Li, Recf: Exploiting response reasoning for correlation filters in real-time UAV tracking, *IEEE Trans. Intell. Transp. Syst.*, **23** (2021), 10469–10480. https://doi.org/10.1109/TITS.2021.3094654

25. Y. Chen, K. Chen, Four mathematical modeling forms for correlation filter object tracking algorithms and the fast calculation for the filter, *Electron. Res. Arch.*, **32** (2024), 4684–4714. https://doi.org/10.3934/era.2024213

26. H. K. Galoogahi, A. Fagg, S. Lucey, Learning background-aware correlation filters for visual tracking, in *Proceedings of the IEEE International Conference on Computer Vision*, (2017), 1135–1143.

27. C. Fu, J. Xu, F. Lin, F. Guo, T. Liu, Z. Zhang, Object saliency-aware dual regularized correlation filter for real-time aerial tracking, *IEEE Trans. Geosci. Remote Sens.*, **58** (2020), 8940–8951. https://doi.org/10.1109/TGRS.2020.2992301

28. F. Lin, C. Fu, Y. He, F. Guo, Q. Tang, Bicf: Learning bidirectional incongruity-aware correlation filter for efficient UAV object tracking, in *2020 IEEE International Conference on Robotics and Automation (ICRA)*, (2020), 2365–2371. https://doi.org/10.1109/ICRA40945.2020.9196530
29. M. Danelljan, G. Bhat, F. S. Khan, M. Felsberg, Eco: Efficient convolution operators for tracking, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (2017), 6638–6646.
30. C. Fu, J. Ye, J. Xu, Y. He, F. Lin, Disruptor-aware interval-based response inconsistency for correlation filters in real-time aerial tracking, *IEEE Trans. Geosci. Remote Sens.*, **59** (2020), 6301–6313. https://doi.org/10.1109/TGRS.2020.3030265
31. Y. Li, C. Fu, F. Ding, Z. Huang, G. Lu, Autotrack: Towards high-performance visual tracking for UAV with automatic spatio-temporal regularization, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2020), 11923–11932.
32. J. Ye, C. Fu, F. Lin, F. Ding, S. An, G. Lu, Multi-regularized correlation filter for UAV tracking and self-localization, *IEEE Trans. Ind. Electron.*, **69** (2021), 6004–6014. https://doi.org/10.1109/TIE.2021.3088366
33. Y. Zhang, Y. F. Yu, K. K. Huang, Y. Wang, Channel attentional correlation filters learning with second-order difference for UAV tracking, *IEEE Geosci. Remote Sens. Lett.*, **20** (2023), 1–5. https://doi.org/10.1109/LGRS.2023.3311441
34. J. Lin, J. Peng, J. Chai, Real-time UAV correlation filter based on response-weighted background residual and spatio-temporal regularization, *IEEE Geosci. Remote Sens. Lett.*, **20** (2023), 1–5. https://doi.org/10.1109/LGRS.2023.3272522
35. Y. F. Yu, Y. Zhang, L. Chen, P. Ge, C. L. P. Chen, Multi-scale enhanced features correlation filters learning with dual second-order difference for UAV tracking, *IEEE Trans. Intell. Veh.*, **9** (2024), 3232–3245. https://doi.org/10.1109/TIV.2024.3355171
36. Z. An, X. Wang, B. Li, J. Fu, Learning spatial regularization correlation filters with the Hilbert-Schmidt independence criterion in RKHS for UAV tracking, *IEEE Trans. Instrum. Meas.*, **72** (2023), 1–12. https://doi.org/10.1109/TIM.2023.3265106
37. L. Chen, Y. Liu, Y. Wang, An efficient spatial-temporal UAV visual tracker with the temporal enhancement model update strategy, *Signal Image Video Process.*, **19** (2025), 217. https://doi.org/10.1007/s11760-024-03772-3
38. Y. F. Yu, X. Tan, Y. Zhang, L. Chen, W. Ding, K-nearest neighbor correlation filters learning with p-laplacian regularization for visual tracking, *Tsinghua Sci. Technol.*, (2025). https://doi.org/10.26599/TST.2025.9010049
39. N. Wang, Y. Song, C. Ma, W. Zhou, W. Liu, H. Li, Unsupervised deep tracking, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2019), 1308–1317.
40. J. Zhao, K. Dai, D. Wang, H. Lu, X. Yang, Online filtering training samples for robust visual tracking, in *Proceedings of the 28th ACM International Conference on Multimedia*, (2020), 1488–1496. https://doi.org/10.1145/3394171.3413930

41. B. Wang, W. Li, B. Zhang, Y. Liu, J. Du, Correlation filters for UAV online tracking based on complementary appearance model and reversibility reasoning, *IEEE Trans. Circuits Syst. Video Technol.*, **34** (2024), 3983–3997. https://doi.org/10.1109/TCSVT.2023.3325672
42. J. Zhang, Y. He, W. Feng, J. Wang, N. N. Xiong, Learning background-aware and spatial-temporal regularized correlation filters for visual tracking, *Appl. Intell.*, **53** (2023), 7697–7712. https://doi.org/10.1007/s10489-022-03868-8
43. Z. Cao, C. Fu, J. Ye, B. Li, Y. Li, Hift: Hierarchical feature transformer for aerial tracking, in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, (2021), 15457–15466.
44. I. Sosnovik, A. Moskalev, A. Smeulders, Scale equivariance improves siamese tracking, in *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, (2021), 2764–2773.
45. Z. Cao, Z. Huang, L. Pan, S. Zhang, Z. Liu, C. Fu, Tctrack: Temporal contexts for aerial tracking, in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2022), 14778–14788.
46. H. Zuo, C. Fu, S. Li, K. Lu, Y. Li, C. Feng, Adversarial blur-deblur network for robust UAV tracking, *IEEE Robot. Autom. Lett.*, **8** (2023), 1101–1108. https://doi.org/10.1109/LRA.2023.3236584
47. B. Yan, H. Peng, J. Fu, D. Wang, H. Lu, Learning spatio-temporal transformer for visual tracking, in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, (2021), 10428–10437.
48. B. Kang, X. Chen, D. Wang, H. Peng, H. Lu, Exploring lightweight hierarchical vision transformers for efficient visual tracking, in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, (2023), 9578–9587.
49. Y. Cui, T. Song, G. Wu, L. Wang, Mixformerv2: Efficient fully transformer tracking, *Adv. Neural Inf. Process. Syst.*, **36** (2023), 58736–58751.
50. Y. Wu, X. Wang, D. Zeng, H. Ye, X. Xie, Q. Zhao, et al., Learning motion blur robust vision transformers with dynamic early exit for real-time UAV tracking, *Expert Syst. Appl.*, **297** (2024), https://doi.org/10.1016/j.eswa.2025.129445
51. S. Li, X. Yang, X. Wang, D. Zeng, H. Ye, Q. Zhao, Learning target-aware vision transformers for real-time UAV tracking, *IEEE Trans. Geosci. Remote Sens.*, **62** (2024), 1–18. https://doi.org/10.1109/TGRS.2024.3417400
52. C. Xue, B. Zhong, Q. Liang, Y. Zheng, N. Li, Y. Xue, et al., Similarity-guided layer-adaptive vision transformer for UAV tracking, in *Proceedings of the Computer Vision and Pattern Recognition Conference*, (2025), 6730–6740.
53. Y. Wu, X. Wang, X. Yang, M. Liu, D. Zeng, H. Ye, et al., Learning occlusion-robust vision transformers for real-time UAV tracking, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2025), 17103–17113.
54. X. Yang, D. Zeng, X. Wang, H. Ye, X. Xie, Q. Zhao, et al., Towards real-time UAV tracking with adaptive and background-aware vision transformers, *IEEE Trans. Circuits Syst. Video Technol.*, **36** (2026), 7395–7410. https://doi.org/10.1109/TCSVT.2026.3652352

55. K. Yin, J. Feng, S. Dong, Y. Long, Mmkltrack: Illumination-robust UAV tracking with progressive localization, *Expert Syst. Appl.*, **300** (2026), 130207. https://doi.org/10.1016/j.eswa.2025.130207