Electronic Research Archive

ERA, 33(8): 4857–4892.
DOI: 10.3934/era.2025219
Received: 09 May 2025
Revised: 25 June 2025
Accepted: 24 July 2025
Published: 20 August 2025

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

*Research article*

# Correlation filter object tracking algorithm based on spatial and channel attention mechanism

**Kaiwei Chen[1], Yingpin Chen[1,2,*], Ronghuan Zhang[1], Yiling Chen[1], Hongshuo Han[1], Yijing He[1], Wenjie Xu[1], Wenbing Ye[1] and Jinghao Li[1]**

[1] School of Physics and Information Engineering, Minnan Normal University, Zhangzhou 363000, China
[2] Key Laboratory of Light Field Manipulation and System Integration Applications in Fujian Province, Minnan Normal University, Zhangzhou 363000, China

* **Correspondence:** Email: cyp1707@mnnu.edu.cn.

**Abstract:** Correlation filter algorithms are widely used in the field of object tracking due to their excellent tracking performance and real-time tracking efficiency. Traditional correlation filter tracking methods focus on three aspects of improving algorithm performance: feature representation, spatial regularization techniques, and temporal smoothness. However, these methods overlook the removal of redundant and interfering information in the filters, as well as the protection of filters in occluded scenes, which leads to poor performance in complex scenarios such as cluttered backgrounds and occlusions. It is noted that interfering information exhibits sparsity but often lacks the structural property of combinatorial sparsity. Inspired by this, this paper proposes a correlation filter object tracking algorithm based on spatial and channel attention mechanisms. Specifically, the algorithm introduces fiber group sparsity constraints in the spatial direction of the filter and slice group sparsity constraints in the row, column, and channel directions. In this way, the structural sparsity property of the filter is further exploited to allocate attention to the spatial and channel features, thereby removing redundant and interfering information. In addition, the reliability analysis of the best candidate samples is performed by a history template pool to decide whether the filter should be updated or not to avoid the tracking failure problem caused by filter degradation in occlusion scenarios. Experiments on several datasets show that the proposed method outperforms other state-of-the-art trackers and achieves good performance in terms of accuracy and robustness.

**Keywords:** correlation filter; object tracking; structural sparsity; attention mechanism; history template pool

## 1. Introduction

Object tracking technologies [1,2] constitute a critical research direction in computer vision, and demonstrate broad application prospects across domains such as disaster rescue and search operations [3], Unmanned Aerial Vehicle (UAV) surveillance [4,5], satellite video monitoring [6,7], and military reconnaissance [8]. The UAV video object tracking task is of significant academic value due to its irreplaceability in many special scenarios. To enhance the performance of UAV visual object trackers, researchers have explored various innovative methodologies, which achieved substantial advancements in recent years. Nevertheless, the complexity of aerial tracking environments, object variability, and strict real-time accuracy requirements make precise and fast UAV object tracking challenging in practical applications.

UAV object tracking [9–11] recognizes and localizes specific target objects via consecutive image frames to adapt to diverse dynamic scenarios. Existing UAV tracking methods can be broadly categorized into two groups: generative methods [12] and discriminative methods [13]. Generative methods typically employ similarity-based measurements. It obtains key features for data association by learning the appearance or movement patterns of the object. Subsequently, the optimal candidate for the next position is identified. However, generative methods solely focus on the object itself while neglecting contextual background information. Discriminative methods treat the tracking task as a classification or regression problem. These approaches specifically extract positive samples from foreground regions and negative samples from background areas to train classifiers for object region identification in subsequent frames. The tracking framework based on a discriminant correlation filter (DCF) [14] has become the mainstream tracking framework in the field of object tracking due to its computational efficiency and robust performance. It is especially suitable for resource-constrained edge computing platforms such as UAVs and autonomous vehicles.

Most DCF trackers focus on a single object template. However, the appearance information carried by a single template is limited and fuzzy, which will increase the difficulty of object tracking. Most DCF trackers update their models at a fixed learning rate, which aggravates error accumulation and leads to model degradation. To avoid filter model degradation, some state-of-the-art DCF trackers introduce temporal regularization [15,16] into filter learning. For example, the spatial-temporal regularized correlation filter (STRCF) [17] tracker introduces a temporal regularization factor based on the spatially regularized correlation filter (SRDCF) [18] tracker to prevent the filter from drifting due to changes in the object's appearance, thus enabling more robust tracking under conditions of occlusion or significant appearance variations. The existing methods primarily focus on utilizing the historical object appearance information [19–21] to determine whether the object is occluded in a new frame. For instance, Huang et al. [22] proposed an anomaly-suppressing correlation filter, which suppresses the anomalous response due to the occlusion scene with a shifted peak energy regularity constraint on the responses of neighboring filters. Some scholars [23] have utilized the statistical properties of the response map to enhance the tracker's immunity to occlusion, such as the average peak energy and the peak sidelobe ratio. Occlusion scenes tend to mislead the tracker to learn the wrong appearance model, which is the main reason for destroying tracking robustness and is one of

the difficulties in object tracking techniques. The algorithm in this paper utilizes the object saliency feature [24] to determine whether an object exists in the search region, and then decides whether to update the filter or not.

To improve the feature representation capabilities, traditional DCF trackers typically employ pre-trained deep neural network features [25–27]. However, when deep features are introduced, different channels exhibit distinct properties. A large number of channels contain information unrelated to the object. As a result, these features may contain a lot of redundant and irrelevant information, which will affect the tracking performance. To solve this problem, the group feature selection method for discriminative correlation filter (GFS-DCF) [28] tracker proposes a joint group feature selection method. It aims to highlight relevant features and reduce the information redundancy of multi-channel features to improve the discriminative ability and filter stability of the selected features. However, the GFS-DCF tracker ignores the compact representation of the structure. The enhanced robust spatial feature selection and correlation filter (EFSCF) [29] imposes joint sparsity constraints along the row and column directions of the filter to enhance its structural sparsity. To this end, this paper proposes a correlation filter object tracking algorithm based on spatial and channel attention mechanisms (SCACF). The algorithm realizes robust visual object tracking by endowing the tracking system with an attention mechanism. The SCACF jointly performs sparsity [30] feature selection in both spatial and channel dimensions to highlight the most discriminative and important visual information about the object.

Despite significant progress in existing UAV tracking frameworks, there are still many challenges that need to be overcome. In particular, traditional UAV correlation filter algorithms use a strategy of frequently updating the object templates and filters. This mechanism is similar to continuously acquiring new information to adapt to a dynamically changing environment. Nevertheless, when the object encounters heavy occlusion, the frequent update strategy may lead to template contamination, which, in turn, triggers tracking performance degradation. In addition, traditional filters often contain a large amount of redundant and irrelevant information when processing multi-channel features. This ignores the sparsity of different channel features, which leads to difficulties for UAV tracking systems to effectively distinguish between interfering objects and real objects. To address the above limitations, the main contributions of this paper are as follows:

1) To fully exploit the structural sparsity property of the filter, different features are assigned with different channel attention, row-group sparsity attention, column-group sparsity attention, and fiber-group sparsity spatial attention. The attention allocation scheme is utilized to accomplish feature reduction and dimensionality reduction, which significantly improves the performance of the tracker.

2) To utilize a reliability analysis of the best candidate samples through an anti-occlusion strategy and an optimization of model updates using a thresholding mechanism to avoid updating the correlation filter at a constant learning rate. This approach takes the diversity of samples into account and effectively avoids the filter degradation problem caused by occlusion. Thus, it is expected to meet the tracking challenges in complex environments.

3) On the benchmark test platforms of the UAV123 [31], UAV20L, DTB70 [32], and OTB [33] datasets, we conducted extensive experimental evaluations to verify the effectiveness of the proposed method. The experimental results show that the SCACF filter outperforms other state-of-the-art trackers and achieves robust UAV object tracking.

The rest of this paper is arranged as follows: Section 2 reviews the visual object tracking based on the DCF; Section 3 details the proposed method; Section 4 shows the experimental results of this

algorithm on multiple datasets and discusses its limitations; and Section 5 draws a brief conclusion and puts forward the prospect of follow-up work.

## 2. Related works

### *2.1. DCF-based visual object tracking*

Bolme et al. proposed the minimum output sum of squared error (MOSSE) filter [34], thus marking the first application of a correlation filter (CF) in the field of object tracking. MOSSE utilizes the fast Fourier transform (FFT) to convert correlation operations into simple entry-wise multiplication operations, and achieves tracking speeds of up to 700 frames per second (FPS). However, the inherent linear nature of the MOSSE algorithm makes its modeling capabilities limited. To address this limitation, Henriques et al. proposed the circulant structure of tracking-by-detection with kernels (CSK) [35] tracker, which extends MOSSE to nonlinear filters by introducing a circulant shift and kernel method. However, CSK is still limited to processing single-channel grayscale features, which restricts its differentiation ability. Henriques et al. introduced the kernelized correlation filter (KCF) [36] tracker, which extends single-channel features to histogram of oriented gradients (HOG) [37] features that significantly improve the filter performance while maintaining a high execution speed. The Staple [38] algorithm applies both CourseNetworking (CN) [39] and HOG features to the correlation filter to improve the robustness of deformation and illumination changes. However, hand-crafted features can only extract a single level of features and often do not effectively represent the multi-level structure and characteristics of the object. To improve the robustness of UAV tracking performance, most DCF trackers use several types of features [40], such as hand-crafted features like HOG, CN, and convolutional neural networks [41,42] (CNN) deep features. Compared with traditional hand-crafted features, deep features have rich semantic information and powerful discriminative capabilities. This enables the learned model to achieve robust tracking in challenging scenarios such as fast motion and light changes.

Localizing an object in consecutive video frames is the core task of visual object tracking. DCF-based methods aim at the initial predicted position of the object in a given video. Suppose that the predicted position of the tracked object has already been determined in the frame $t$. To predict the position of the object in the next frame (i.e., to localize the tracked object in the frame $t+1$, the DCF training filter $\mathcal{W}_t \in \mathbb{R}^{N\times N\times C}$ is employed with a pair of training samples $\{\mathcal{X}_t, Y\}$, where $Y \in \mathbb{R}^{N\times N}$ is the optimal object detection response map, and $\mathcal{X}_t \in \mathbb{R}^{N\times N\times C}$ is a tensor consisting of the $C$ channel features extracted from the current frame. To solve the multichannel correlation filter $\mathcal{W}_t$, the object tracking problem is transformed into the following least squares problem with a penalty term:

$$\mathcal{W}_t = \underset{\mathcal{W}_t}{argmin}\frac{1}{2}\left\|\sum_{k=1}^{C} W_t^{\{k\}} * X_t^{\{k\}} - Y\right\|_2^2 + R(\mathcal{W}_t), \tag{1}$$

where $*$ is the circular convolution operator, $W_t^{\{k\}} \in \mathbb{R}^{N\times N}$ is the discriminant correlation filter for the $k$th channel, $X_t^{\{k\}} \in \mathbb{R}^{N\times N}$ denotes the $k$th channel feature, and $R(\mathcal{W}_t) = \lambda\sum_{k=1}^{C}\left\|W_t^{\{k\}}\right\|_2^2$ is the regularization term. In the frequency domain, closed solutions can be obtained with relatively few operations. In the tracking phase of the DCF algorithm, the filter learned in the first frame can be

used to locate the tracking object in the second frame and update the filter learned in the other frames as follows:

$$\mathcal{W}_t = \alpha\mathcal{W} + (1-\alpha)\mathcal{W}_{t-1}, \tag{2}$$

where $\alpha \in [0,1]$ is a predefined update rate.

It is assumed that the search window in the frame $t+1$ is known. First, the multi-channel feature of the $t+1$ frame is extracted. Then, the multi-channel feature of the $t+1$ frame is calculated with the correlation filter $\widehat{W}_t^{\{k\}}$ learned in the frame $t$, which can effectively estimate the frequency domain response mapping as follows:

$$\hat{R} = \sum_{k=1}^{C} \hat{X}_{t+1}^{\{k\}} \odot \widehat{W}_t^{\{k\}}, \tag{3}$$

where $\odot$denotes the entry-wise multiplication operation, and the superscript $\wedge$ denotes the discrete Fourier transform (DFT). Equation (3) is the maximum response mapping obtained by an inverse DFT (i.e., the corresponding position of the predicted object).

### *2.2. Deep learning-based visual object tracking*

In recent years, deep learning-based [43–46] tracking methods have achieved state-of-the-art tracking performance. Deep learning-based tracking methods can usually be categorized into the following two groups: online fine-tuning tracking methods and offline pre-trained tracking methods. One example of online fine-tuning tracking methods is the multi-domain convolutional neural network (MDNet) tracker [47], which pre-trains the network through multi-domain learning, thereby fine-tuning domain-specific layers online during tracking to adapt to changes in the appearance of the object. The real-time MDNet (RT-MDNet) tracker [48] optimizes computational efficiency based on MDNet and achieves real-time tracking. Although this type of tracking method achieves robust tracking by dynamically updating the model, it is computationally expensive. One example of offline pre-trained tracking methods [49] is the fully-convolutional Siamese (SiamFC) network [50], which trains a Siamese network offline, thereby utilizing the structure of a Siamese network and the characteristics of a fully convolutional network to achieve real-time and accurate tracking results. The recurrent aggregation Siamese network (RASNet) [51] introduces spatial and channel attention in the Siamese network, thereby combining residual learning to enhance feature representation. The dynamic Siamese network (DSiamM) [52] introduces an object appearance transformation layer and a background suppression layer to adapt to the object's changing appearance and background interference in the video sequence. The Siamese region proposal network (SiamRPN) [53] combines Siamese networks and an RPN to effectively deal with the object's appearance changes and occlusion problems, thus realizing high-accuracy object tracking. SiamRPN++ [54] uses a spatially aware sampling strategy to mitigate the translational invariance problem and ResNet-50 as a backbone network to increase the network depth. However, Siamese network-based tracking methods only focus on the object and ignore the rich background information. It is easy to misjudge when there are similar objects in the video. To solve this problem, the Transformer model [55,56] is used in the field of visual object tracking. However, the deep learning model requires large memory and computational resources and are not suitable for UAV tracking, which has limited onboard resources and requires low power consumption. Despite

these challenges, deep learning-based methods still have huge potential in the field of UAV tracking.

### *2.3. Attention mechanism-based correlation filter tracking method*

DCF-based tracking methods enhance the representation of object features by introducing an attention mechanism to improve the robustness and accuracy of tracking. An aberrance repressed correlation filter (ARCF) [22] introduces a channel attention mechanism, thereby enhancing the discriminative ability of object features by dynamically adjusting the weights of different channel features. The discriminative correlation filter with channel and spatial reliability (CSR-DCF) [57] combines channel reliability and spatial reliability mechanisms to enhance tracking performance in complex environments. Deformable Siamese attention networks (SiamAttn) [58] combine the self-attention mechanism in a Siamese network framework to optimize the extraction and matching process of object features. Stark [59] introduces a spatio-temporal attention mechanism that focuses on both the spatial location and temporal continuity of the object. The visual attention learning and an anti-occlusion mechanism (VALACF) [60] introduces a visual attention mechanism and a local perception strategy to dynamically focus on key regions of the object and optimize the spatial distribution of features. The deep-feature-based asymmetrical background-aware correlation filter (DeepABCF) [21] combines deep features extracted by deep learning with a binary correlation filter and introduces the attention mechanism to dynamically weigh the features. These tracking methods based on the attention mechanism perform well in complex scenes (e.g., occlusion, fast motion, background clutter, etc.) and are an important research direction in the field of object tracking.

## 3. Proposed methods

### *3.1. Group feature selection for DCF*

In DCF-based visual object tracking, multi-channel features are usually extracted from a larger search window to capture contextual information about the object. However, only a smaller window region is adopted in the subsequent filter and response computation. This leads to a large amount of unutilized information, increasing the irrelevance and redundancy of the image features, thus reducing the efficiency and robustness of the tracking algorithm. Spatial feature selection or regularization is widely adopted in existing DCF-based trackers to solve this problem. However, the current methods mainly focus on the spatial dimension and fail to adequately address the information redundancy and noise between feature channels.

To this end, this paper proposes a correlation filter method based on the spatial and channel attention mechanism (SCACF). Compared to traditional DCF algorithms, SCACF employs a form of fiber group sparsity in the spatial and three-slice group sparsity in the channel for dimensionality reduction of multichannel features in the traditional DCF optimization task. Joint spatial and channel group sparsity feature selection effectively solves the problem of information redundancy and irrelevance in the high-dimensional multichannel. In addition, to address the problem of creating too much variability in the prediction, the SCACF tracker introduces a temporal constraint term to smooth the filter between temporal frames. The anti-occlusion strategy introduced by SCACF can fully exploit the feature diversity of the object, specifically including the appearance diversity, the temporal diversity, and scale diversity of the object. The strategy solves the tracking drift problem in occlusion

scenarios by evaluating the reliability of the optimal candidate sample. The workflow of the SCACF tracker is shown in Figure 1. Assuming that the current frame learned by the correlation filter is frame $t$, the subscript "$t$" is omitted from the following equations for brevity. According to Eq (1), the objective function of SCACF can be obtained as follows:

$$\mathcal{W} = \underset{\mathcal{W}}{argmin} \frac{1}{2} \left\| \sum_{k=1}^{C} W_t^{\{k\}} * X_t^{\{k\}} - Y \right\|_2^2 + \lambda_1 R_{Fiber}(\mathcal{W}) + \lambda_2 R_{Slice}(\mathcal{W}) + \frac{\lambda_3}{2} R_T(\mathcal{W}), \quad (4)$$

where $R_{Fiber}(\mathcal{W})$ is a fiber group sparsity constraint for the spatial feature selection, $R_{Slice}(\mathcal{W})$ is a slice group sparsity constraint for the channel selection, $R_T(\mathcal{W})$ is a temporal regularization term, and $\lambda_i$ is a balancing parameter.

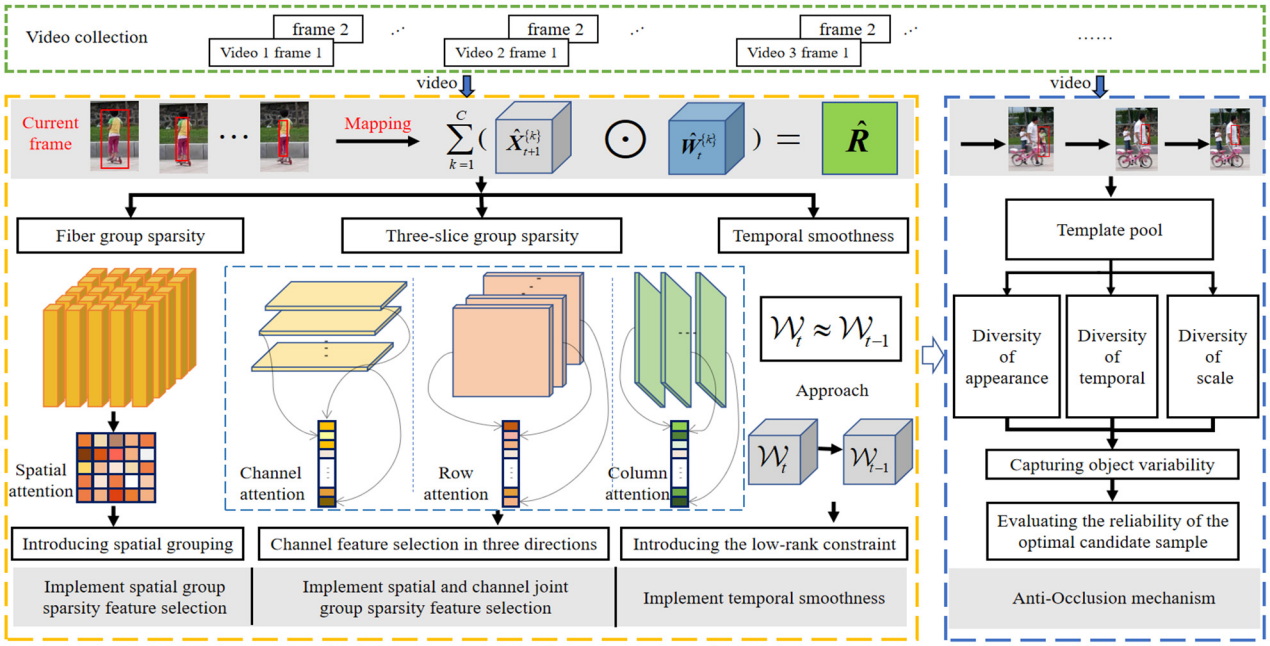



**Figure 1.** General structure diagram of the proposed method.

### *3.2. Fiber group sparsity constraint*

The purpose of introducing feature grouping in the model is to perform spatial and channel regularization of group features, thereby considering that most of the current features are represented in multi-channel form, and the tracked object is spatially coherent. In this paper, group sparsity features are adopted for $R_{Fiber}(\mathcal{W})$ and $R_{Slice}(\mathcal{W})$ by assigning different variables to specific groups. The group sparsity constraints are applied in both the spatial and channel dimensions, thus realizing the spatial and channel selection. This strategy has been shown to be effective in visual data science. To achieve fiber group sparsity selection in the spatial domain, the spatial fiber group sparsity constraint is defined as follows:

$$R_{Fiber}(\mathcal{W}) = \sum_{i=1}^{N} \sum_{j=1}^{N} \left\| w_{ij} \right\|_2. \quad (5)$$

$\mathcal{W} \in \mathbb{R}^{N\times N\times C}$ is expressed as the vector $\boldsymbol{w}_{ij} = [\boldsymbol{W}^{\{1\}}(i,j), \boldsymbol{W}^{\{2\}}(i,j),\text{L}\ ,\boldsymbol{W}^{\{C\}}(i,j)]^{\text{T}}$. The spatial location grouping attributes are obtained through the $\ell_2$ paradigm. Then, the correlation filter is optimized for regularization using the implicit $\ell_1$ paradigm for all spatial grouping properties. Sparsity is injected into the spatial domain by grouping across channels. This structured spatial sparsity facilitates robust group feature selection, thus reflecting the common action of features in the spatial domain.

### *3.3. Slice group sparsity constraint*

In this paper, to solve the above problems, such as the redundancy of channel information, we realize the slice group sparsity on the channel. The channel slice group sparsity constraint is defined as follows:

$$R_{Slice}(\mathcal{W}) = \sum_{i=1}^{H}\left\|W^{(i)}\right\|_2 + \sum_{j=1}^{W}\left\|W^{[j]}\right\|_2 + \sum_{k=1}^{C}\left\|W^{\{k\}}\right\|_2, \tag{6}$$

where $\left\{W^{(i)}\right\}_{i=1}^{H}$ is the row group sparsity attention, $\left\{W^{[j]}\right\}_{j=1}^{W}$ is the column group sparsity attention, and $\left\{W^{\{k\}}\right\}_{k=1}^{C}$ is the channel attention. The grouping properties of the slice group sparsity constraint are obtained by the Frobenius norm. The implicit constraint of Eq (6) is a sparsity-inducing $\ell_1$ paradigm.

In this paper, the regularization term is used to adaptively introduce the temporal low-rank property to achieve temporal smoothness. The temporal constraint term is defined as follows:

$$R_T(\mathcal{W}) = \sum_{k=1}^{C}\left\|W_t^{\{k\}} - W_{t-1}^{\{k\}}\right\|_2^2. \tag{7}$$

### *3.4. Solver for the proposed model*

Since Eq (4) is a convex function, this paper uses the augmented Lagrange method proposed by Lin et al. to optimize Eq (4). By introducing the slack variable $\mathcal{W}' = \mathcal{W}$, the Lagrange function is obtained as follows:

$$\begin{aligned} L =& \frac{1}{2}\left\|\sum_{k=1}^{C}\boldsymbol{W}_t^{\{k\}} * \boldsymbol{X}_t^{\{k\}} - \boldsymbol{Y}\right\|_2^2 + \lambda_1\left(\sum_{i=1}^{H}\left\|\boldsymbol{W}'^{(i)}\right\|_2 + \sum_{j=1}^{W}\left\|\boldsymbol{W}'^{[j]}\right\|_2 + \sum_{k=1}^{C}\left\|\boldsymbol{W}'^{\{k\}}\right\|_2\right) + \lambda_2\sum_{i=1}^{N}\sum_{j=1}^{N}\left\|\boldsymbol{w}_{ij}'^{(t)}\right\|_2 \\ &+ \left\langle \hat{\boldsymbol{W}}_t^{\{k\}} - \hat{\boldsymbol{W}}_t'^{\{k\}}, \hat{\boldsymbol{\Gamma}}^{\{k\}}\right\rangle + \frac{\lambda_3}{2}\sum_{k=1}^{C}\left\|\hat{\boldsymbol{W}}_t^{\{k\}} - \hat{\boldsymbol{W}}_{t-1}^{\{k\}}\right\|_2^2 + \frac{\mu}{2}\sum_{k=1}^{C}\left\|\hat{\boldsymbol{W}}_t^{\{k\}} - \hat{\boldsymbol{W}}_t'^{\{k\}}\right\|_2^2. \end{aligned} \tag{8}$$

The above objective function is obtained by matching squares as follows:

$$
\begin{aligned}
L =& \frac{1}{2}\left\|\sum_{k=1}^{C}\boldsymbol{W}_t^{\{k\}} * \boldsymbol{X}_t^{\{k\}} - \boldsymbol{Y}\right\|_2^2 + \lambda_1\left(\sum_{i=1}^{H}\left\|\boldsymbol{W}^{'(i)}\right\|_2 + \sum_{j=1}^{W}\left\|\boldsymbol{W}^{'[j]}\right\|_2 + \sum_{k=1}^{C}\left\|\boldsymbol{W}^{'\{k\}}\right\|_2\right) \\
&+\lambda_2\sum_{i=1}^{N}\sum_{j=1}^{N}\left\|\boldsymbol{w}_{ij}^{'(t)}\right\|_2 + \frac{\mu}{2}\left\langle 2\left(\hat{\boldsymbol{W}}_t^{\{k\}} - \hat{\boldsymbol{W}}_t^{'\{k\}}\right), \frac{\hat{\boldsymbol{\Gamma}}^{\{k\}}}{\mu}\right\rangle + \frac{\lambda_3}{2}\sum_{k=1}^{C}\left\|\hat{\boldsymbol{W}}_t^{\{k\}} - \hat{\boldsymbol{W}}_{t-1}^{\{k\}}\right\|_2^2 \\
&+\frac{\mu}{2}\sum_{k=1}^{C}\left\|\hat{\boldsymbol{W}}_t^{\{k\}} - \hat{\boldsymbol{W}}_t^{'\{k\}}\right\|_2^2 + \frac{\mu}{2}\left\|\frac{\hat{\boldsymbol{\Gamma}}^{\{k\}}}{\mu}\right\|_2^2 - \frac{\mu}{2}\left\|\frac{\hat{\boldsymbol{\Gamma}}^{\{k\}}}{\mu}\right\|_2^2.
\end{aligned} \tag{9}
$$

Equation (9) can be simplified as follows:

$$
\begin{aligned}
L =& \frac{1}{2}\left\|\sum_{k=1}^{C}\boldsymbol{W}_t^{\{k\}} * \boldsymbol{X}_t^{\{k\}} - \boldsymbol{Y}\right\|_2^2 + \lambda_1\left(\sum_{i=1}^{H}\left\|\boldsymbol{W}^{'(i)}\right\|_2 + \sum_{j=1}^{W}\left\|\boldsymbol{W}^{'[j]}\right\|_2 + \sum_{k=1}^{C}\left\|\boldsymbol{W}^{'\{k\}}\right\|_2\right) \\
&+\lambda_2\sum_{i=1}^{N}\sum_{j=1}^{N}\left\|\boldsymbol{w}_{ij}^{'(t)}\right\|_2 + \frac{\lambda_3}{2}\sum_{k=1}^{C}\left\|\hat{\boldsymbol{W}}_t^{\{k\}} - \hat{\boldsymbol{W}}_{t-1}^{\{k\}}\right\|_2^2 + \frac{\mu}{2}\sum_{k=1}^{C}\left\|\hat{\boldsymbol{W}}_t^{\{k\}} - \hat{\boldsymbol{W}}_t^{'\{k\}} + \frac{\hat{\boldsymbol{\Gamma}}^{\{k\}}}{\mu}\right\|_2^2 - \frac{\mu}{2}\left\|\frac{\hat{\boldsymbol{\Gamma}}^{\{k\}}}{\mu}\right\|_2^2.
\end{aligned} \tag{10}
$$

According to Parseval's theorem, Eq (10) is partially transformed to the frequency domain to obtain Eq (11) as follows:

$$
\begin{aligned}
L =& \frac{1}{2N^2}\left\|\sum_{k=1}^{C}\hat{\boldsymbol{W}}_t^{\{k\}} \text{ e } \hat{\boldsymbol{X}}_t^{\{k\}} - \hat{\boldsymbol{Y}}\right\|_2^2 + \lambda_1\left(\sum_{i=1}^{H}\left\|\boldsymbol{W}^{'(i)}\right\|_2 + \sum_{j=1}^{W}\left\|\boldsymbol{W}^{'[j]}\right\|_2 + \sum_{k=1}^{C}\left\|\boldsymbol{W}^{'\{k\}}\right\|_2\right) \\
&+\lambda_2\sum_{i=1}^{N}\sum_{j=1}^{N}\left\|\boldsymbol{w}_{ij}^{'(t)}\right\|_2 + \frac{\lambda_3}{2}\sum_{k=1}^{C}\left\|\hat{\boldsymbol{W}}_t^{\{k\}} - \hat{\boldsymbol{W}}_{t-1}^{\{k\}}\right\|_2^2 + \frac{\mu}{2}\sum_{k=1}^{C}\left\|\hat{\boldsymbol{W}}_t^{\{k\}} - \hat{\boldsymbol{W}}_t^{'\{k\}} + \frac{\hat{\boldsymbol{\Gamma}}^{\{k\}}}{\mu}\right\|_2^2 - \frac{\mu}{2N^2}\left\|\frac{\hat{\boldsymbol{\Gamma}}^{\{k\}}}{\mu}\right\|_2^2.
\end{aligned} \tag{11}
$$

1) The solution of $W_t^{\{k\}}$

First, we extract data from $Y$ at spatial point $(i,j)$, then obtain vectors $\widehat{w}_{ij} = [\widehat{W}_t^{\{1\}}(i,j), \widehat{W}_t^{\{2\}}(i,j), \cdots, \widehat{W}_t^{\{C\}}(i,j)]^T$ and $\widehat{w}_{ij}^{(t-1)} = [\widehat{W}_{t-1}^{\{1\}}(i,j), \widehat{W}_{t-1}^{\{2\}}(i,j), \cdots, \widehat{W}_{t-1}^{\{C\}}(i,j)]^T$ by taking the filter along the channel direction at the spatial point $(i,j)$. Similarly, we can obtain $\hat{x}_{ij}$, $\hat{y}_{ij}$, $\widehat{w}'_{ij}$, and $\hat{\gamma}_{ij}$. For the variable $\widehat{w}_{ij}$, the objective function can be expressed as follows:

$$
\mathcal{L}_{\widehat{w}_{ij}} = \frac{1}{2N^2}\left\|\hat{x}_{ij}^T \times \widehat{w}_{ij} - \hat{y}_{ij}\right\|_2^2 + \frac{\lambda_3}{2}\left\|\widehat{w}_{ij} - \widehat{w}_{ij}^{(t-1)}\right\|_2^2 + \frac{\mu}{2}\left\|\widehat{w}_{ij} - \widehat{w}'_{ij} + \frac{\widehat{\gamma}_{ij}}{\mu}\right\|_2^2. \tag{12}
$$

This can be obtained by making $\frac{\partial \mathcal{L}_{\widehat{w}_{ij}}}{\partial \widehat{w}_{ij}} = 0$, that is,

$$
\left(\hat{x}_{ij}^T\right)^H \times \left(\hat{x}_{ij}^T \times \widehat{w}_{ij} - \hat{y}_{ij}\right) + \lambda_3 N^2\left(\widehat{w}_{ij} - \widehat{w}_{ij}^{(t-1)}\right) + \mu N^2\left(\widehat{w}_{ij} - \widehat{w}'_{ij} + \hat{\gamma}_{ij}\right) = 0. \tag{13}
$$

Then, we have the following:

$$\left(\hat{x}_{ij}^{*}\hat{x}_{ij}^{T}+(\lambda_3+\mu)N^2I\right)\widehat{w}_{ij}=\left(\hat{x}_{ij}^{*}\hat{y}_{ij}+\lambda_3N^2\widehat{w}_{ij}^{(t-1)}+\mu N^2\widehat{w}'_{ij}-N^2\hat{\gamma}_{ij}\right). \tag{14}$$

We use the Sherman-Morrison formula $(uv^T+A)^{-1}=A^{-1}-(1+v^TA^{-1}u)^{-1}(A^{-1}uv^TA^{-1})$, which can be derived as follows:

$$\left(\hat{x}_{ij}^{*}\hat{x}_{ij}^{T}+(\lambda_3+\mu)N^2I\right)^{-1}=\frac{I}{(\lambda_3+\mu)N^2}-\left(1+\frac{\hat{x}_{ij}^{T}}{(\lambda_3+\mu)N^2}\hat{x}_{ij}^{*}\right)^{-1}\left(\frac{\hat{x}_{ij}^{*}\hat{x}_{ij}^{T}}{(\lambda_3+\mu)^2N^4}\right). \tag{15}$$

This can be derived by bringing Eq (15) into Eq (14) as follows:

$$\begin{aligned}\hat{\boldsymbol{w}}_{ij}&=\left(\hat{\boldsymbol{x}}_{ij}^{*}\hat{\boldsymbol{x}}_{ij}^{\mathrm{T}}+\left(\lambda_3+\mu\right)N^2\boldsymbol{I}\right)^{-1}\left(\hat{\boldsymbol{x}}_{ij}^{*}\hat{y}_{ij}+\lambda_3N^2\hat{\boldsymbol{w}}_{ij}^{(t-1)}+\mu N^2\hat{\boldsymbol{w}}_{ij}^{'}-N^2\hat{\boldsymbol{\gamma}}_{ij}\right)\\&=\left[\frac{\boldsymbol{I}}{\left(\lambda_3+\mu\right)N^2}-\left(1+\frac{\hat{\boldsymbol{x}}_{ij}^{\mathrm{T}}}{\left(\lambda_3+\mu\right)N^2}\hat{\boldsymbol{x}}_{ij}^{*}\right)^{-1}\left(\frac{\hat{\boldsymbol{x}}_{ij}^{*}\hat{\boldsymbol{x}}_{ij}^{\mathrm{T}}}{\left(\lambda_3+\mu\right)^2N^4}\right)\right]\\&\quad\left(\hat{\boldsymbol{x}}_{ij}^{*}\hat{y}_{ij}+\lambda_3N^2\hat{\boldsymbol{w}}_{ij}^{(t-1)}+\mu N^2\hat{\boldsymbol{w}}_{ij}^{'}-N^2\hat{\boldsymbol{\gamma}}_{ij}\right)\\&=\frac{1}{\left(\lambda_3+\mu\right)N^2}\left[\boldsymbol{I}-\left(\frac{\hat{\boldsymbol{x}}_{ij}^{*}\hat{\boldsymbol{x}}_{ij}^{\mathrm{T}}}{\hat{\boldsymbol{x}}_{ij}^{T}\hat{\boldsymbol{x}}_{ij}^{*}+\left(\lambda_3+\mu\right)N^2}\right)\right]\left(\hat{\boldsymbol{x}}_{ij}^{*}\hat{y}_{ij}+\lambda_3N^2\hat{\boldsymbol{w}}_{ij}^{(t-1)}+\mu N^2\hat{\boldsymbol{w}}_{ij}^{'}-N^2\hat{\boldsymbol{\gamma}}_{ij}\right).\end{aligned} \tag{16}$$

2) The solution of $W_t^{'\{k\}}$

According to the previous deduction, the objective function of $W_t^{'\{k\}}$ can be obtained as follows:

$$\mathcal{L}_{\widehat{w}_{ij}}=\lambda_1\left(\sum_{i=1}^{H}\left\|\widehat{W}^{(i)}\right\|_2+\sum_{j=1}^{W}\left\|\widehat{W}^{[j]}\right\|_2+\sum_{k=1}^{C}\left\|\widehat{W}^{\{k\}}\right\|_2\right)+\lambda_2\sum_{i=1}^{N}\sum_{j=1}^{N}\left\|w_{ij}^{'(t)}\right\|_2+\frac{\mu}{2}\sum_{k=1}^{C}\left\|\widehat{w}_{ij}-\widehat{w}'_{ij}+\frac{\widehat{\gamma}_{ij}}{\mu}\right\|_2^2. \tag{17}$$

According to the $\ell_{21}$ shrinkage, it can be derived as follows:

$$w_{ij}^{'(t)}=max\left(0,1-\frac{\lambda_1}{\mu\left\|P^{(i)}\right\|_2}-\frac{\lambda_1}{\mu\left\|P^{[j]}\right\|_2}-\frac{\lambda_1}{\mu\left\|P^{\{k\}}\right\|_2}-\frac{\lambda_2}{\mu\left\|p_{ij}\right\|_2}\right)\left(w_{ij}^{\{k\}}+\frac{\gamma_{ij}^{\{k\}}}{\mu}\right), \tag{18}$$

where $p_{ij}^{\{k\}}=w_{ij}^{\{k\}}+\frac{\gamma_{ij}^{\{k\}}}{\mu}$, $P^{(i)}$ is the $i$th mode 1 slice of the composition tensor of $p_{ij}$, $P^{[j]}$ is the $j$th mode 2 slice of the composition tensor of $p_{ij}$, and $P^{\{k\}}$ denotes the $k$th mode 3 slice of the composition tensor by $p_{ij}^{\{k\}}$.

3) The solution of $\hat{\Gamma}^{\{k\}}$

Similarly, the objective function of $\hat{\Gamma}^{\{k\}}$ is as follows:

$$L_{\hat{\Gamma}^{\{k\}}}=\left\langle\hat{\boldsymbol{W}}_t^{\{k\}}-\hat{\boldsymbol{W}}_t^{'\{k\}},\ \hat{\boldsymbol{\Gamma}}^{\{k\}}\right\rangle. \tag{19}$$

According to the gradient ascent method, $\hat{\Gamma}^{\{k\}}$ is obtained as follows:

$$\hat{F}^{\{k\}} = \hat{F}^{\{k\}} + \mu\left(\widehat{W}_t^{\{k\}} - \widehat{W}'^{\{k\}}_t\right). \tag{20}$$

### *3.5. Anti-Occlusion strategy*

Traditional UAV tracking methods have significant technical limitations when facing scenarios such as partial occlusion and complete occlusion. The tracker is prone to misjudging the occlusion as an object, thus affecting the object's appearance model update. When the visual information relied on by traditional methods fails, the absence of a reasonable occlusion recovery mechanism will lead to a tracking failure. For this reason, this paper proposes to construct a historical template pool. It evaluates the reliability of the current best candidate samples through the historical optimal samples to avoid the problem of filter degradation in occlusion scenarios. The implementation scheme of the candidate sample reliability assessment mechanism is analyzed in detail below.

To construct the template pool, the patch of the first frame is taken as the history sample (i.e., $f_n = x^{(1)}(n = 1,2,\cdots,N)$), where $x^{(1)}$ denotes the first frame patch, and $f_n$ denotes the $n$th column vector of the template pool $M$. Starting from the second frame, assuming that the patch of the optimal sample is $b$, the HOG features of $M$ and $b$ are extracted, as shown in Eqs (21) and (22), respectively:

$$l_{f_n} = \frac{HOG(f_n)}{max(HOG(f_n))}, \tag{21}$$

$$l_b = \frac{HOG(b)}{max(HOG(b))}, \tag{22}$$

where $HOG$ denotes the directional gradient histogram extraction operator, $HOG(f_n)$ denotes the HOG features of the historical samples, and $HOG(b)$ denotes the HOG features of the best candidate samples.

If the similarity between the best candidate sample and all the samples in the template pool is lower than a preset threshold, the sample is determined to be an unreliable sample. This mechanism effectively avoids misleading the filter design due to complex scenarios such as occlusion or an object moving out of the field of view. When the similarity between the best candidate sample and the largest in the template pool is greater than the human-set threshold, the sample is judged to be a reliable sample. At this point, the filter and object appearance models can be updated. The best candidate sample is incorporated into the template pool, which is shown in Eq (23):

$$max\left(cos\left(l_{f_n}, b\right)\right) > \tau, \tag{23}$$

where $\tau$ is a threshold value that takes the range of $[0,1]$.

The anti-occlusion mechanism evaluates the best candidate samples for each frame by a threshold. This prevents the filter from being contaminated by the occlusion to update and optimize the performance of the filter. The mechanism can tap the diversity of the object and effectively address the tracking drift and failure caused by occlusion, violent motion, and other challenges. It promises to realize more intelligent and robust UAV object tracking.

## 4. Experiments

In this section, we conduct extensive and comprehensive experiments on three challenging UAV

benchmarks and OTB benchmarks. First, we present the detailed experimental setup, including the evaluation metrics, parameters, benchmark tests, and platform. Next, we compare the proposed tracker with other state-of-the-art trackers. Then, the effectiveness of the proposed tracker is verified by ablation experiments. Finally, we analyze the limitations of the proposed tracker.

### *4.1. Experimental setup*

This section provides information on the evaluation metrics, parameters, UAV benchmarks, and platform for the experiment.

1) Evaluation metrics

In the experiment, we adopted the one pass evaluation (OPE) strategy. We combine five main metrics to evaluate the performance of different algorithms: center point error (CPE), overlap rate (OR), distance precision (DP), area under the curve (AUC), and frames per second (FPS). The CPE is the Euclidean distance in pixels between the predicted object center and the real object center. The OR is the degree of overlap between the predicted bounding box and the real object bounding box, measured by the intersection over union (IoU) threshold. We use the DP/AUC as a combined evaluation metric for the precision plot /success rate plot. The DP denotes the precision score; it is the percentage of video frames where the distance between the center of the predicted bounding box and the center of the real bounding box is within a threshold of 20 pixels. The AUC is the area under the success curve. The FPS is a metric used to measure the speed of the object tracking algorithm and represents the number of frames processed per second.

The CPE is defined as follows:

$$d_c = \sqrt{(x - x_c)^2 + (y - y_c)^2}, \tag{24}$$

where $(x, y)$ denotes the predicted object center position, and $(x_c, y_c)$ denotes the real object center position.

The precision plot is a percentage curve of video frames with a center position error less than a given threshold $p$:

$$P = \frac{\#(d_c < p)}{N}, \tag{25}$$

where $N$ is the number of frames of the video, and $\#(d_c < p)$ denotes the number of frames where the center point error is less than the threshold $p$. The precision plot of the video is shown on the horizontal axis. The horizontal axis of the precision plot is the threshold from 0 to 50 pixels and the vertical axis represents the precision. The DP denotes the precision with a threshold $p$ of 20 pixels.

The OR is defined as follows:

$$OR = \frac{|D \cap D_g|}{|D \cup D_g|}, \tag{26}$$

where $D$ denotes the predicted object bounding box, and $D_g$ denotes the real object bounding box. The OR uses $IoU \geq 0.5$ as a threshold to determine whether the prediction is correct or not.

The success rate plot is a percentage curve of video frames whose overlap precision is greater than a given threshold $s$:

$$SR = \frac{\#(OR>s)}{N}, \tag{27}$$

where $\#(OR > s)$ denotes the number of frames where the overlap rate is greater than the threshold $s$. The horizontal axis of the success rate plot is the threshold IoU (from 0 to 1) and the vertical axis is the success rate $SR$.

The FPS is defined as follows:

$$FPS = \frac{N}{T}, \tag{28}$$

where $T$ denotes the total time spent processing these frames.

2) Parameters

We set the balance parameters as $\lambda_1 = 10$ and $\lambda_2 = 1$ to realize the allocation of the spatial and channel attention. For hand-crafted features, we set the parameters to $\lambda_3 = 16$, and $\alpha = 0.9$. For deep features, we set the parameters as $\lambda_3 = 12$, and $\alpha = 0.05$.

3) UAV benchmarks

With the rapid development of UAV technology, UAV is increasingly used in aerial photography, monitoring, logistics, and distribution. However, effective object tracking of UAV-captured videos still faces many challenges. To verify the effectiveness of the algorithms in this paper, experiments are conducted on three well-known UAV benchmarks: UAV123, UAV20L, and DTB70.

The UAV123 dataset is an object-tracking dataset composed of video sequences captured by UAVs. It contains 123 high-definition video sequences and rich annotation information. This dataset is known for its clean background and diverse perspective changes. The UAV123 dataset covers diverse environmental conditions and object types, which can meet the needs of different application scenarios. It is widely used in object-tracking algorithm research, UAV application development, and other fields.

The UAV20L dataset is a long-time tracking subset of the UAV123 dataset. It contains 20 long video sequences with significant time spans. Evaluating the object tracking algorithm on the UAV20L dataset can comprehensively verify the performance of the algorithm in long-term tracking, complex scenes, multi-object tracking, and UAV perspectives.

The DTB70 dataset is another influential dataset in the field of UAV object tracking. It contains 70 high-quality short video sequences and covers a variety of challenging scenarios, such as fast movement, occlusion, and appearance change, which can comprehensively evaluate the performance of the algorithm.

These datasets provide strong support for research in the field of UAV video tracking and promote the development of related technologies.

4) Platform

The proposed SCACF is implemented on the MATLAB 2018b platform. The computer has an Intel Xeon CPU E3-1230 v6 (3.50 GHz) processor and an NVIDIA Quadro P2000 graphics card.

### *4.2. Quantitative analysis*

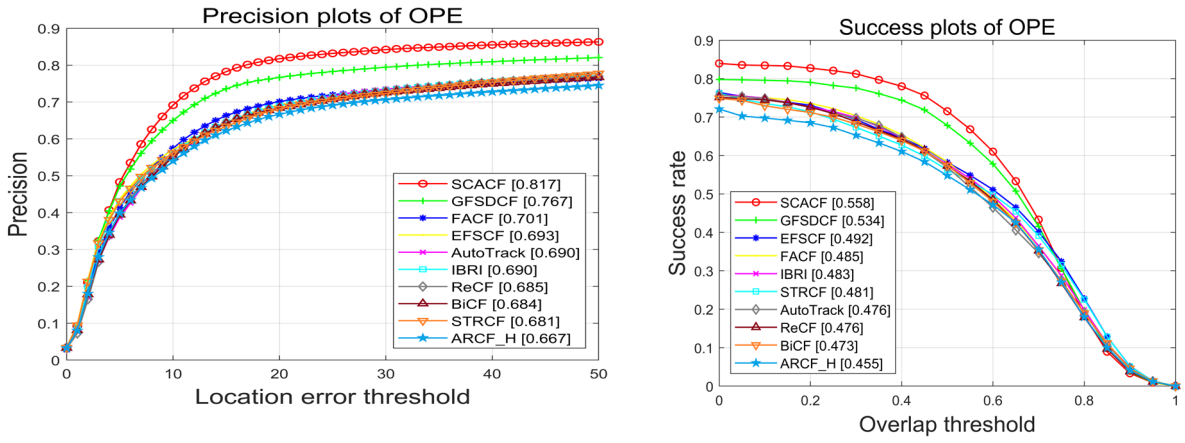
To validate the performance of the proposed tracker in UAV-specific scenarios, comprehensive experimental evaluations are conducted on three UAV tracking benchmarks, UAV123, UAV20L, and DTB70, as well as on the OTB dataset. For the UAV tracking benchmarks, SCACF is compared with

other state-of-the-art tracking algorithms, including GFSDCF [28], EFSCF [29], STRCF [17], ARCF_H [22], IBRI [61], BiCF [62], ReCF [63], AutoTrack [64], and FACF [65]. For the OTB dataset, SCACF is compared with GFSDCF [28], EFSCF [29], STRCF [17], ARCF_H [22], IBRI [61], BiCF [62], CSR-DCF [57], Staple [38], and BACF [66].

#### 4.2.1. UAV123

Comprehensive evaluation: Due to the characteristics of UAV photography, the object in the video sequences of the UAV123 dataset often faces challenges such as fast motion, occlusion, and illumination variations. It is suitable to evaluate and develop advanced object-tracking algorithms. Figure 2 demonstrates the precision and success plots of the proposed SCACF with nine other state-of-the-art object tracking algorithms on the UAV123 dataset. As shown in Figure 2(a), the SCACF ranks first with 81.7% precision, which is 5% better than the second-ranked GFSDCF. FACF ranks third with 70.1% precision. As shown in Figure 2(b), the SCACF is ranked first with a 55.8% success rate, the GFSDCF has a 53.4% success rate, and the EFSCF ranks third with a 49.2% success rate. The comprehensive evaluation results show that the SCACF outperforms the other algorithms in UAV vision tasks across the board. It excels in both precision and success rate.

Attribute-based evaluation: The UAV123 dataset covers a wide range of environments and weather conditions, such as cities, villages, coastlines, forests, and ski resorts. Different weather conditions include sunny, cloudy, rainy, snowy, nighttime, etc. These environmental factors can affect the quality of images and feature extraction, thus making object detection and tracking tasks more difficult. We use 12 challenging attributes on the UAV123 dataset to evaluate the accuracy and robustness of the SCACF in various challenging scenarios: scale variation (SV), aspect ratio change (ARC), low resolution (LR), fast motion (FM), full occlusion (FO), partial occlusion (PO), out-of-view (OV), background clutter (BC), illumination variation (IV), viewpoint change (VC), camera motion (CM), and similar object (SO).




(a) Precision plot (b) Success plot

**Figure 2.** Precision and success plots of our proposed tracker and nine other state-of-the-art trackers on UAV123.

Tables 1 and 2 demonstrate the precision and success rate of SCACF with the other 9 state-of-the-art tracking algorithms under each challenging attribute, respectively. The top three algorithms in each attribute are displayed in red, green, and blue, respectively. The experimental results show that the SCACF demonstrates strong adaptability and stability when dealing with complex tasks, thus fully reflecting its excellent performance. Notably, the SCACF maintains high precision and success rates in challenging scenarios such as light change, occlusion, and background clutter.

**Table 1.** Precision of 12 challenging attributes on UAV123.

| Attribute | SCACF | GFSDCF | FACF | AutoTrack | ARCF_H | ReCF | IBRI | BiCF | EFSCF | STRCF |
|---|---|---|---|---|---|---|---|---|---|---|
| SV | 0.794 | 0.738 | 0.663 | 0.651 | 0.626 | 0.645 | 0.651 | 0.643 | 0.655 | 0.643 |
| ARC | 0.788 | 0.719 | 0.626 | 0.628 | 0.586 | 0.621 | 0.616 | 0.600 | 0.605 | 0.586 |
| LR | 0.717 | 0.635 | 0.629 | 0.595 | 0.534 | 0.602 | 0.611 | 0.602 | 0.616 | 0.589 |
| FM | 0.761 | 0.732 | 0.582 | 0.542 | 0.552 | 0.573 | 0.567 | 0.569 | 0.561 | 0.554 |
| FO | 0.649 | 0.605 | 0.477 | 0.464 | 0.431 | 0.485 | 0.489 | 0.504 | 0.487 | 0.488 |
| PO | 0.752 | 0.691 | 0.605 | 0.586 | 0.571 | 0.591 | 0.605 | 0.583 | 0.617 | 0.587 |
| OV | 0.715 | 0.681 | 0.584 | 0.562 | 0.542 | 0.563 | 0.569 | 0.580 | 0.546 | 0.570 |
| BC | 0.739 | 0.667 | 0.563 | 0.579 | 0.526 | 0.562 | 0.578 | 0.556 | 0.532 | 0.502 |
| IV | 0.843 | 0.765 | 0.630 | 0.629 | 0.586 | 0.646 | 0.609 | 0.618 | 0.586 | 0.538 |
| VC | 0.802 | 0.732 | 0.622 | 0.624 | 0.596 | 0.630 | 0.610 | 0.628 | 0.596 | 0.581 |
| CM | 0.800 | 0.762 | 0.684 | 0.660 | 0.646 | 0.653 | 0.674 | 0.651 | 0.668 | 0.658 |
| SO | 0.786 | 0.723 | 0.698 | 0.661 | 0.684 | 0.698 | 0.703 | 0.634 | 0.717 | 0.648 |

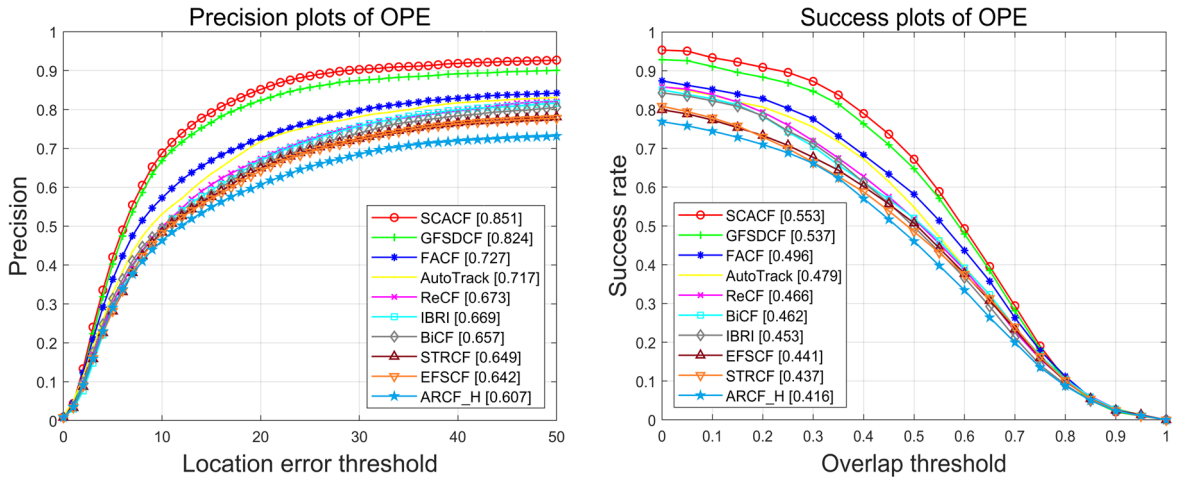
**Table 2.** Success rate of 12 challenging attributes on UAV123.

| Attribute | SCACF | GFSDCF | FACF | AutoTrack | ARCF_H | ReCF | IBRI | BiCF | EFSCF | STRCF |
|---|---|---|---|---|---|---|---|---|---|---|
| SV | 0.533 | 0.505 | 0.454 | 0.444 | 0.422 | 0.443 | 0.451 | 0.441 | 0.460 | 0.448 |
| ARC | 0.508 | 0.475 | 0.410 | 0.415 | 0.384 | 0.409 | 0.413 | 0.397 | 0.413 | 0.398 |
| LR | 0.428 | 0.374 | 0.362 | 0.338 | 0.285 | 0.340 | 0.356 | 0.332 | 0.352 | 0.337 |
| FM | 0.478 | 0.460 | 0.376 | 0.363 | 0.353 | 0.370 | 0.371 | 0.364 | 0.354 | 0.347 |
| FO | 0.353 | 0.327 | 0.243 | 0.233 | 0.212 | 0.246 | 0.261 | 0.254 | 0.251 | 0.258 |
| PO | 0.498 | 0.464 | 0.400 | 0.392 | 0.369 | 0.395 | 0.412 | 0.390 | 0.424 | 0.402 |
| OV | 0.470 | 0.462 | 0.407 | 0.401 | 0.380 | 0.400 | 0.411 | 0.415 | 0.397 | 0.410 |
| BC | 0.472 | 0.433 | 0.334 | 0.345 | 0.315 | 0.345 | 0.350 | 0.341 | 0.335 | 0.318 |
| IV | 0.558 | 0.513 | 0.397 | 0.404 | 0.389 | 0.472 | 0.415 | 0.408 | 0.390 | 0.354 |
| VC | 0.536 | 0.506 | 0.419 | 0.424 | 0.402 | 0.434 | 0.421 | 0.429 | 0.421 | 0.406 |
| CM | 0.552 | 0.536 | 0.478 | 0.465 | 0.450 | 0.460 | 0.475 | 0.457 | 0.474 | 0.470 |
| SO | 0.530 | 0.490 | 0.468 | 0.447 | 0.448 | 0.467 | 0.477 | 0.433 | 0.491 | 0.451 |

#### 4.2.2. DTB70

Comprehensive evaluation: The performance of this paper's tracking algorithm is evaluated on the DTB dataset for UAV-specific complex scenarios. Figure 3 demonstrates the precision and success plots of the SCACF against nine other state-of-the-art object tracking algorithms on the DTB dataset. As shown in Figure 3, the proposed method has the highest precision and success rate of 85.1%

and 55.3%, respectively. Compared to the second-ranked GFSDCF, the SCACF improves 2.7% and 1.6% in terms of precision and success rate, respectively. It improves 12.4% and 5.7% over the third-ranked FACF. The experimental results on the DTB dataset show that the SCACF has good overall tracking performance and is suitable for UAV tracking.




(a) Precision plot (b) Success plot

**Figure 3.** Precision and success plots of our proposed tracker and nine other state-of-the-art trackers on DTB70.

Attribute-based evaluation: We used 11 attributes on the DTB70 dataset to evaluate the reliability of SCACF in complex scenes: SV, aspect ratio variation (ARV), occlusion (OCC), deformation (DEF), fast camera motion (FCM), in-plane rotation (IPR), out-of-plane rotation (OPR), OV, BC, small object appearance (SOA), and motion blur (MB). Table 3 shows the precision of the SCACF with the other 9 state-of-the-art tracking algorithms under each challenging attribute. Table 4 demonstrates the success rate of the SCACF with the other 9 state-of-the-art tracking algorithms under each challenging attribute. The experimental results show that the SCACF performs well in complex scenarios and significantly outperforms the other 9 state-of-the-art tracking algorithms. The algorithm ranks first in precision and success rate for all 11 challenging attributes on the DTB dataset.

**Table 3.** Precision of 11 challenging attributes on DTB70.

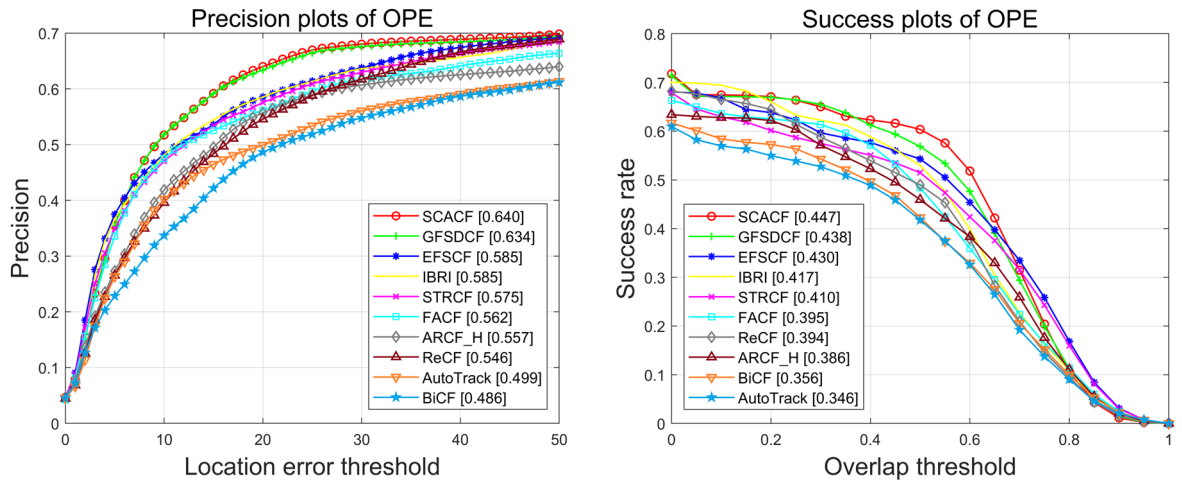
| Attribute | SCACF | GFSDCF | FACF | AutoTrack | ARCF_H | ReCF | IBRI | BiCF | EFSCF | STRCF |
|---|---|---|---|---|---|---|---|---|---|---|
| SV | 0.820 | 0.815 | 0.676 | 0.688 | 0.560 | 0.626 | 0.609 | 0.590 | 0.570 | 0.568 |
| ARV | 0.749 | 0.731 | 0.617 | 0.609 | 0.431 | 0.557 | 0.539 | 0.541 | 0.492 | 0.492 |
| OCC | 0.762 | 0.719 | 0.635 | 0.631 | 0.546 | 0.633 | 0.528 | 0.558 | 0.617 | 0.617 |
| DEF | 0.816 | 0.811 | 0.689 | 0.670 | 0.427 | 0.534 | 0.585 | 0.597 | 0.555 | 0.554 |
| FCM | 0.894 | 0.867 | 0.773 | 0.746 | 0.654 | 0.720 | 0.713 | 0.685 | 0.693 | 0.713 |
| IPR | 0.813 | 0.800 | 0.676 | 0.685 | 0.547 | 0.620 | 0.639 | 0.615 | 0.587 | 0.586 |
| OPR | 0.552 | 0.535 | 0.511 | 0.439 | 0.262 | 0.384 | 0.402 | 0.372 | 0.368 | 0.385 |
| OV | 0.853 | 0.841 | 0.705 | 0.690 | 0.671 | 0.571 | 0.660 | 0.587 | 0.652 | 0.652 |
| BC | 0.888 | 0.849 | 0.681 | 0.635 | 0.555 | 0.577 | 0.643 | 0.579 | 0.623 | 0.611 |
| SOA | 0.859 | 0.804 | 0.753 | 0.731 | 0.679 | 0.743 | 0.697 | 0.681 | 0.663 | 0.677 |
| MB | 0.886 | 0.847 | 0.781 | 0.703 | 0.590 | 0.655 | 0.694 | 0.633 | 0.684 | 0.689 |

**Table 4.** Success rate of 11 challenging attributes on DTB70.

| Attribute | SCACF | GFSDCF | FACF | AutoTrack | ARCF_H | ReCF | IBRI | BiCF | EFSCF | STRCF |
|---|---|---|---|---|---|---|---|---|---|---|
| SV | 0.536 | 0.534 | 0.478 | 0.493 | 0.406 | 0.496 | 0.450 | 0.482 | 0.422 | 0.417 |
| ARV | 0.479 | 0.472 | 0.408 | 0.406 | 0.314 | 0.403 | 0.360 | 0.398 | 0.353 | 0.347 |
| OCC | 0.506 | 0.476 | 0.440 | 0.415 | 0.354 | 0.403 | 0.365 | 0.372 | 0.410 | 0.400 |
| DEF | 0.534 | 0.529 | 0.459 | 0.452 | 0.308 | 0.406 | 0.394 | 0.444 | 0.389 | 0.390 |
| FCM | 0.579 | 0.562 | 0.529 | 0.497 | 0.444 | 0.481 | 0.480 | 0.472 | 0.462 | 0.467 |
| IPR | 0.518 | 0.513 | 0.459 | 0.454 | 0.383 | 0.433 | 0.433 | 0.439 | 0.404 | 0.393 |
| OPR | 0.376 | 0.372 | 0.366 | 0.343 | 0.228 | 0.327 | 0.280 | 0.354 | 0.254 | 0.257 |
| OV | 0.532 | 0.511 | 0.463 | 0.407 | 0.424 | 0.364 | 0.457 | 0.389 | 0.431 | 0.424 |
| BC | 0.541 | 0.519 | 0.439 | 0.394 | 0.354 | 0.378 | 0.408 | 0.381 | 0.380 | 0.369 |
| SOA | 0.569 | 0.538 | 0.502 | 0.473 | 0.434 | 0.474 | 0.446 | 0.444 | 0.443 | 0.447 |
| MB | 0.569 | 0.551 | 0.537 | 0.468 | 0.395 | 0.441 | 0.447 | 0.448 | 0.447 | 0.447 |

#### 4.2.3. UAV20L

Comprehensive evaluation: The UAV20L dataset contains long-time video sequences, which can effectively evaluate the robustness of this paper's algorithm in long-time tracking. Figure 4 demonstrates the precision and success plots of this paper's algorithm with nine other state-of-the-art object tracking algorithms on the UAV20L dataset. As shown in Figure 4, in terms of the precision and success plots, the SCACF achieves the best performance of 0.640 and 0.447, respectively. The experimental results show that the SCACF satisfies the long-term tracking requirements in practical applications.




(a) Precision plot (b) Success plot

**Figure 4.** Precision and success plots of our proposed tracker and nine other state-of-the-art trackers on UAV20L.

Attribute-based evaluation: The UAV20L dataset has the same challenging attributes as the UAV123 dataset. Table 5 lists the precision of the 10 different tracking algorithms. Table 6 lists the success rate of 10 different tracking algorithms. The experimental results on the UAV20L dataset validate the effectiveness of the SCACF in long-duration tracking tasks and further highlight its wide applicability in practical applications.

**Table 5.** Precision of 12 challenging attributes on UAV20L.

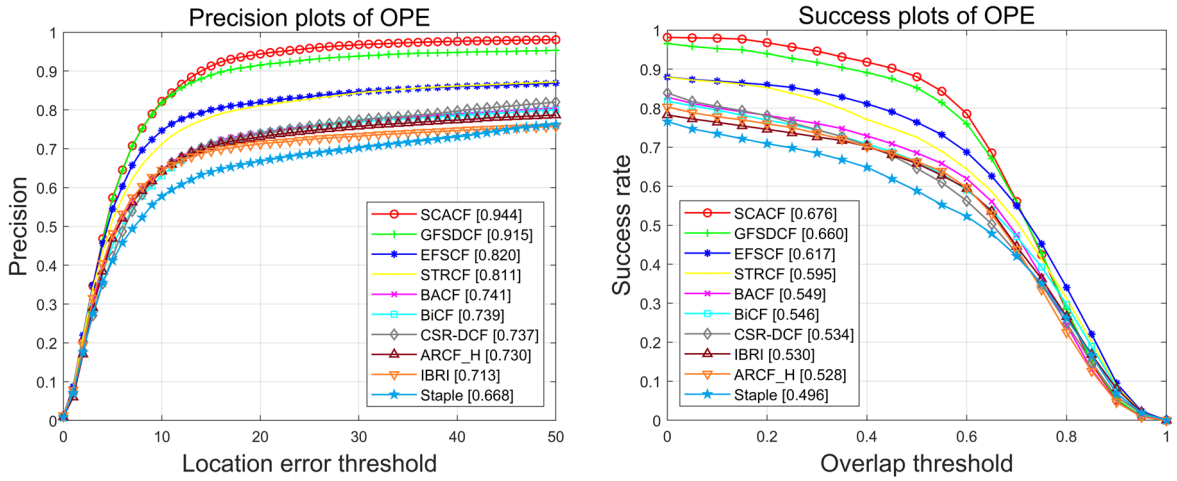
| Attribute | SCACF | GFSDCF | FACF | AutoTrack | ARCF_H | ReCF | IBRI | BiCF | EFSCF | STRCF |
|---|---|---|---|---|---|---|---|---|---|---|
| SV | 0.621 | 0.614 | 0.539 | 0.473 | 0.534 | 0.530 | 0.563 | 0.488 | 0.564 | 0.553 |
| ARC | 0.552 | 0.545 | 0.455 | 0.407 | 0.454 | 0.466 | 0.490 | 0.395 | 0.485 | 0.472 |
| LR | 0.532 | 0.528 | 0.488 | 0.414 | 0.440 | 0.475 | 0.537 | 0.385 | 0.535 | 0.513 |
| FM | 0.550 | 0.553 | 0.433 | 0.382 | 0.354 | 0.520 | 0.518 | 0.430 | 0.488 | 0.506 |
| FO | 0.430 | 0.423 | 0.403 | 0.401 | 0.378 | 0.388 | 0.406 | 0.385 | 0.382 | 0.406 |
| PO | 0.606 | 0.599 | 0.545 | 0.476 | 0.543 | 0.515 | 0.567 | 0.452 | 0.576 | 0.564 |
| OV | 0.611 | 0.606 | 0.537 | 0.478 | 0.538 | 0.493 | 0.522 | 0.482 | 0.532 | 0.525 |
| BC | 0.416 | 0.404 | 0.375 | 0.371 | 0.329 | 0.345 | 0.376 | 0.345 | 0.330 | 0.330 |
| IV | 0.638 | 0.621 | 0.508 | 0.452 | 0.488 | 0.452 | 0.513 | 0.413 | 0.465 | 0.429 |
| VC | 0.578 | 0.574 | 0.465 | 0.407 | 0.465 | 0.435 | 0.466 | 0.406 | 0.457 | 0.441 |
| CM | 0.621 | 0.614 | 0.539 | 0.473 | 0.534 | 0.522 | 0.563 | 0.459 | 0.564 | 0.553 |
| SO | 0.625 | 0.613 | 0.523 | 0.447 | 0.558 | 0.524 | 0.566 | 0.460 | 0.572 | 0.547 |

**Table 6.** Success rate of 12 challenging attributes on UAV20L.

| Attribute | SCACF | GFSDCF | FACF | AutoTrack | ARCF_H | ReCF | IBRI | BiCF | EFSCF | STRCF |
|---|---|---|---|---|---|---|---|---|---|---|
| SV | 0.433 | 0.423 | 0.380 | 0.327 | 0.368 | 0.388 | 0.403 | 0.357 | 0.415 | 0.393 |
| ARC | 0.382 | 0.376 | 0.327 | 0.280 | 0.319 | 0.335 | 0.352 | 0.292 | 0.355 | 0.331 |
| LR | 0.312 | 0.302 | 0.292 | 0.238 | 0.240 | 0.278 | 0.316 | 0.223 | 0.321 | 0.293 |
| FM | 0.316 | 0.300 | 0.263 | 0.231 | 0.201 | 0.316 | 0.307 | 0.241 | 0.270 | 0.243 |
| FO | 0.232 | 0.232 | 0.216 | 0.199 | 0.199 | 0.201 | 0.220 | 0.199 | 0.211 | 0.217 |
| PO | 0.421 | 0.410 | 0.372 | 0.317 | 0.373 | 0.366 | 0.399 | 0.329 | 0.420 | 0.401 |
| OV | 0.427 | 0.411 | 0.364 | 0.319 | 0.377 | 0.360 | 0.371 | 0.352 | 0.392 | 0.375 |
| BC | 0.268 | 0.268 | 0.245 | 0.222 | 0.210 | 0.215 | 0.246 | 0.210 | 0.226 | 0.227 |
| IV | 0.464 | 0.465 | 0.380 | 0.327 | 0.385 | 0.339 | 0.393 | 0.325 | 0.373 | 0.344 |
| VC | 0.426 | 0.416 | 0.354 | 0.303 | 0.334 | 0.340 | 0.353 | 0.323 | 0.363 | 0.332 |
| CM | 0.433 | 0.423 | 0.378 | 0.328 | 0.378 | 0.376 | 0.402 | 0.338 | 0.412 | 0.392 |
| SO | 0.464 | 0.460 | 0.405 | 0.351 | 0.441 | 0.428 | 0.447 | 0.390 | 0.461 | 0.439 |

### 4.2.4. OTB50

The OTB50 dataset contains 50 video sequences. Each video sequence is labeled with 11 challenge attributes: FM, BC, MB, SV, DEF, IV, IPR, LR, OCC, OPR, and OV). OTB50 is a landmark dataset in the field of object tracking and provides important support for the evaluation and comparison of algorithms.




(a) Precision plot (b) Success plot

**Figure 5.** Precision and success plots of our proposed tracker and nine other state-of-the-art trackers on OTB50.

Comprehensive evaluation: Figure 5 demonstrates the precision and success rate plots of this paper's algorithm against nine other state-of-the-art target tracking algorithms on the OTB50 dataset. From the plots, the SCACF outperforms the other algorithms with 94.4% precision and a 67.6% success rate. The DP/AUC scores are 2.9%/1.6% higher than the baseline algorithm GFCDCF, respectively.

Attribute-based evaluation: To further evaluate the performance of the SCACF on the OTB50 dataset, Table 7 shows the precision performance of each algorithm under different challenge attributes. Table 8 shows the success rate performance of each algorithm under different challenge attributes. The SCACF outperforms the other algorithms in terms of precision and success rates for each attribute. Among them, under the OPR attribute, the SCACF achieves a precision and success rate of 0.928 and 0.672, respectively. Under the OCC attribute, the SCACF achieves a precision and success rate of 0.918 and 0.667, respectively.

**Table 7.** Precision of 11 challenging attributes on OTB50.

| Attribute | SCACF | GFSDCF | EFSCF | CSR-DCF | ARCF_H | Staple | IBRI | BiCF | BACF | STRCF |
|---|---|---|---|---|---|---|---|---|---|---|
| IV | 0.969 | 0.955 | 0.771 | 0.722 | 0.703 | 0.642 | 0.705 | 0.721 | 0.685 | 0.755 |
| OPR | 0.928 | 0.883 | 0.799 | 0.671 | 0.666 | 0.605 | 0.654 | 0.675 | 0.692 | 0.776 |
| SV | 0.933 | 0.897 | 0.785 | 0.684 | 0.667 | 0.605 | 0.673 | 0.681 | 0.687 | 0.777 |
| OCC | 0.918 | 0.866 | 0.802 | 0.647 | 0.656 | 0.643 | 0.644 | 0.666 | 0.687 | 0.777 |
| DEF | 0.904 | 0.848 | 0.768 | 0.732 | 0.679 | 0.643 | 0.657 | 0.649 | 0.669 | 0.755 |
| MB | 0.934 | 0.910 | 0.751 | 0.714 | 0.658 | 0.604 | 0.654 | 0.706 | 0.652 | 0.727 |
| FM | 0.937 | 0.924 | 0.742 | 0.734 | 0.707 | 0.640 | 0.683 | 0.712 | 0.732 | 0.761 |
| IPR | 0.941 | 0.925 | 0.778 | 0.702 | 0.665 | 0.608 | 0.648 | 0.701 | 0.705 | 0.743 |
| OV | 0.953 | 0.925 | 0.763 | 0.686 | 0.666 | 0.669 | 0.647 | 0.714 | 0.711 | 0.720 |
| BC | 0.973 | 0.913 | 0.832 | 0.704 | 0.722 | 0.624 | 0.694 | 0.747 | 0.702 | 0.841 |
| LR | 0.973 | 0.950 | 0.751 | 0.677 | 0.692 | 0.610 | 0.711 | 0.774 | 0.741 | 0.737 |

**Table 8.** Success rate of 11 challenging attributes on OTB50.

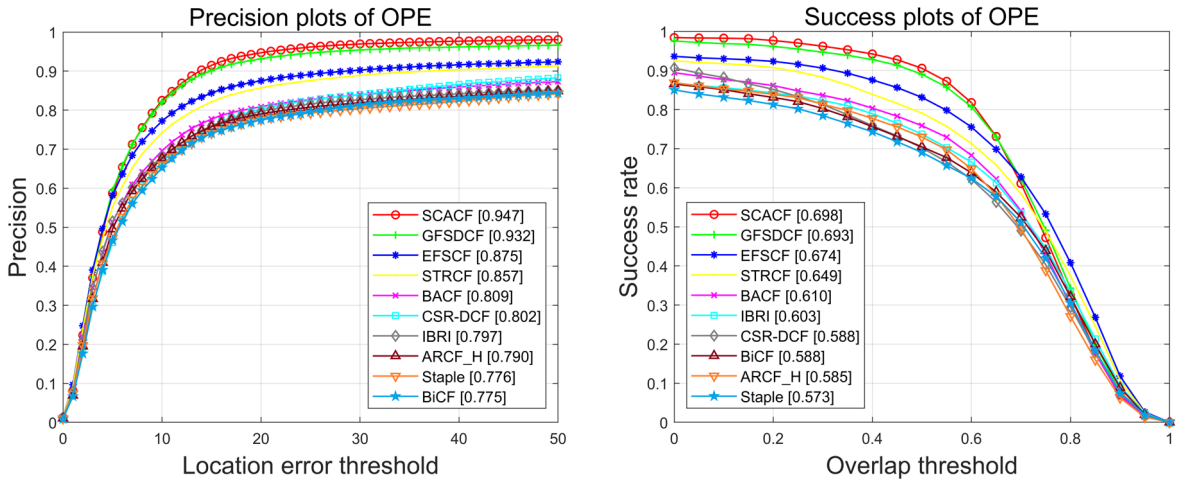
| Attribute | SCACF | GFSDCF | EFSCF | CSR-DCF | ARCF_H | Staple | IBRI | BiCF | BACF | STRCF |
|---|---|---|---|---|---|---|---|---|---|---|
| IV | 0.703 | 0.701 | 0.597 | 0.540 | 0.522 | 0.489 | 0.527 | 0.545 | 0.517 | 0.559 |
| OPR | 0.672 | 0.644 | 0.598 | 0.482 | 0.480 | 0.445 | 0.485 | 0.494 | 0.514 | 0.562 |
| SV | 0.677 | 0.652 | 0.591 | 0.499 | 0.476 | 0.445 | 0.500 | 0.510 | 0.511 | 0.574 |
| OCC | 0.667 | 0.634 | 0.597 | 0.474 | 0.470 | 0.469 | 0.475 | 0.475 | 0.505 | 0.558 |
| DEF | 0.635 | 0.609 | 0.562 | 0.524 | 0.494 | 0.486 | 0.482 | 0.471 | 0.503 | 0.516 |
| MB | 0.665 | 0.659 | 0.583 | 0.546 | 0.502 | 0.447 | 0.493 | 0.533 | 0.496 | 0.560 |
| FM | 0.678 | 0.674 | 0.584 | 0.552 | 0.537 | 0.483 | 0.521 | 0.549 | 0.556 | 0.586 |
| IPR | 0.655 | 0.647 | 0.583 | 0.505 | 0.482 | 0.443 | 0.485 | 0.513 | 0.524 | 0.557 |
| OV | 0.676 | 0.662 | 0.569 | 0.489 | 0.466 | 0.455 | 0.484 | 0.507 | 0.502 | 0.533 |
| BC | 0.701 | 0.669 | 0.627 | 0.533 | 0.528 | 0.476 | 0.531 | 0.557 | 0.522 | 0.608 |
| LR | 0.691 | 0.650 | 0.559 | 0.417 | 0.460 | 0.399 | 0.514 | 0.550 | 0.532 | 0.538 |

#### 4.2.5. OTB2015

The OTB2015 dataset is an extended version of the OTB50 dataset. It contains 100 video sequences that cover a wide range of challenging scenarios and complex situations. The dataset provides rich data and evaluation metrics and has become a standard dataset widely used in object-

tracking research.

Comprehensive evaluation: We compared the proposed method with nine other state-of-the-art object tracking algorithms on OTB2015. The precision and success plots are shown in Figure 6. In terms of the DP and AUC, our SCACF tracker performs the best (0.947 and 0.698), followed by the GFSDCF tracker (0.932 and 0.693) and the EFSCF tracker (0.875 and 0.674).




(a) Precision plot (b) Success plot

**Figure 6.** Precision and success plots of our proposed tracker and nine other state-of-the-art trackers on OTB2015.

Attribute-based evaluation: The OTB2015 dataset has the same challenge attributes as the OTB50 dataset. Table 9 shows the precision performance of each algorithm under different challenge attributes. Table 10 shows the success rate performance of each algorithm under different challenging attributes. Under the IV attribute, the SCACF has the second-highest AUC score after the GFSDCF. Under other attributes, the SCACF obtains the highest DP and AUC scores.

**Table 9.** Precision of 11 challenging attributes on OTB2015.

| Attribute | SCACF | GFSDCF | EFSCF | CSR-DCF | ARCF_H | Staple | IBRI | BiCF | BACF | STRCF |
|---|---|---|---|---|---|---|---|---|---|---|
| IV | 0.963 | 0.954 | 0.840 | 0.779 | 0.769 | 0.756 | 0.768 | 0.768 | 0.782 | 0.817 |
| OPR | 0.954 | 0.929 | 0.861 | 0.760 | 0.737 | 0.725 | 0.747 | 0.721 | 0.767 | 0.838 |
| SV | 0.940 | 0.918 | 0.833 | 0.739 | 0.736 | 0.711 | 0.744 | 0.729 | 0.755 | 0.828 |
| OCC | 0.912 | 0.880 | 0.835 | 0.795 | 0.676 | 0.707 | 0.710 | 0.684 | 0.714 | 0.795 |
| DEF | 0.928 | 0.897 | 0.835 | 0.777 | 0.740 | 0.728 | 0.748 | 0.679 | 0.747 | 0.823 |
| MB | 0.940 | 0.921 | 0.818 | 0.741 | 0.718 | 0.672 | 0.714 | 0.728 | 0.716 | 0.800 |
| FM | 0.946 | 0.937 | 0.792 | 0.766 | 0.758 | 0.710 | 0.727 | 0.748 | 0.787 | 0.802 |
| IPR | 0.957 | 0.947 | 0.829 | 0.781 | 0.750 | 0.752 | 0.741 | 0.769 | 0.777 | 0.796 |
| OV | 0.954 | 0.932 | 0.802 | 0.691 | 0.674 | 0.668 | 0.652 | 0.709 | 0.748 | 0.766 |
| BC | 0.959 | 0.921 | 0.890 | 0.778 | 0.803 | 0.749 | 0.788 | 0.797 | 0.801 | 0.872 |
| LR | 0.973 | 0.950 | 0.751 | 0.677 | 0.692 | 0.610 | 0.711 | 0.774 | 0.741 | 0.737 |

To comprehensively evaluate the performance of the SCACF in different tracking environments, extensive object-tracking algorithm experiments are conducted on the OTB dataset. The results show that the SCACF can better cope with complex tracking environments. Utilizing the proposed attention mechanism and anti-occlusion strategy, the SCACF can effectively avoid contaminating the filter with interfering information. Its performance is superior in challenging scenarios such as rotation, occlusion, and background clutter.

**Table 10.** Success rate of 11 challenging attributes on OTB2015.

| Attribute | SCACF | GFSDCF | EFSCF | CSR-DCF | ARCF_H | Staple | IBRI | BiCF | BACF | STRCF |
|---|---|---|---|---|---|---|---|---|---|---|
| IV | 0.723 | 0.727 | 0.672 | 0.594 | 0.591 | 0.578 | 0.596 | 0.596 | 0.608 | 0.638 |
| OPR | 0.695 | 0.680 | 0.650 | 0.547 | 0.535 | 0.525 | 0.553 | 0.536 | 0.570 | 0.618 |
| SV | 0.692 | 0.678 | 0.637 | 0.532 | 0.531 | 0.509 | 0.555 | 0.549 | 0.564 | 0.623 |
| OCC | 0.688 | 0.669 | 0.649 | 0.536 | 0.516 | 0.529 | 0.542 | 0.523 | 0.554 | 0.604 |
| DEF | 0.668 | 0.656 | 0.626 | 0.563 | 0.543 | 0.535 | 0.562 | 0.517 | 0.559 | 0.593 |
| MB | 0.700 | 0.699 | 0.653 | 0.588 | 0.554 | 0.521 | 0.562 | 0.582 | 0.557 | 0.635 |
| FM | 0.699 | 0.697 | 0.628 | 0.586 | 0.573 | 0.541 | 0.566 | 0.591 | 0.599 | 0.629 |
| IPR | 0.678 | 0.673 | 0.621 | 0.550 | 0.542 | 0.539 | 0.550 | 0.563 | 0.573 | 0.591 |
| OV | 0.700 | 0.690 | 0.617 | 0.520 | 0.486 | 0.476 | 0.498 | 0.526 | 0.547 | 0.584 |
| BC | 0.700 | 0.683 | 0.683 | 0.572 | 0.592 | 0.561 | 0.600 | 0.605 | 0.605 | 0.648 |
| LR | 0.691 | 0.650 | 0.559 | 0.417 | 0.460 | 0.399 | 0.514 | 0.550 | 0.532 | 0.538 |

### *4.3. Qualitative analysis*

This section selects five representative video sequences (Person1_s, Group3_4, Person21, Car7, and Wakeboard6) on the UAV123 dataset to intuitively demonstrate the tracking performance of the proposed algorithm in UAV-specific scenarios. The SCACF tracker is qualitatively compared to the other 8 state-of-the-art trackers (GFSDCF, FACF, ReCF, IBRI, BiCF, EFSCF, STRCF, and AutoTrack) as well as to the real object position (Ground Truth) of the video sequence. These video sequences encompass the typical challenges of object tracking, such as illumination changes, FM, OCC, rotation, BC, and similar objects. Figure 7 intercepts some frames of the video sequences to visualize the tracking under different tracking challenges.

The SCACF can achieve robust tracking in these UAV challenge scenarios while it remains stable in the case of the tracking failure of other trackers. Specifically, the following is observed:

1) Illumination variation: In Figure 7(a), the SCACF can adapt to the drastic variation of illumination and accurately track the object, while other trackers are prone to drift or lose the object due to illumination variation.

2) FM: As shown in Figure 7(d), in the case of the FM of the object, the SCACF can continuously lock on the object by virtue of its efficient filter updating mechanism. In contrast, except for the IBRI and FACF trackers, the other trackers lose the object due to an insufficient response speed.

3) OCC: In the video sequence Group3_4, the object is occluded by trees. The SCACF utilizes an anti-occlusion strategy to maintain tracking even when the object is partially or completely occluded, while other trackers are prone to tracking failure in occlusion situations.

4) Rotation and BC: The SCACF effectively suppresses the background interference information

by the group sparsity constraint. Even in situations of rotation or background clutter, the SCACF can accurately track the object.

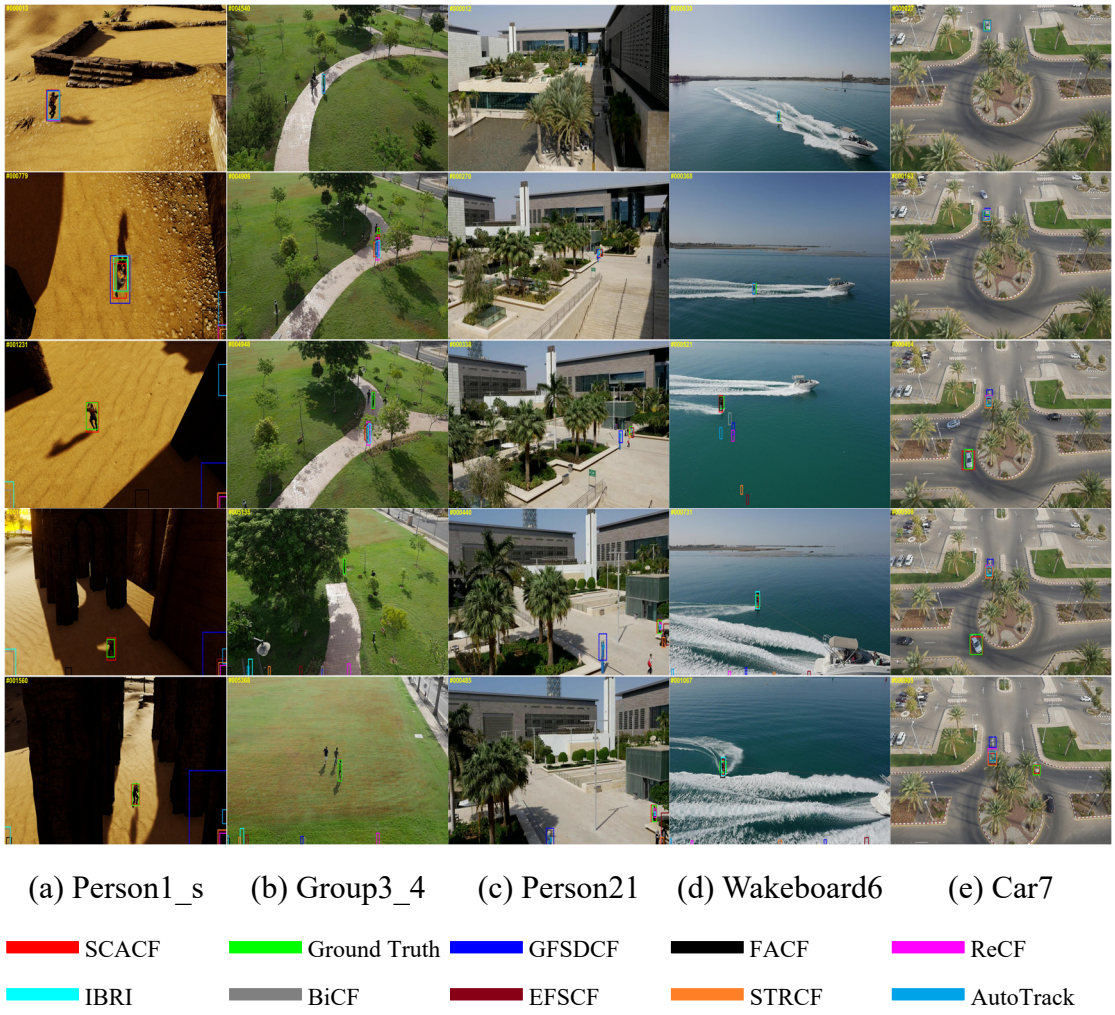



**Figure 7.** Visualize the tracking effect on different video sequences.

5) Similar object: In Figure 7(e), the object car and the background car are white. The SCACF can distinguish between the object and a similar background target, while other trackers are prone to drift under a similar object interference.

Benefiting from the proposed group sparsity constraint and anti-occlusion strategy, our tracker can suppress redundant and interfering information. This effectively avoids the effects caused by losing the object in the field of view (e.g., occlusion). Experiments further demonstrate the superior performance of the SCACF in UAV-specific scenarios.

### *4.4. Ablation experiments*

We performed ablation experiments on the OTB2015 dataset to further prove the effectiveness of the proposed method. To visualize the superiority of each mechanism, we visualized the average overlap rate and average center point error of the SCACF and control trackers.

4.4.1. Ablation experiment of group sparsity constraints

To verify the effectiveness of fiber group sparsity and slice group sparsity, this section compares the tracking results of the proposed SCACF and the BACF without group sparsity constraints in complex scenes. As shown in Figure 8, the red box represents the tracking results of the SCACF. The green box represents the real object position (Ground Truth) of the video sequence Biker. The blue box represents the tracking result of the BACF. In the initial frame to frame 47, the object moves forward, and both algorithms achieve good tracking results. In frames 66 to 71, the object is rotated by 90 degrees. The SCACF with group sparsity feature selection can better exploit the structural sparsity property of the filter and suppress the redundant information in the background to achieve accurate tracking. In contrast, the BACF is contaminated by the background interference information, thus failing to capture the object. At frame 140, the object has been rotated by 180 degrees. The SCACF and Ground Truth continue to robustly track the object. while the BACF fails to track.

The experimental results show that the SCACF with the introduction of group sparsity constraints can effectively cope with object rotation and achieve stable and accurate object tracking. The algorithm demonstrates excellent tracking performance in complex scenarios, which fully verifies the competitiveness of group sparsity constraints in the object-tracking task.

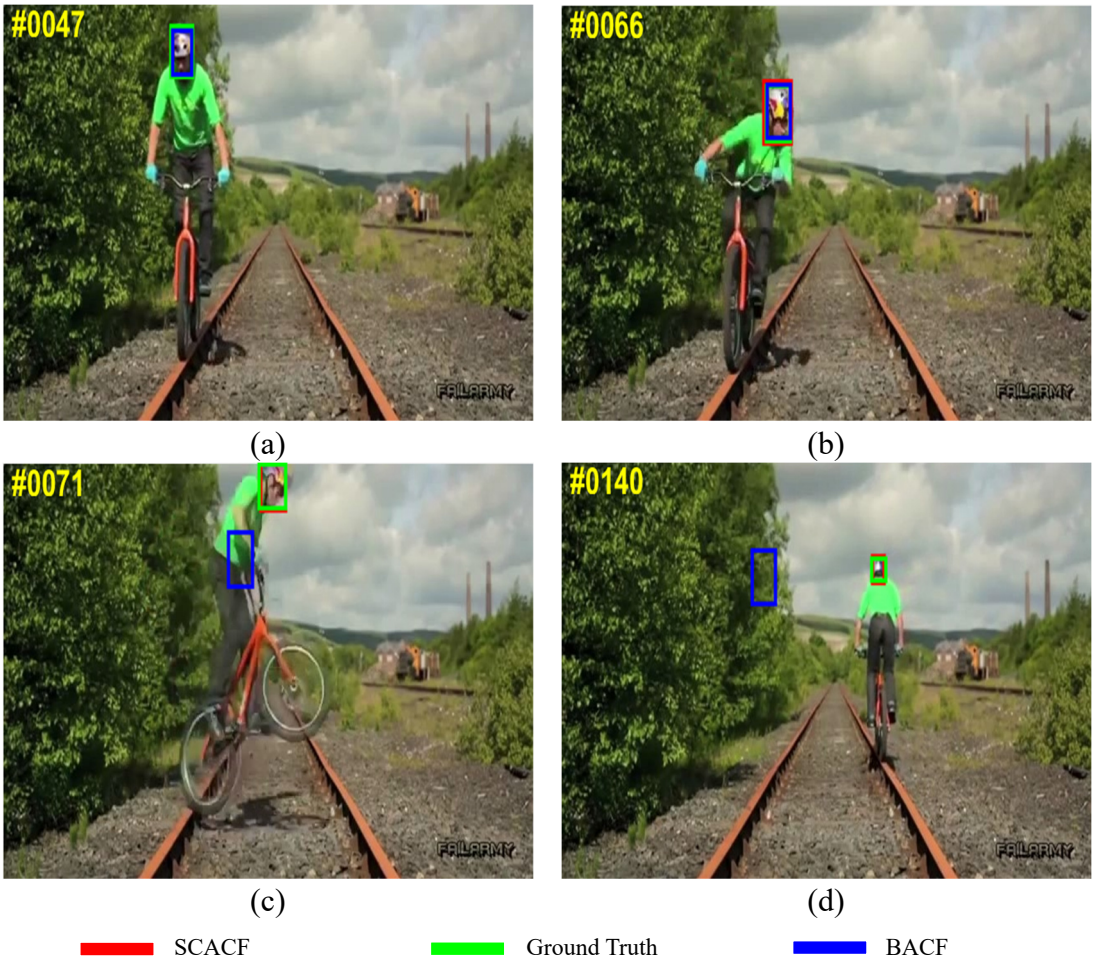



**Figure 8.** Comparison experiment with or without group sparsity constraints.

4.4.2. Ablation experiment of deep features

To evaluate the tracking performance improvement by deep features in complex scenarios, this section conducts comparative experiments between the SCACF and the SCACF (HC) based on hand-crafted features. As shown in Figure 9, the red box represents the tracking results of the SCACF. The green box represents the real object position (Ground Truth) of the video sequence Jogging-1. The blue box represents the tracking results of the SCACF (HC). In frames 13 to 68, both algorithms achieve normal tracking of the object. However, as the complexity of the scene increases, the performance differences between the two algorithms gradually appear. In frames 68 to 84, the SCACF (HC) mistakenly recognizes the utility pole as the object due to its limited feature discrimination capability, which leads to a tracking failure. On the other hand, the SCACF shows excellent robustness and a high accuracy tracking ability due to the strong discriminative ability of its deep features. At frame 244, the SCACF(HC) suffers from tracking drift because it is unable to reposition the exact location of the object. In contrast, the SCACF and Ground Truth remain consistent, thus accurately identifying and tracking the object stably.

The experimental results show that the introduction of deep features significantly improves the tracking performance of the SCACF in complex scenes. Compared with the SCACF (HC) based on hand-crafted features, the SCACF can more accurately capture the semantic information of the object, thus maintaining higher tracking precision and stability.

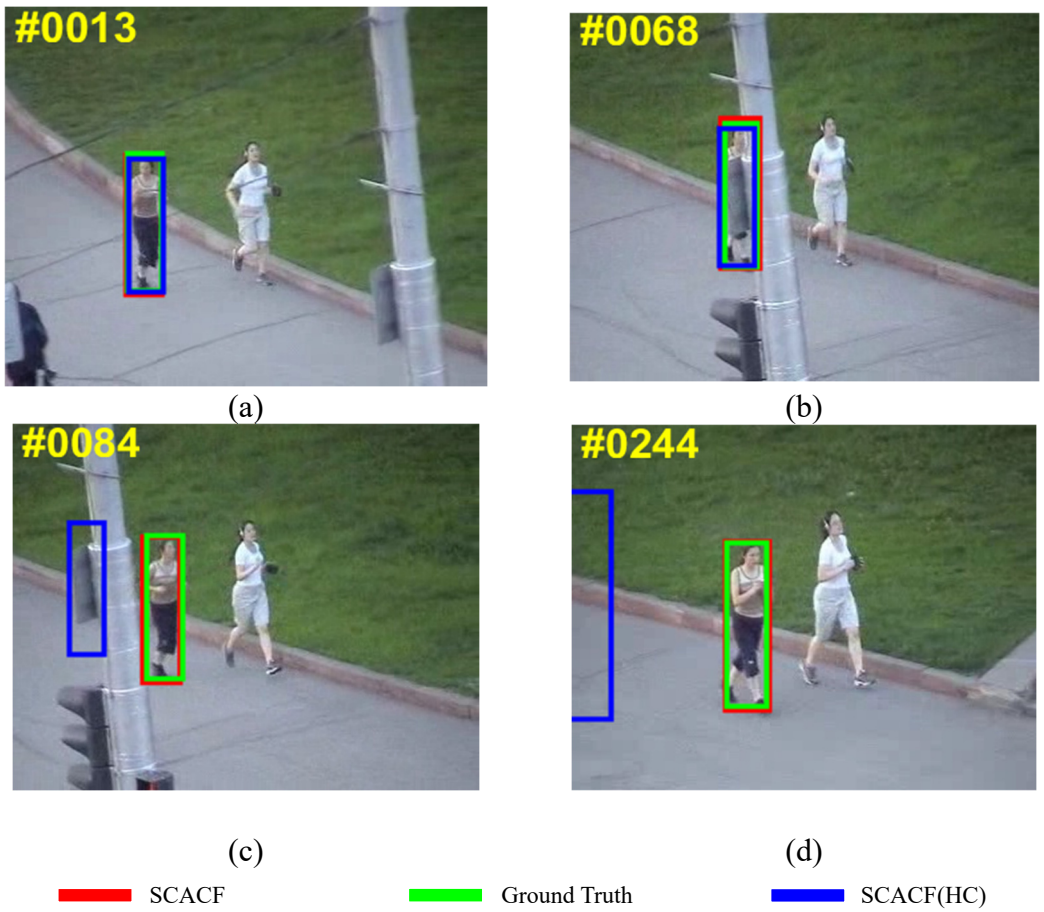



**Figure 9.** Comparison experiment with or without deep features.

4.4.3. Ablation experiment of anti-occlusion strategy

To verify the effect of the anti-occlusion strategy on the tracking performance, this section compares the performance of the SCACF with the anti-occlusion strategy and the GFSDCF without the anti-occlusion strategy in occlusion scenarios. As shown in Figure 10, the red box represents the tracking results of the SCACF. The green box represents the real object position (Ground Truth) of the video sequence HuMan3. The blue box represents the tracking results of the GFSDCF. At frame 15, both algorithms achieve excellent tracking of the object. However, when the object is occluded by a passerby, the GFSDCF (blue box) mistakenly recognizes the passerby as the object due to the lack of an anti-occlusion strategy, which results in a tracking failure. In contrast, the SCACF (red box) effectively resists interference and continues to accurately track the object by virtue of its anti-occlusion strategy. In frames 37 to 148, the object continuously passes through complex occlusion scenarios such as traffic lights, pedestrians, road signs, etc. The SCACF still performs well and can stably track the object. In contrast, the GFSDCF has completely lost the object because it cannot cope with the occlusion problem. In the subsequent frames, the SCACF and Ground Truth always successfully lock the object.

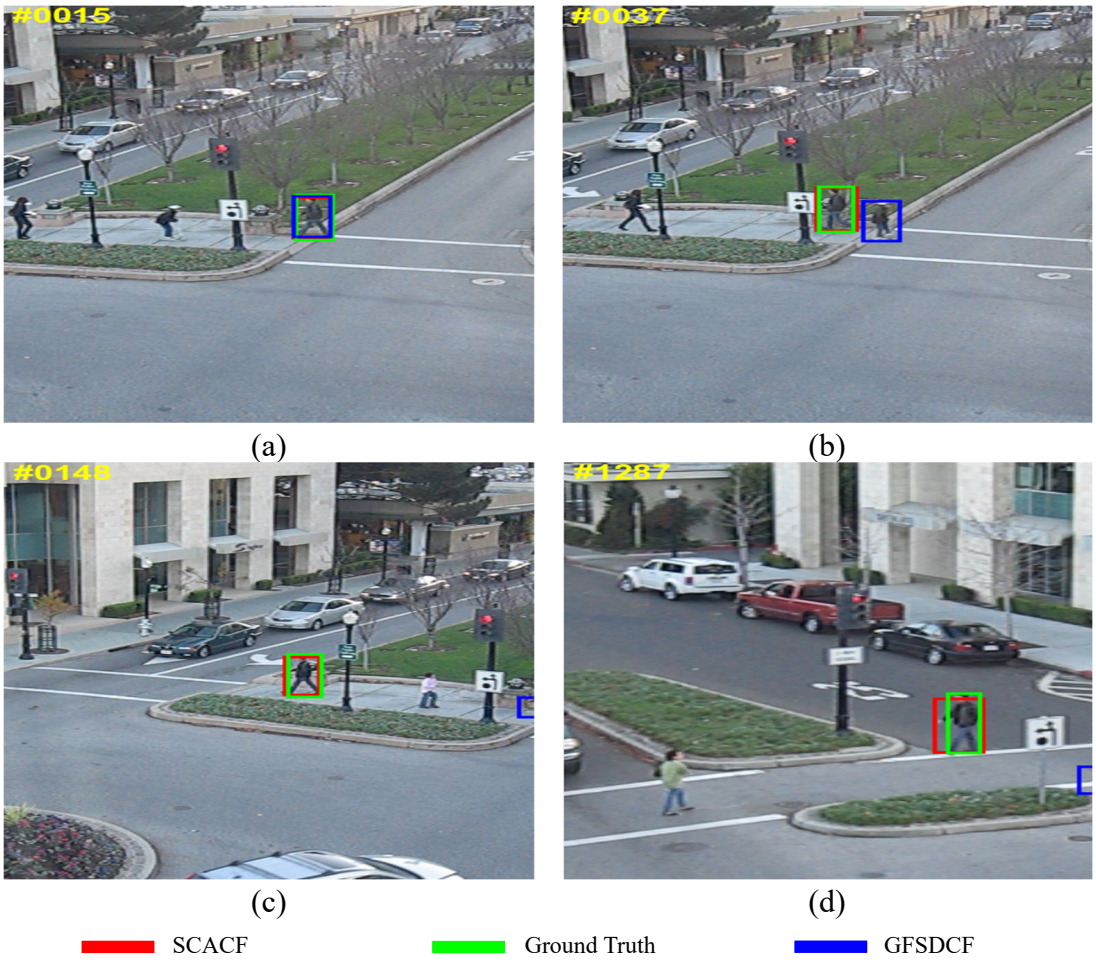



**Figure 10.** Comparison experiment with or without anti-occlusion strategy.

The experimental results show that the introduction of the anti-occlusion strategy significantly

improves the robustness of the SCACF in occlusion scenarios. The algorithm can maintain accurate tracking of the object when the object is partially or completely occluded, which effectively solves the tracking drift problem caused by occlusion. This improvement makes the SCACF more adaptable in practical applications, especially in dynamic and complex scenarios of UAV tracking.

#### 4.4.4. Overlap rate and center point error visualization

In Figure 11, we visualize the average tracking overlap rate and average center point error of the video sequences of the three ablation experiments to further validate the performance advantages of the SCACF. The higher average tracking overlap rate indicates an improved tracking performance. The smaller average center point error indicates the smaller deviation of the algorithm's tracking results from the real position of the object.

In the video sequence Biker, the average tracking overlap rate of the SCACF is 0.73, which is significantly higher than that of the BACF (0.38). Meanwhile, the average center point error of the SCACF is 2.64 pixels, which is much lower than that of the BACF (81.78 pixels). The results show that the group sparsity constraint is one of the key factors to improve the performance of the object-tracking algorithm.

In the video sequence Jogging-1, the average tracking overlap rate of the SCACF is 0.71, while that of the SCACF (HC) is only 0.18. Meanwhile, the average center point error of the SCACF is 6.24 pixels, which is significantly better than that of the 95.63 pixels of the SCACF (HC). This comparison verifies the important role of deep features in improving the tracking performance.

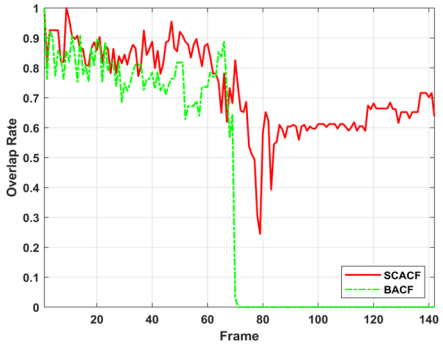
In the video sequence HuMan3, the average tracking overlap rate of the SCACF is 0.60, which is much higher than that of the GFSDCF (0.01). Meanwhile, the average center point error of the SCACF is 7.02 pixels, which is significantly lower than that of the GFSDCF (250.98 pixels). This result fully demonstrates the effectiveness of the anti-occlusion strategy in dealing with occluded scenes.

Table 11 lists the average tracking overlap rates of the SCACF and other tracking algorithms for six typical video sequences. The SCACF ranks first with an overall average tracking overlap rate of 0.697. Table 12 lists the average center point error of the SCACF and other tracking algorithms for six typical video sequences. The total mean pixel error of the SCACF is 6.727 pixels, which is ranked first among the ten algorithms by a significant margin.

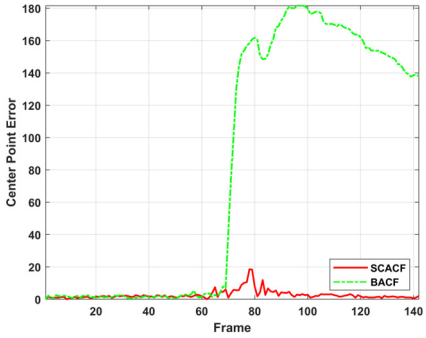
In summary, SCACF performs well in two key metrics: average tracking overlap rate and average center point error. The algorithm provides strong technical support for object-tracking tasks in complex scenarios.

**Table 11.** Average tracking overlap rate of each tracking algorithm in some videos.

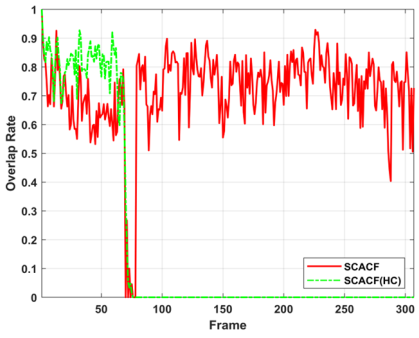
| Video | SCACF | GFSDCF | EFSCF | CSR-DCF | ARCF_H | Staple | IBRI | BiCF | BACF | STRCF |
|---|---|---|---|---|---|---|---|---|---|---|
| Biker | 0.73 | 0.70 | 0.39 | 0.23 | 0.35 | 0.26 | 0.40 | 0.42 | 0.38 | 0.38 |
| HuMan3 | 0.60 | 0.01 | 0.63 | 0.02 | 0.02 | 0.02 | 0.02 | 0.02 | 0.02 | 0.62 |
| Boy | 0.85 | 0.82 | 0.83 | 0.76 | 0.77 | 0.82 | 0.83 | 0.82 | 0.76 | 0.84 |
| Coke | 0.66 | 0.60 | 0.56 | 0.54 | 0.55 | 0.57 | 0.53 | 0.58 | 0.58 | 0.56 |
| Ironman | 0.55 | 0.42 | 0.16 | 0.14 | 0.12 | 0.09 | 0.15 | 0.52 | 0.14 | 0.10 |
| Walking2 | 0.79 | 0.78 | 0.81 | 0.35 | 0.78 | 0.78 | 0.79 | 0.68 | 0.79 | 0.81 |
| Average | 0.697 | 0.555 | 0.563 | 0.340 | 0.432 | 0.423 | 0.453 | 0.507 | 0.445 | 0.552 |



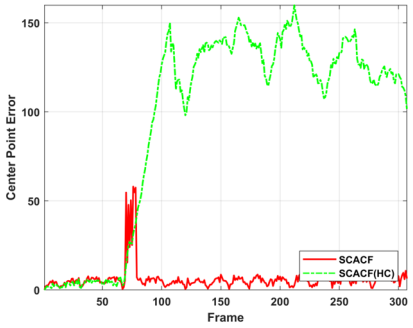

(a) Overlap rate of the Biker video sequence



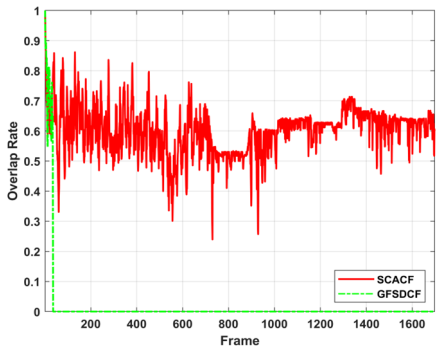

(b) Center point error of the Biker video sequence



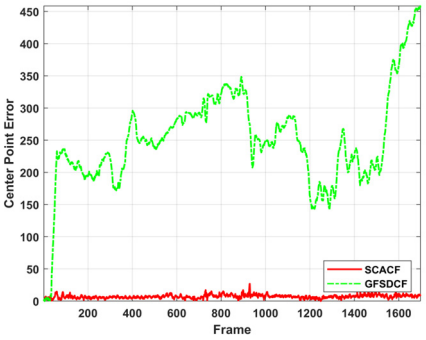

(c) Overlap rate of the Jogging-1 video sequence




(d) Center point error of the Jogging-1 video sequence




(e) Overlap rate of the HuMan3 video sequence




(f) Center point error of the HuMan3 video sequence

**Figure 11.** Visualize the overlap rate and center point error of the video sequences.

**Table 12.** Average center point error of each tracking algorithm in some videos.

| Video | SCACF | GFSDCF | EFSCF | CSR-DCF | ARCF_H | Staple | IBRI | BiCF | BACF | STRCF |
|---|---|---|---|---|---|---|---|---|---|---|
| Biker | 2.64 | 2.54 | 37.76 | 31.80 | 87.70 | 79.43 | 49.30 | 34.54 | 81.78 | 36.99 |
| HuMan3 | 7.02 | 250.98 | 3.67 | 256.42 | 238.81 | 303.73 | 246.72 | 243.60 | 213.63 | 3.58 |
| Boy | 1.77 | 2.35 | 1.93 | 3.46 | 1.51 | 2.49 | 2.20 | 2.60 | 1.50 | 1.86 |
| Coke | 10.16 | 11.68 | 13.51 | 14.62 | 14.11 | 12.20 | 14.50 | 12.54 | 12.86 | 14.05 |
| Ironman | 15.42 | 37.73 | 57.35 | 60.61 | 222.99 | 81.95 | 196.58 | 30.02 | 174.30 | 233.89 |
| Walking2 | 3.35 | 4.06 | 2.70 | 40.19 | 3.57 | 3.43 | 3.07 | 7.90 | 2.37 | 2.43 |
| Average | 6.727 | 51.557 | 19.487 | 67.850 | 94.782 | 80.538 | 85.395 | 55.200 | 81.073 | 48.800 |

### *4.5. Limitations*

Table 13 shows the FPS of each algorithm for 10 video sequences on the UAV123 dataset. The top three algorithms in the FPS performance evaluation for each video sequence are labeled in red, green, and blue, respectively. We visually compare the real-time performance of each algorithm. The results show that SCACF underperforms in terms of FPS.

Figure 12 visualizes three tracking failure cases of the SCACF. The red box represents the tracking results of the SCACF, and the green box represents the real object position (Ground Truth) of the video sequence. The experimental results show that the anti-occlusion strategy of the SCACF fails to play a full role in scenes where the object undergoes prolonged OCC or completely disappears from the field of view, which results in tracking failures. Although the SCACF performs well in most complex scenarios, there is still room to improve its performance in extreme cases (e.g., long-time occlusion or object disappearance reappearance).

**Table 13.** FPS of some video sequences on UAV123.

| Video | SCACF | GFSDCF | FACF | AutoTrack | ARCF_H | ReCF | IBRI | BiCF | EFSCF | STRCF |
|---|---|---|---|---|---|---|---|---|---|---|
| Bird1_1 | 2.58 | 4.04 | 47.37 | 6.26 | 6.34 | 38.61 | 13.65 | 24.29 | 18.12 | 18.81 |
| Bird1_2 | 2.86 | 4.20 | 3.29 | 6.52 | 6.31 | 46.41 | 14.27 | 30.97 | 19.09 | 19.21 |
| Car6_5 | 2.05 | 2.83 | 33.35 | 5.05 | 6.30 | 35.56 | 10.14 | 24.37 | 14.44 | 14.84 |
| Car7 | 2.43 | 3.61 | 53.10 | 5.78 | 6.74 | 46.50 | 14.51 | 31.69 | 18.42 | 18.41 |
| Car11 | 2.58 | 3.55 | 57.71 | 5.89 | 10.20 | 72.92 | 21.20 | 43.39 | 28.08 | 29.31 |
| Group3_2 | 2.29 | 2.97 | 55.28 | 5.69 | 7.73 | 64.44 | 20.95 | 40.37 | 27.63 | 22.40 |
| Group3_4 | 2.18 | 2.99 | 53.77 | 5.56 | 6.74 | 67.66 | 18.09 | 41.77 | 23.29 | 24.11 |
| Person22 | 1.97 | 2.32 | 55.05 | 5.35 | 9.01 | 69.63 | 20.68 | 40.85 | 29.21 | 29.76 |
| Uav1_2 | 1.61 | 1.84 | 43.68 | 5.60 | 6.28 | 38.85 | 14.64 | 25.65 | 17.59 | 17.76 |
| Person1_s | 1.44 | 1.55 | 41.48 | 5.60 | 6.19 | 37.09 | 12.94 | 23.66 | 16.51 | 17.22 |

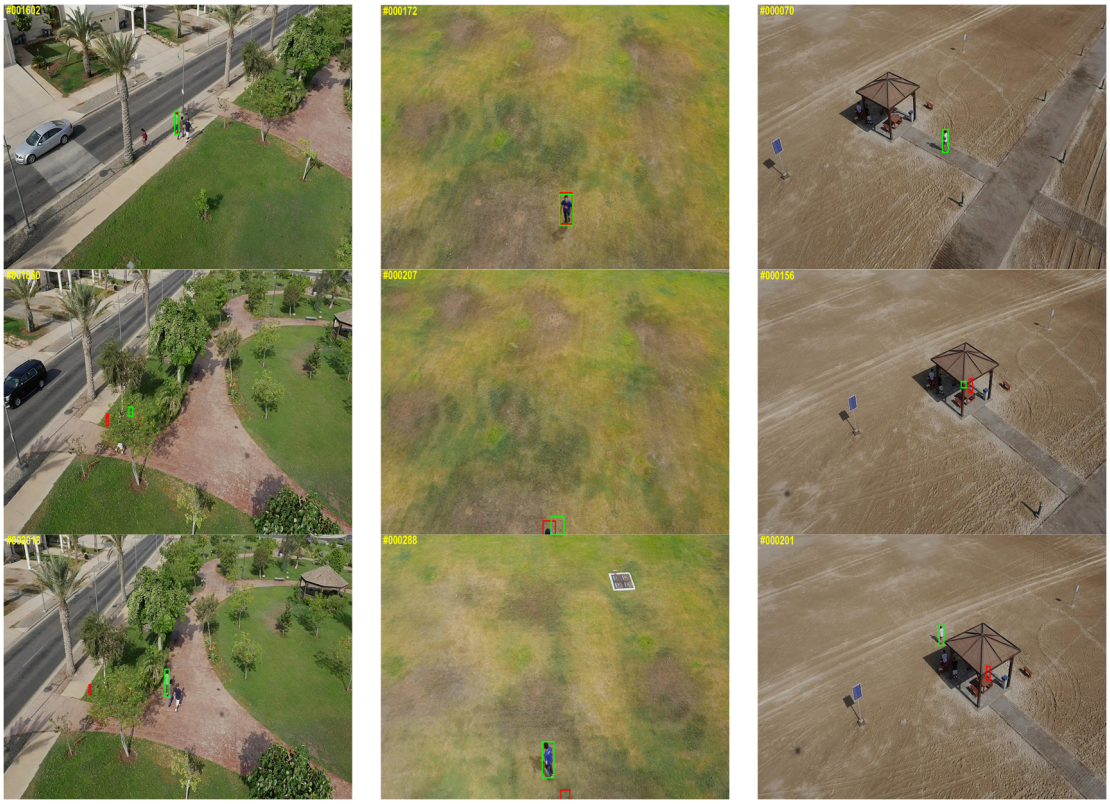



**Figure 12.** Visualize the tracking failure cases of SCACF.

## 5. Conclusions

To realize robust UAV tracking, this paper proposed a correlation filter algorithm for UAV tracking based on a spatial and channel attention mechanism (SCACF). The SCACF captures the key structural features of the object through the group sparsity feature selection strategy in the space and channel and fully exploits the structural sparsity property of the filter. This strategy effectively removes redundant and interfering information, thus enhancing the filter discrimination ability. In addition, the SCACF constructs a history template pool by screening reliable samples. This method avoids the filter degradation problem while ensuring the variety of templates, which further enhances the stability of the algorithm. Extensive experiments on the UAV123, UAV20L, and DTB70 UAV benchmark datasets showed that SCACF performed excellently in the UAV object tracking task. Its precision and robustness were better than the existing state-of-the-art algorithms, which fully verify its effectiveness in practical applications. Meanwhile, experiments on the OTB50 and OTB2015 datasets further demonstrated that the SCACF could adapt to diverse tracking scenarios. Overall, the experimental results demonstrate its potential in the field of generalized object tracking. However, the SCACF was also found to run slowly in the experiments. Future research will dynamically adjust the algorithm complexity for the object motion trajectory. The tracking speed will be significantly improved while maintaining high precision to meet the demand for real-time performance in practical applications.

## 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 is supported by the Natural Science Foundation of Fujian Province (2024J01820, 2024J01821, 2024J01822), Natural Science Foundation Project of Zhangzhou City (ZZ2023J37), the Principal Foundation of Minnan Normal University (KJ19019), the High-level Science Research Project of Minnan Normal University (GJ19019), Research Project on Education and Teaching of Undergraduate Colleges and Universities in Fujian Province (FBJY20230083), Guangdong Province Natural Science Foundation (2024A1515011766), the State Key Laboratory Major Special Projects of Jilin Province Science and Technology Development Plan (SKL202402024) and National Independent Innovation Demonstration Zone System (Fuzhou, Xiamen, Quanzhou) Innovation Platform Project (3502ZCQXT2024006).

## Conflict of interest

The authors declare there is no conflict of interest.

## References

1. A. Kumar, R. Vohra, R. Jain, M. Li, C. Gan, D. K. Jain, Correlation filter based single object tracking: A review, *Inf. Fusion*, **112** (2024), 102562. https://doi.org/10.1016/j.inffus.2024.102562
2. S. Arthanari, D. Elayaperumal, Y. H. Joo, Learning temporal regularized spatial-aware deep correlation filter tracking via adaptive channel selection, *Neural Networks*, **186** (2025), 107210. https://doi.org/10.1016/j.neunet.2025.107210
3. X. F. Zhu, X. J. Wu, T. Xu, Z. H. Feng, J. Kittler, Robust visual object tracking via adaptive attribute-aware discriminative correlation filters, *IEEE Trans. Multimedia*, **24** (2021), 301–312. https://doi.org/10.1109/TMM.2021.3050073
4. S. Ma, B. Zhao, Z. Hou, W. Yu, L. Pu, X. Yang, SOCF: A correlation filter for real-time UAV tracking based on spatial disturbance suppression and object saliency-aware, *Expert Syst. Appl.*, **238** (2024), 122131. https://doi.org/10.1016/j.eswa.2023.122131
5. S. Li, Y. Liu, Q. Zhao, Z. Feng, Learning residue-aware correlation filters and refining scale for real-time UAV tracking, *Pattern Recognit.*, **127** (2022), 108614. https://doi.org/10.1016/j.patcog.2022.108614
6. P. Lai, M. Zhang, G. Cheng, S. Li, X. Huang, J. Han, Target-aware transformer for satellite video object tracking, *IEEE Trans. Geosci. Remote Sens.*, **62** (2023), 1–10. https://doi.org/10.1109/TGRS.2023.3339658
7. Y. Li, N. Wang, W. Li, X. Li, M. Rao, Object tracking in satellite videos with distractor–occlusion-aware correlation particle filters, *IEEE Trans. Geosci. Remote Sens.*, **62** (2024), 1–12. https://doi.org/10.1109/TGRS.2024.3353298

8. X. Qu, Z. Ma, H. Zhang, X. Sun, X. Yang, Target tracking method based on scale-adaptive rotation kernelized correlation filter for through-the-wall radar, *IEEE Signal Process Lett.*, **32** (2025), 1001–1005. https://doi.org/10.1109/LSP.2025.3540954
9. 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
10. Z. Chen, L. J. Liu, Z. Yu, Learning dynamic distractor-repressed correlation filter for real-time UAV tracking, *IEEE Signal Process Lett.*, **32** (2025), 616–620. https://doi.org/10.1109/LSP.2024.3522850
11. 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
12. P. Feng, C. Xu, Z. Zhao, F. Liu, J. Guo, C. Yuan, et al., A deep features based generative model for visual tracking, *Neurocomputing*, **308** (2018), 245–254. https://doi.org/10.1016/j.neucom.2018.05.007
13. 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), 1446–1475. https://doi.org/10.3934/era.2025068
14. 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
15. H. Zhu, H. Peng, G. Xu, L. Deng, Y. Cheng, A. Song, Bilateral weighted regression ranking model with spatial-temporal correlation filter for visual tracking, *IEEE Trans. Multimedia*, **24** (2021), 2098–2111. https://doi.org/10.1109/TMM.2021.3075876
16. Y. Liang, Y. Liu, Y. Yan, L. Zhang, H. Wang, Robust visual tracking via spatio-temporal adaptive and channel selective correlation filters, *Pattern Recognit.*, **112** (2021), 107738. https://doi.org/10.1016/j.patcog.2020.107738
17. F. Li, C. Tian, W. Zuo, L. Zhang, M. H. Yang, Learning spatial-temporal regularized correlation filters for visual tracking, in *IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, (2018), 4904–4913. https://doi.org/10.1109/CVPR.2018.00515
18. M. Danelljan, G. Hager, F. S. Khan, M. Felsberg, Learning spatially regularized correlation filters for visual tracking, in *IEEE International Conference on Computer Vision*, IEEE, (2015), 4310–4318. https://doi.org/10.1109/ICCV.2015.490
19. K. Nai, S. Chen, Learning a novel ensemble tracker for robust visual tracking, *IEEE Trans. Multimedia*, **26** (2023), 3194–3206. https://doi.org/10.1109/TMM.2023.3307939
20. Q. Hu, H. Wu, J. Wu, J. Shen, H. Hu, Y. Chen, et al., Spatio-temporal self-learning object tracking model based on anti-occlusion mechanism, *Eng. Lett.*, **31** (2023), 1141–1150.
21. Y. Chen, H. Wu, Z. Deng, J. Zhang, H. Wang, L. Wang, et al., Deep-feature-based asymmetrical background-aware correlation filter for object tracking, *Digital Signal Process.*, **148** (2024), 104446. https://doi.org/10.1016/j.dsp.2024.104446
22. Z. Huang, C. Fu, Y. Li, F. Lin, P. Lu, Learning aberrance repressed correlation filters for real-time UAV tracking, in *IEEE International Conference on Computer Vision*, IEEE, (2019), 2891–2900. https://doi.org/10.1109/ICCV.2019.00298

23. J. Liao, C. Qi, J. Cao, Temporal constraint background-aware correlation filter with saliency map, *IEEE Trans. Multimedia*, **23** (2020), 3346–3361. https://doi.org/10.1109/TMM.2020.3023794
24. P. Yang, Q. Wang, J. Dou, L. Dou, SDCS-CF: Saliency-driven localization and cascade scale estimation for visual tracking, *J. Visual Commun. Image Represent.*, **98** (2024), 104040. https://doi.org/10.1016/j.jvcir.2023.104040
25. M. Danelljan, A. Robinson, F. S. Khan, M. Felsberg, Beyond correlation filters: Learning continuous convolution operators for visual tracking, in *European Conference on Computer Vision*, Springer, (2016), 472–488. https://doi.org/10.1007/978-3-319-46454-1_29
26. M. Danelljan, G. Bhat, F. S. Khan, M. Felsberg, Eco: Efficient convolution operators for tracking, in *IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, (2017), 6638–6646. https://doi.org/10.48550/arXiv.1611.09224
27. J. Valmadre, L. Bertinetto, J. Henriques, A. Vedaldi, P. H. Torr, End-to-end representation learning for correlation filter based tracking, in *IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, (2017), 2805–2813. https://doi.org/10.1109/CVPR.2017.531
28. T. Xu, Z. H. Feng, X. J. Wu, J. Kittler, Joint group feature selection and discriminative filter learning for robust visual object tracking, in *IEEE International Conference on Computer Vision*, IEEE, (2019), 7950–7960. https://doi.org/10.1109/ICCV.2019.00804
29. J. Wen, H. Chu, Z. Lai, T. Xu, L. Shen, Enhanced robust spatial feature selection and correlation filter learning for UAV tracking, *Neural Networks*, **161** (2023), 39–54. https://doi.org/10.1016/j.neunet.2023.01.003
30. Y. Q. Su, F. Xu, Z. S. Wang, M. C. Sun, H. Zhao, A context constraint and sparse learning based on correlation filter for high-confidence coarse-to-fine visual tracking, *Expert Syst. Appl.*, **268** (2025), 126225. https://doi.org/10.1016/j.eswa.2024.126225
31. M. Mueller, N. G. Smith, B. Ghanem, A benchmark and simulator for UAV tracking, in *European Conference on Computer Vision*, Springer, (2016), 445–461. https://doi.org/10.1007/978-3-319-46448-0_27
32. S. Li, D. Y. Yeung, Visual object tracking for unmanned aerial vehicles: A benchmark and new motion models, in *AAAI Conference on Artificial Intelligence*, AAAI Press, (2017), 4140–4146. https://doi.org/10.1609/aaai.v31i1.11205
33. Y. Wu, J. Lim, M. H. Yang, Online object tracking: A benchmark, in *IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, (2013), 2411–2418. https://doi.org/10.1109/CVPR.2013.312
34. D. S. Bolme, J. R. Beveridge, B. A. Draper, Y. M. Lui, Visual object tracking using adaptive correlation filters, in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, IEEE, (2010), 2544–2550. https://doi.org/10.1109/CVPR.2010.5539960
35. J. F. Henriques, R. Caseiro, P. Martins, J. Batista, Exploiting the circulant structure of tracking-by-detection with kernels, in *European Conference on Computer Vision*, Springer, (2012), 702–715. https://doi.org/10.1007/978-3-642-33765-9
36. J. F. Henriques, R. Caseiro, P. Martins, J. Batista, High-speed tracking with kernelized correlation filters, *IEEE Trans. Pattern Anal. Mach. Intell.*, **37** (2014), 583–596. https://doi.org/10.1109/TPAMI.2014.2345390
37. N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, IEEE, (2005), 886–893. https://doi.org/10.1109/CVPR.2005.177

38. L. Bertinetto, J. Valmadre, S. Golodetz, O. Miksik, P. H. S. Torr, Staple: Complementary learners for real-time tracking, in *IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, (2016), 1401–1409. https://doi.org/10.1109/CVPR.2016.156
39. J. Van De Weijer, C. Schmid, J. Verbeek, D. Larlus, Learning color names for real-world applications, *IEEE Trans. Image Process.*, **18** (2009), 1512–1523. https://doi.org/10.1109/TIP.2009.2019809
40. M. Fiaz, A. Mahmood, S. Javed, S. K. Jung, Handcrafted and deep trackers: Recent visual object tracking approaches and trends, *ACM Comput. Surv.*, **52** (2019), 1–44. https://doi.org/10.48550/arXiv.1812.07368
41. M. Oquab, L. Bottou, I. Laptev, J. Sivic, Learning and transferring mid-level image representations using convolutional neural networks, in *IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, (2014), 1717–1724. https://doi.org/10.1109/CVPR.2014.222
42. E. Shelhamer, J. Long, T. Darrell, Fully convolutional networks for semantic segmentation, *IEEE Trans. Pattern Anal. Mach. Intell.*, **39** (2016), 640–651. https://doi.org/10.48550/arXiv.1411.4038
43. K. Chen, L. Wang, H. Wu, C. Wu, Y. Liao, Y. Chen, et al., Background-aware correlation filter for object tracking with deep CNN features, *Eng. Lett.*, **32** (2024), 1351–1363.
44. D. Yuan, X. Chang, P. Y. Huang, Q. Liu, Z. He, Self-supervised deep correlation tracking, *IEEE Trans. Image Process.*, **30** (2020), 976–985. https://doi.org/10.1109/TIP.2020.3037518
45. X. Li, C. Ma, B. Wu, Z. He, M. H. Yang, Target-aware deep tracking, in *IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, (2019), 1369–1378. https://doi.org/10.1109/CVPR.2019.00146
46. K. Qian, J. Shen, S. Wang, Y. Wu, G. Lu, SiamUF: SiamCar based small UAV tracker using dense U-shape deep features in near infrared videos, *Opt. Lasers Eng.*, **186** (2025), 108825. https://doi.org/10.1016/j.optlaseng.2025.108825
47. H. Nam, B. Han, Learning multi-domain convolutional neural networks for visual tracking, in *IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, (2016), 4293–4302. https://doi.org/10.1109/CVPR.2016.465
48. I. Jung, J. Son, M. Baek, B. Han, Real-time MDNet, in *European Conference on Computer Vision*, Springer, (2018), 83–98. https://doi.org/10.1007/978-3-030-01225-0_6
49. K. Yang, Z. He, W. Pei, Z. Zhou, X. Li, D. Yuan, et al., SiamCorners: Siamese corner networks for visual tracking, *IEEE Trans. Multimedia*, **24** (2021), 1956–1967. https://doi.org/10.1109/TMM.2021.3074239
50. L. Bertinetto, J. Valmadre, J. F. Henriques, A. Vedaldi, P. H. Torr, Fully-convolutional Siamese networks for object tracking, in *European Conference on Computer Vision*, Springer, (2016), 850–865. https://doi.org/10.48550/arXiv.1606.09549
51. Q. Wang, Z. Teng, J. Xing, J. Gao, W. Hu, S. Maybank, Learning attentions: Residual attentional Siamese network for high performance online visual tracking, in *IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, (2018), 4854–4863. https://doi.org/10.1109/CVPR.2018.00510
52. Q. Guo, W. Feng, C. Zhou, R. Huang, L. Wan, S. Wang, Learning dynamic Siamese network for visual object tracking, in *IEEE International Conference on Computer Vision*, IEEE, (2017), 1763–1771. https://doi.org/10.1109/ICCV.2017.196

53. B. Li, J. Yan, W. Wu, Z. Zhu, X. Hu, High performance visual tracking with Siamese region proposal network, in *IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, (2018), 8971–8980. https://doi.org/10.1109/CVPR.2018.00935
54. B. Li, W. Wu, Q. Wang, F. Zhang, J. Xing, J. Yan, Siamrpn++: Evolution of Siamese visual tracking with very deep networks, in *IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, (2019), 4282–4291. http://dx.doi.org/10.1109/CVPR.2019.00441
55. K. Yang, Q. Li, C. Tian, H. Zhang, A. Shi, J. Li, DeforT: Deformable transformer for visual tracking, *Neural Networks*, **176** (2024), 106380. https://doi.org/10.1016/j.neunet.2024.106380
56. 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
57. A. Lukezic, T. Vojir, L. C. Zajc, J. Matas, M. Kristan, Discriminative correlation filter with channel and spatial reliability, in *IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, (2017), 6309–6318. https://doi.org/10.1007/s11263-017-1061-3
58. Y. Yu, Y. Xiong, W. Huang, M. R. Scott, Deformable Siamese attention networks for visual object tracking, in *IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, (2020), 6728–6737. https://doi.org/10.1109/CVPR42600.2020.00676
59. B. Yan, H. Peng, J. Fu, D. Wang, H. Lu, Learning spatio-temporal transformer for visual tracking, in *IEEE International Conference on Computer Vision*, IEEE, (2021), 10448–10457. https://doi.org/10.48550/arXiv.2103.17154
60. Y. Huang, Y. Chen, C. Lin, Q. Hu, J. Song, Visual attention learning and antiocclusion-based correlation filter for visual object tracking, *J. Electron. Imaging*, **32** (2023), 013023. https://doi.org/10.1117/1.JEI.32.1.013023
61. 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** (2021), 6301–6313. https://doi.org/10.1109/TGRS.2020.3030265
62. F. Lin, C. Fu, Y. He, F. Guo, Q. Tang, BiCF: Learning bidirectional incongruity-aware correlation filter for efficient UAV object tracking, in *IEEE International Conference on Robotics and Automation*, IEEE, (2020), 2365–2371. https://doi.org/10.1109/ICRA40945.2020.9196530
63. 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
64. Y. Li, C. Fu, F. Ding, Z. Huang, G. Lu, AutoTrack: Towards high-performance visual tracking for UAV with automatic spatio-temporal regularization, in *IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, (2020), 11923–11932. https://doi.org/10.1109/CVPR42600.2020.01194
65. F. Zhang, S. Ma, L. Yu, Y. Zhang, Z. Qiu, Z. Li, Learning future-aware correlation filters for efficient UAV tracking, *Remote Sens.*, **13** (2021), 4111. https://doi.org/10.3390/rs13204111

66. H. K. Galoogahi, A. Fagg, S. Lucey, Learning background-aware correlation filters for visual tracking, in *IEEE International Conference on Computer Vision*, IEEE, (2017), 1135–1143. https://doi.org/10.1109/ICCV.2017.129


AIMS Press