Journal of Real-Time Image Processing (2023) 20:51
https://doi.org/10.1007/s11554-023-01306-7

RESEARCH

Learning discriminative correlation filters via saliency-aware channel selection for robust visual object tracking

Sugang Ma[1,2] · Zhixian Zhao[1] · Lei Pu[3] · Zhiqiang Hou[1] · Lei Zhang[4] · Xiangmo Zhao[2]

Received: 8 November 2022 / Accepted: 18 April 2023 / Published online: 5 May 2023

Abstract

In recent years, discriminative correlation filters (DCF) with deep features have achieved excellent results in visual object tracking tasks. These trackers usually use multi-channel features of the fixed layer of the pre-trained network model to represent the target. However, the multi-channel features contain many interfering channels that are not conducive to object representation, resulting in overfitting and high computational complexity. To solve this problem, we research the correlation between multi-channel deep features and target saliency information and propose a novel DCF tracking method based on saliency-aware and adaptive channel selection. Specifically, we adaptively select the most representative feature channels to represent the target by calculating the energy mean ratio of the saliency-aware region to the search region, reducing the feature dimension and improving the tracking efficiency. Then, according to the feedback, the selected channels are given different weights to further enhance the discrimination of the filter. In addition, an adaptive update strategy is designed to alleviate the model degradation problem according to the fluctuation of feature maps in the recent frames. Finally, we use the alternating direction method of multipliers (ADMM) to optimize the proposed tracker model. Extensive experimental results on five well-known tracking benchmark datasets have verified the superiority of the proposed tracker with many state-of-the-art deep features-based trackers, and the running speed of the algorithm can basically meet the real-time requirements.

Keywords Correlation filters · Visual tracking · Saliency detection · Channel selection

1 Introduction

Visual object tracking has always been a critical task in computer vision, and it is the premise of many higher-level image processing tasks. Object tracking technology is to use the target and background information of the initial video frame to predict the position and scale of the target in the subsequent frames, which is widely used in video surveillance, intelligent transportation, intelligent medical, and other practical scenarios [1–3]. In recent years, the object tracking algorithm has achieved outstanding results in tracking performance. However, it is still challenging to accurately position the target in complex environments such as scale variation, illumination variation, and object occlusion.

The object tracking algorithms based on discriminative correlation filters (DCF) have received extensive attention due to their excellent tracking performance and efficient computational efficiency [4]. The characteristic of the DCF method is to collect samples by cyclic matrix and transform the correlation operation in the time domain into point-wise multiplication in the frequency domain by Fast Fourier Transform (FFT), which dramatically reduces the computational complexity and improves the speed of the algorithm. Most of the early correlation filter algorithms used handcrafted features such as histogram of oriented gradients (HOG) and color names (CN) to represent targets and showed favorable tracking results and excellent computational efficiency, reaching state-of-the-art at that time

✉ Zhixian Zhao
zhaozhixian1@126.com

[1] School of Computer Science and Technology, Shaanxi Key Laboratory of Network Data Analysis and Intelligent Processing,Xi'an University of Posts and Telecommunications, Xi'an 710121, China

[2] School of Information Engineering, Chang'an University, Xi'an 710064, China

[3] School of Operational Support, Rocket Force Engineering University, Xi'an 710025, China

[4] School of Automation, Northwestern Polytechnical University, Xi'an 710072, China

[5–7]. Convolutional neural network (CNN) features have a more robust representation power than handcrafted features, so many researchers have introduced multi-channel CNN features into the correlation filters framework [8–10].

It is indisputable that the success of recent DCF-based trackers is mainly due to the use of deep CNN features. As a result, researchers have proposed some methods to exploit the potential of deep features. Some trackers utilize principal component analysis (PCA) methods [8, 11] to reduce or compress the deep feature dimension but still need to address the high computational and memory costs required to extract deep features. Some algorithms improve tracking efficiency by using attention mechanisms [12] and assigning weights [13, 14] to deep feature channels. Nevertheless, the number of feature channels used by these algorithms is still significant, and the computational efficiency needs to be improved. Many researchers have recently introduced saliency detection into the correlation filter tracking framework and developed many advanced tracking algorithms with good results. However, most of these algorithms use image saliency information to either construct spatial or temporal regularization terms in the DCF model [15–17] to alleviate the boundary effect problem or to achieve reinforcement learning of the target appearance without paying attention to the correlation between multi-channel deep features and target saliency information [18, 19]. Recent studies [20, 21] have shown that different feature channels have different characteristics and contributions in the tracking process, especially multi-channel depth features. Deep features may contain many interference channels with irrelevant and redundant information to the target, and directly fusing all the hundreds and thousands of dimensional deep feature channels may produce severe overfitting, which leads to degradation of tracking performance.

Based on the above discussion, this paper proposes a new saliency-aware channel selection discriminative correlation filter (SCDCF) for robust visual tracking. The overall framework of the proposed tracker is shown in Fig. 1. SCDCF includes three stages: training, detection, and update. Firstly, we obtain the multi-channel deep features containing the energy of the target saliency-aware region through the saliency detection and feature extraction process. All channels are evaluated according to the proposed saliency-aware average energy ratio (SAER) indicator to obtain effective feature channels that pay more attention to the target information. Channels are given different weights according to their importance, and the final feature training filter is obtained to reduce the filter dimension and improve its discrimination power. Then, the selected feature channels and the trained filter perform correlation operations to obtain a response map to locate the target position. Finally, the proposed model updating mechanism is used for adaptive updating to avoid

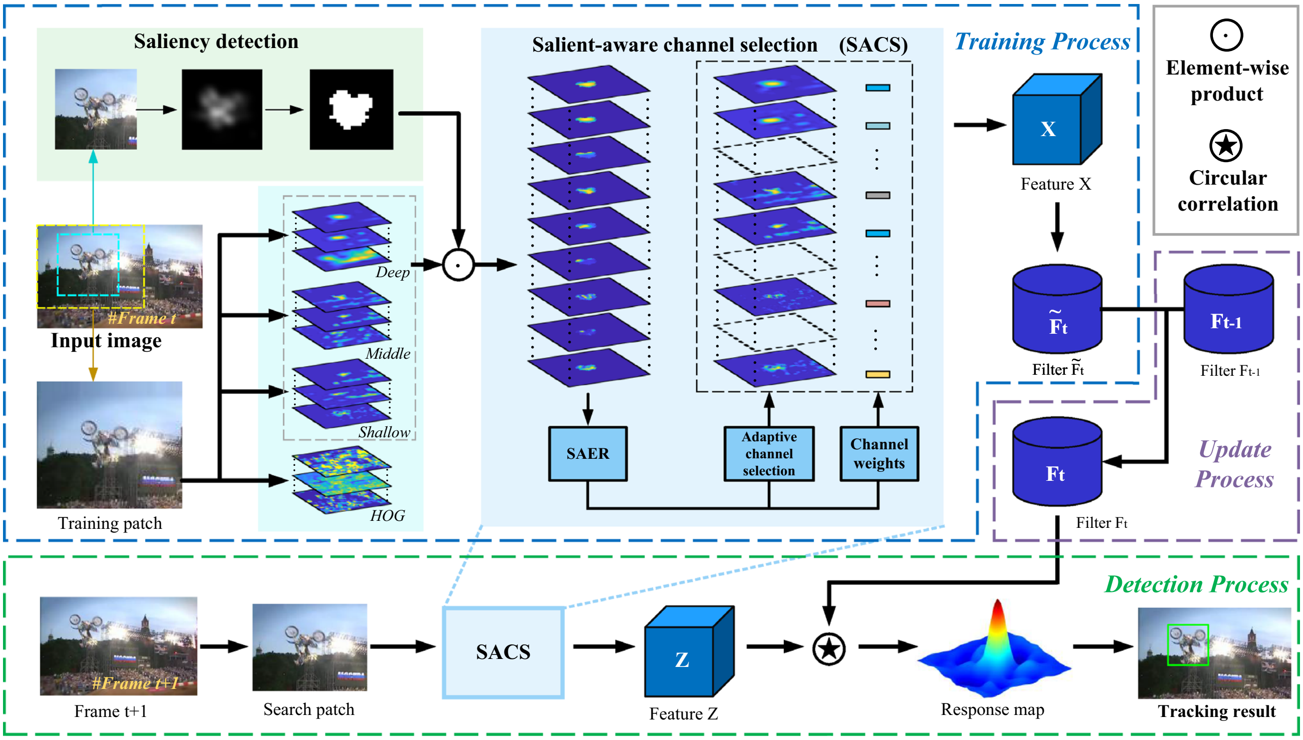


Fig. 1 The overall framework of SCDCF tracker. SCDCF includes three processes: training, detection, and update, which are marked by blue, green, and purple boxes, respectively

model degradation. The ADMM [22] algorithm is used to accelerate the solution of the proposed SCDCF model.

The main work of this paper is summarized as follows:

1. The saliency detection method is introduced to obtain the saliency information of the target. The feature energy of the saliency-aware region is calculated according to the target mask to highlight the target appearance and suppress the interference of the background information in the bounding box during the tracking process.
2. A new channel evaluation indicator is proposed to evaluate the importance of feature channels. Based on this, an adaptive channel selection mechanism is designed to select effective feature channels, reduce the feature dimension and enhance the discrimination ability of filters. According to the score, channel reliability is judged, and different weights are assigned to improve the representation ability of features.
3. An adaptive model updating mechanism is designed to judge the reliability of tracking results according to the fluctuation of the response map in the near time frame to ensure the accuracy of the target representation of the appearance model and alleviate the problem of model degradation.
4. The proposed trackers are evaluated on five public tracking datasets, including OTB2013 [23], OTB2015 [24], TC128 [25], UAV123 [26], and VOT2018 [27]. Experimental results show SCDCF is superior to many advanced trackers.

2 Related work

Early DCF algorithms mostly use handcrafted features to represent targets, such as color, texture, and edge features. The MOOSE [28] tracker, which initially introduced correlation filter theory into the field of object tracking, only used grayscale features to describe the target, and the tracking speed can reach hundreds of frames per second. Subsequently, Henriques et al. [29] incorporated multi-channel HOG features into the correlation filter framework and improved the algorithm accuracy by mapping linear space to high dimensional space through kernel functions. Danelljan et al. [30] extended the original RGB color space to 11 dimensions, trained correlation filters using color names (CN) features containing rich color information. Many subsequent trackers [5, 7] utilize complementary handcrafted features to describe the target to enhance the feature representation power. Recently, due to the excellent performance of handcrafted features in computational efficiency and accuracy, DCF trackers based on handcrafted features have shown significant advantages in aerial target scenarios and are widely used in unmanned aerial vehicles (UAV) platforms [15, 31, 32].

Deep features show strong representation ability with the rapid development of neural networks. Many trackers use multi-channel convolutional features extracted by deep neural network models to represent targets. Ma et al. [10] used the VGG-19 network to extract the multi-layer convolution features of the target and achieved precise positioning of the target according to the characteristics of different layers of features. Danelljan et al. [9] proposed the DeepSRDCF tracker based on spatially regularized discriminative correlation filters (SRDCF [33]) combined with convolutional features for modeling. The C-COT [34] tracker used deep neural network to extract features, obtained feature maps of continuous spatial domains by interpolation operations, and applied Hessian matrices to achieve sub-pixel accuracy localization of target positions. Noting the interfering channels and running speed problems caused by multi-channel deep features, many tracking algorithms use attention mechanisms, feature compression, and other methods to alleviate them to achieve robust and fast tracking.

Saliency detection is to simulate the human visual attention mechanism to detect the most interesting and visually expressive areas in the image. It is widely used in visual tasks such as object detection, semantic segmentation, and image caption. Many recent works have applied it to object tracking and achieved good results. For example, some trackers introduce image saliency detection into the regularization term of the DCF formula to alleviate the boundary effect problem. According to the characteristics of aerial object tracking, Fu et al. [15] used the dual regularization strategy to construct the target saliency regularization model to achieve accurate real-time tracking of aerial objects. Feng et al. [16] integrated saliency information and target change information into the spatial weight map and proposed a dynamic saliency spatial regularization correlation filter method. Yang et al. [17] introduced two saliency information extraction methods in the regularization process and proposed co-saliency spatio-temporal regularization correlation filters. In addition, some researchers have also used saliency information to highlight image saliency regions for reinforcing learning of target appearance [18, 19].

Although the tracking performance of the above DCF trackers has been improved, the correlation between target saliency information and feature channel information is not considered. Therefore, we investigate the relationship between target saliency region information and multi-channel deep features and propose a new channel selection method based on image saliency information. By combining saliency detection with feature channel selection, we can accurately highlight the target region and suppress the interference of background information in the target tracking frame, reduce the dimension of feature channels and improve the discriminant ability of the filter.

3 Proposed method

In this section, we first briefly review the discriminative correlation filters and describe the saliency-aware detection mechanism used. Then we propose an adaptive channel selection method and our SCDCF model and use the ADMM method for optimization. Finally, we develop a new model update strategy.

3.1 Revisit of DCF

Given the initial target position in the first frame, the task of object tracking is to estimate the target position in subsequent frames. To locate the target position in the $(t+1)$th frame, DCF uses the training sample $\{X_t, Y\}$ of the t-th frame to learn the multi-channel correlation filter, where $X_t \in \mathbb{R}^{W\times H\times C}$ is defined as a C-dimensional channel feature with width W and height H, and Y is the expected response map of the corresponding Gaussian shape. To obtain a multi-channel correlation filter, DCF expresses the objective as a regularized least squares problem:

$$\tilde{F}_t = \underset{F_t}{\arg\min} \left\| \sum_{i=1}^{C} F_t^i \otimes X_t^i - Y \right\|_2^2 + \lambda \sum_{i=1}^{C} \left\| F_t^i \right\|_2^2, \tag{1}$$

where $\otimes$ denotes circular correlation operator, $X_t^i \in \mathbb{R}^{W\times H}$ and $F_t^i \in \mathbb{R}^{W\times H}$ represent the i-th channel of X_t and F_t, $\lambda \sum_{i=1}^{C} \|F_t^i\|_2^2$ is a regularization term, and λ is the regularization parameter. The task can be transformed into the Fourier domain to derive the closed-form solution of Eq. 1 as follows:

$$\hat{F}_t^i = \frac{\left(\hat{X}_t^i\right)^* \odot \hat{Y}}{\left(\hat{X}_t^i\right)^* \odot \hat{X}_t^i + \lambda}, \tag{2}$$

where $\hat{\cdot}$ stands for the Discrete Fourier Transform (DFT), $\cdot^*$ indicates the complex conjugate operator, and $\odot$ represents the element-wise product operator.

According to the feature vector $Z \in \mathbb{R}^{W\times H\times C}$ extracted from the candidate images of $(t+1)$ frame, the response map $R \in \mathbb{R}^{W\times H}$ can be obtained by the following equation:

$$R = \mathcal{F}^{-1}\left(\sum_{i=1}^{C} \hat{Z}^i \odot \hat{F}_t^i \right), \tag{3}$$

where $\mathcal{F}^{-1}$ denotes the inverse DFT. The target position in $(t+1)$ frame is determined by the peak position in response map R.

3.2 Background-aware correlation filter

The overall objective function of Background-Aware Correlation Filter (BACF) can be expressed as:

$$\underset{F}{\arg\min} \frac{1}{2} \left\| \sum_{i=1}^{C} X^i \otimes \left(P^{\top} F^i\right) - Y \right\|_2^2 + \frac{\lambda}{2} \sum_{i=1}^{C} \left\| F^i \right\|_2^2 \tag{4}$$

where $X^i \in \mathbb{R}^T$ (T is the number of X pixels), P is a binary matrix that is used to crop N ($N << T$) elements in feature samples X, and $P^{\top}$ is the conjugate transpose of P.

The traditional correlation filter algorithm performs the cyclic shift operation on the positive sample extracted from the image target to obtain negative samples to train the filter. It does not model the real background information, which may lead to boundary effects and model drift problems. The handcrafted feature-based BACF uses a clipping matrix to crop negative samples from real background information to train filters, significantly improving the sample quality and quantity. Unfortunately, BACF uses handcrafted features to represent the target and treats all spatial feature channels equally, which cannot accurately identify the appearance changes of the target. In addition, BACF also expands the search area to deal with fast tracking problems, but it also introduces more background interference, which limits the improvement of algorithm performance. Therefore, we introduce multi-channel deep features into the BACF framework to improve the accuracy of target appearance modeling and use saliency-aware detection and channel selection mechanisms to reduce the interference of background clutter during tracking.

3.3 Saliency-aware detection mechanism

The existing advanced feature channel selection methods [20, 35] filter the channels according to the feature response in the rectangular target box. These methods improve the quality of the used feature channels to a certain extent but still introduce some background information to interfere with the learning of the filter. This paper aims to calculate the energy more suitable for the target appearance contour area for channel selection. Therefore, we introduce the saliency detection [36] and design a saliency-aware detection mechanism. As shown in Fig. 2, firstly, according to the target region bounding box, i.e., red box, the region near the target is selected as the saliency detection region, i.e., blue box. Then the blue box region is detected to obtain the saliency map, and the target region mask is generated after threshold mapping. The generated mask can be used to segment the target and surrounding region robustly. Combined

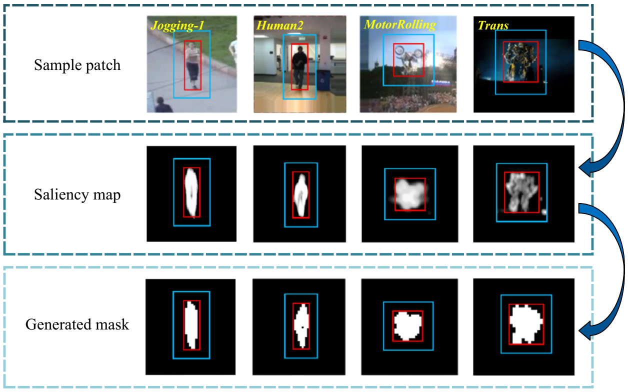


Fig. 2 Visualization of the mask generation process. From top to bottom, the images denote sample patch, saliency map, and generated mask. From left to right, the four sequences from OTB2015 dataset are *Jogging-1*, *Human2*, *MotorRolling*, and *Trans* respectively

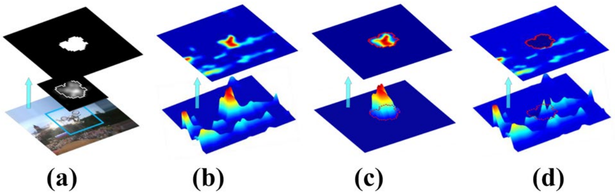


Fig. 3 **a** Schematic diagram of the process of mask generation. **b** Feature map of search region. **c** Feature map of saliency-aware region. **d** Feature map of background region

with the extracted search area features, the multi-channel saliency-aware region features that are more focused on the target saliency-aware region are obtained, which can significantly shield the noise caused by the background information in the tracking box. We use this characteristic of saliency detection to calculate the target saliency region (as shown in Fig. 3(a)). According to the feature map extracted from the search region (as shown in Fig. 3(b)), we can obtain the saliency-aware region feature map (as shown in Fig. 3(c)) and the background region feature map (as shown in Fig. 3(d)) of each channel.

3.4 Adaptive feature channels selection

The achievements of deep DCF trackers in recent years are largely attributed to the use of multi-channel convolutional features, but due to the limited number of training samples for visual tracking, the deep networks used to extract convolutional features are often pre-trained in other computer vision tasks, such as VGGNet or MobileNet, which are trained on ImageNet [37]. Using the deep network trained by general targets to extract multi-channel features of specific targets, hundreds of channels may contain a large number of interference channels, which may not contain target area information or contain more background information, affecting the learning of correlation filters. Figure 4 shows the difference between the efficient channels and the interfering channels. Since the DCF tracker obtains the response map by extracting the search area features according to the target position in the

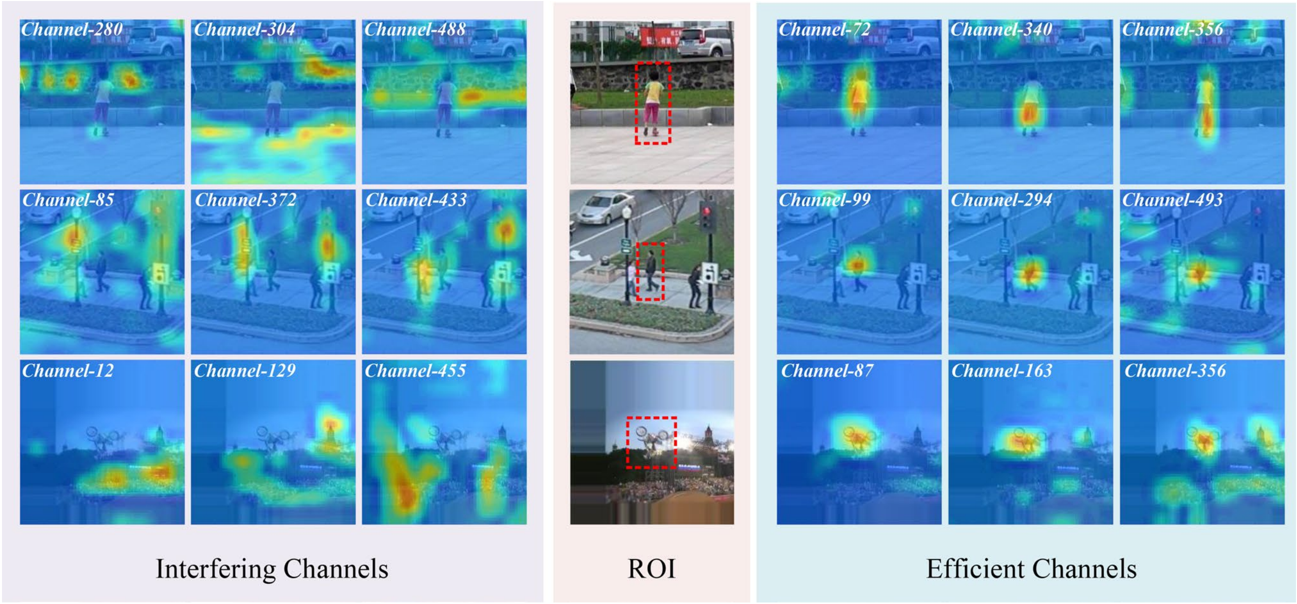


Fig. 4 Visualization of feature maps for different channels of sequences. From top to bottom, the three sequences from OTB2015 dataset are *Girl2*, *Human3*, and *MotorRolling*, respectively

previous frame, the feature channels that are beneficial for tracking should focus more on the energy of the target area, containing larger target area energy and smaller energy of other search areas.

Combining the analysis in Sect. 3.3, we propose a new feature channel evaluation indicator. As shown in Fig. 3c and 3d, we divide the feature map after saliency detection into the target saliency-aware region feature map and the background region feature map. The feature channel is evaluated by calculating the average energy ratio of these two parts. The proposed SAER indicator is defined as follows:

$$SAER(i) = \frac{E_O(X^i)}{E_B(X^i)}, i = 1, 2, \cdots, C, \tag{5}$$

where X^i denotes the ith channel of feature $X \in \mathbb{R}^{W\times H\times C}$. We define $E_O(X^i)$ as the average energy value of the target saliency-aware region O:

$$E_O(X^i) = \frac{\sum_{(p,q)\in O} V(p,q)}{Area(O)}, \tag{6}$$

where $V(p, q)$ is defined as the feature energy value of position (p, q), $Area(O)$ represents the area of region O. Similarly, $E_B(X^i)$ is defined as the average energy value of background region:

$$E_B(X^i) = \frac{\sum_{(p,q)\in S} V(p,q) - \sum_{(p,q)\in O} V(p,q)}{Area(S) - Area(O)}, \tag{7}$$

where S denotes the search region. We judge the confidence of the feature channel according to the SAER index. The higher the SAER score, the richer the target information contained in the channel, and the smaller the SAER score, indicating that the channel contains more background interference. Therefore, we calculate the SAER scores for all channels and adaptively select channels with scores higher than a given threshold for filter learning to reduce the interference of invalid feature channels.

On the other hand, in recent years, the channel attention mechanism has been widely used in computer vision tasks. It judges the importance of feature channels by modeling them and assigns greater weights to more important channels to enhance the discrimination of filters. Therefore, we combine the idea of channel attention with the proposed saliency-aware channel selection mechanism, use SAER to judge the importance of the channel, and assign different weights to the selected channels so as to improve tracking efficiency and alleviate the shortcomings of channel attention. After the salience-aware channel selection, the effective feature with higher discriminative power and the score sequence A containing SAER scores of each channel are obtained, then the weight w^i of the i-th channel can be expressed as:

$$w^i = 1 + \frac{1}{2} \times \frac{A(i) - \min(A)}{\max(A) - \min(A)}. \tag{8}$$

3.5 Modeling and optimization of the SCDCF

Using the proposed feature channel selection method, we can obtain the effective feature X_E that focus more on the target information and the corresponding weight sequences w. Therefore, the proposed SCDCF model can be expressed as:

$$\arg\min_F \frac{1}{2}\left\|\sum_{i=1}^{D} w^i X_E^i \otimes \left(P^\top F^i\right) - Y\right\|_2^2 + \frac{\lambda}{2}\sum_{i=1}^{D}\left\|F^i\right\|_2^2, \tag{9}$$

where X_E^i and w^i denote the i-th feature channel of the effective feature X_E and its weight. After channel selection, the number of channels is reduced from C to D.

To improve the computational efficiency, we use X_S to represent the final feature used to train the filter in the optimization process, that is, $X_S^i = w_i \times X_E^i$, and X_S^i represents the i-th feature channel of feature X_S. The conversion of Eq. 9 to the frequency domain can be expressed as:

$$\begin{aligned} &\arg\min_{\hat{F},\hat{G}} \frac{1}{2}\left\|\hat{X}_S \odot \hat{G} - Y\right\|_2^2 + \frac{\lambda}{2}\|F\|_2^2 \\ &\qquad s.t., \ \hat{G} = \sqrt{T}HP^\top F, \end{aligned} \tag{10}$$

where $\hat{G} = [\hat{G}^1, \hat{G}^2, \dots, \hat{G}^D]$ is an auxiliary variable matrix, H is the orthonormal $T \times T$ matrix of complex basis vectors for mapping any T-dimensional vector to the Fourier domain (e.g., $\hat{g} = \sqrt{T}Hg$). We employ the augmented Lagrangian method to optimize Eq. 10:

$$\begin{aligned} \mathcal{L}(F, \hat{G}, \hat{\vartheta}) =& \frac{1}{2}\left\|\hat{X}_S \odot \hat{G} - \hat{Y}\right\|_2^2 + \frac{\lambda}{2}\|F\|_2^2 \\ &+ \hat{\vartheta}(\hat{G} - \sqrt{T}HP^\top F)^\top \\ &+ \frac{\eta}{2}\left\|\hat{G} - \sqrt{T}HP^\top F\right\|_2^2, \end{aligned} \tag{11}$$

where $\hat{\vartheta} = [\hat{\vartheta}^1, \hat{\vartheta}^2, \dots, \hat{\vartheta}^D]^\top \in \mathbb{R}^{T\times D}$ denotes the Lagrangian multiplier and η is the penalty parameter.

The ADMM algorithm can be applied to Eq. 11 to split it into three independent subproblems, each of which has a closed solution:

$$\begin{cases} F = \frac{\lambda}{2}\|F\|_2^2 + \hat{\vartheta}\left(\hat{G} - \sqrt{T}HP^{\top}F\right)^{\top} \\ \quad + \frac{\eta}{2}\left\|\hat{G} - \sqrt{T}HP^{\top}F\right\|_2^2 \\ \hat{G} = \frac{1}{2}\left\|\hat{X}_s \odot \hat{G} - \hat{Y}\right\|_2^2 + \hat{\vartheta}\left(\hat{G} - \sqrt{T}HP^{\top}F\right)^{\top} \\ \quad + \frac{\eta}{2}\left\|\hat{G} - \sqrt{T}HP^{\top}F\right\|_2^2 \\ \hat{\vartheta}^{l+1} = \hat{\vartheta}^{l} + \eta\left(\hat{G}^{l+1} - F^{l+1}\right) \end{cases} . \tag{12}$$

Then, the individual subproblems are solved iteratively as follows:

Subproblem F: The optimal solution can be easily obtained as follows:

$$F_{opt} = \frac{\vartheta + \eta G}{\eta + \lambda/T}, \tag{13}$$

where $G = \frac{1}{\sqrt{T}}HP^{\top}\hat{G}$ and $\eta = \frac{1}{\sqrt{T}}HP^{\top}\hat{\eta}$. F_{opt} is obtained by Inverse Fast Fourier Transform(IFFT) of $\hat{G}$ and $\hat{\eta}$.

Subproblem $\hat{G}$: For $\hat{G}$, since each pixel is independent, it can be decomposed into T small subproblems. The closed solution can be obtained as follows:

$$\begin{aligned} \hat{G}(k)_{opt} =& \frac{1}{\eta}\left(\frac{1}{T}\hat{X}_S(k)\hat{Y}(k) + \eta\hat{F}(k) - \hat{\vartheta}(k)\right) \\ & - \frac{\hat{X}_S(k)}{\eta b}\left(\frac{1}{T}\hat{U}_X(k)\hat{Y}(k) + \eta\hat{U}_F(k) - \hat{U}_{\vartheta}(k)\right), \end{aligned} \tag{14}$$

where $\hat{U}_X(k) = \hat{X}_S(k)^{\top}\hat{X}_S(k)$, $\hat{U}_F(k) = \hat{X}_S(k)^{\top}\hat{F}(k)$, $\hat{U}_{\vartheta}(k) = \hat{X}_S(k)^{\top}\hat{\vartheta}(k)$ and $b = \hat{U}_X(k) + \eta T$.

Updating other variables: The Lagrange multiplier $\hat{\vartheta}$ and penalty parameter η are updated as:

$$\begin{cases} \hat{\vartheta}^{l+1} = \hat{\vartheta}^{l} + \eta(\hat{G}^{l+1} - F^{l+1}) \\ \eta^{l} = \min(\eta_{\max}, \delta\eta^{l}) \end{cases}, \tag{15}$$

where l represents the number of iterations and δ is the scale factor.

3.6 Adaptive model Update

The traditional DCF algorithm uses linear interpolation to update the filter for each frame. This strategy of updating each frame can slowly learn the latest changes of the target by combining historical and current information. However, if the model continues to be updated under complicated situations such as severe target occlusion may introduce a large amount of interference information that is detrimental to the tracking process, resulting in model drift. To address these issues, researchers have developed two confidence indicators: APCE (Average Peak to Correlation Energy) [38] and PSR (Peak-to-Sidelobe Ratio) [28], which are used to analyze the similarity and peak intensity of the response map. Inspired by APCE and PSR, SCDCF uses the proposed RFM (response map fluctuation) metric to determine the fluctuation degree of the response map and sets the conditions for model updating according to the feedback.

$$RMF = \frac{R_{\max} - R_{\min}}{\sqrt{\frac{1}{W\times H}\left(\sum_{i,j}^{W,H}\left(R_{i,j} - R_{mean}\right)^2\right)}}, \tag{16}$$

where $R_{i,j}$ is defined as the response value of position (i, j), $R_{\max}$, $R_{\min}$, and R_{mean} are the maximum, minimum, and average values in the response map $R \in \mathbb{R}^{W\times H}$. Then the average RFM value $RMF_{mean} = (1/n)\sum_1^n RMF_n$ of nearly n frames is obtained, and whether to update is judged by comparing the current frame score RFM with RMF_{mean}. Therefore, the model update for SCDCF can be expressed as:

$$\begin{cases} \text{Update,} & RMF_t > \varphi RMF_{mean} \\ \text{Noupdate,} & RMF_t \le \varphi RMF_{mean} \end{cases}, \tag{17}$$

where φ is the ratio factor.

4 Experimental results

4.1 Implementation details

Platform: The proposed tracker is implemented in MATLAB 2018a on a PC with an Intel(R) Xeon(R) Gold 6136CPU at 3.00GHz, 512 G RAM and a single NVIDIA TITAN V GPU. The MatConvNet [39] toolbox is used to extract deep features from pre-trained CNN networks.

Parameters: To guarantee the fairness and objectivity of the evaluation, we follow some key parameters in the standard DCF method [7, 11] to construct tracker. For target localization, we use HOG features and shallow layer (conv3-4), middle layer (conv4-3), and deep layer (conv5-1) features of the VGG-16 network to represent the target. We set the learning rate σ=0.0135, and the SAER threshold in Sect. 3.4 is 1.37. For model optimization, we set the number of iterations l of ADMM to be 2, the penalty parameter η to be 1, and the $\eta_{\max}$ and δ in Eq. 15 to be 10^4 and 10, respectively. For model updating, We set the ratio factor φ=0.65 in Eq. 17, refer to [40] to set the number of recent frames n=5. In addition, for some parameters of scale estimation, we refer to the ASRCF [11] tracker, and the remaining parameters are consistent with the BACF [7] tracker.

4.2 Experiment datasets and evaluation metrics

We evaluate the effectiveness of the proposed tracking method on five public tracking datasets, including OTB2013 [23], OTB2015 [24], TC128 [25], UAV123 [26], and VOT2018 [27]. For the OTB2013, OTB2015, TC128, and UAV123 datasets, we use the precision and success plots of the one-pass evaluation (OPE) strategy to measure the performance of the tracker. The precision plot reports the proportion of video frames whose distance between the bounding box predicted by the tracker and the manually labeled actual bounding box is less than a certain threshold. The success plot reports the proportion of video frames whose overlap rate is greater than a certain threshold between the predicted bounding box and the real bounding box. We use the distance precision (DP) with a threshold of 20 pixels in the precision plot and the area under the curve (AUC) of the success plot to evaluate the tracker. The overlap precision (OP) is the corresponding score of the success plot when the overlap rate threshold is set to 0.5. In addition, the center position error (CLE) measures the average Euclidean distance between the center position of the predicted bounding box and the real bounding box, and the speed of the tracker is shown in frames per second (FPS). For the VOT2018 dataset, we analyze the tracker performance using three metrics, expected average overlap (EAO), Accuracy, and Robustness.

4.3 Qualitative valuation

We select 10 representative sequences with different challenge attributes from the OTB2015 dataset for qualitative evaluation of our tracker, and the results are shown in Fig. 5. Comparison algorithms include DCF trackers based on deep features (i.e., DeepSTRCF [41], DeepSRDCF [9], C-COT [34], and HCF [10]) and DCF trackers based on handcrafted features (i.e., MCCT-H [40], ECO-HC [8], BACF [7], Staple [5], and KCF [29]). When the target is severely disturbed by the surrounding background (i.e., *Box*, *Matrix*), our approach performs well due to the saliency-aware detection mechanism that shields the background noise and makes the learned filter more focused on the target information. When the target is deformed and rotated (i.e., *MotorRolling*, *Dragonbaby*, *Jump*, *Sylvester*), SCDCF achieves accurate tracking in continuous frames because it uses the channel selection strategy to remove a large number of redundant channels and uses the channel adaptive weighting method to improve the representation of the features used effectively. Especially in the *Jump* sequence, which is more difficult to track, only the SCDCF tracker successfully tracks the target in several

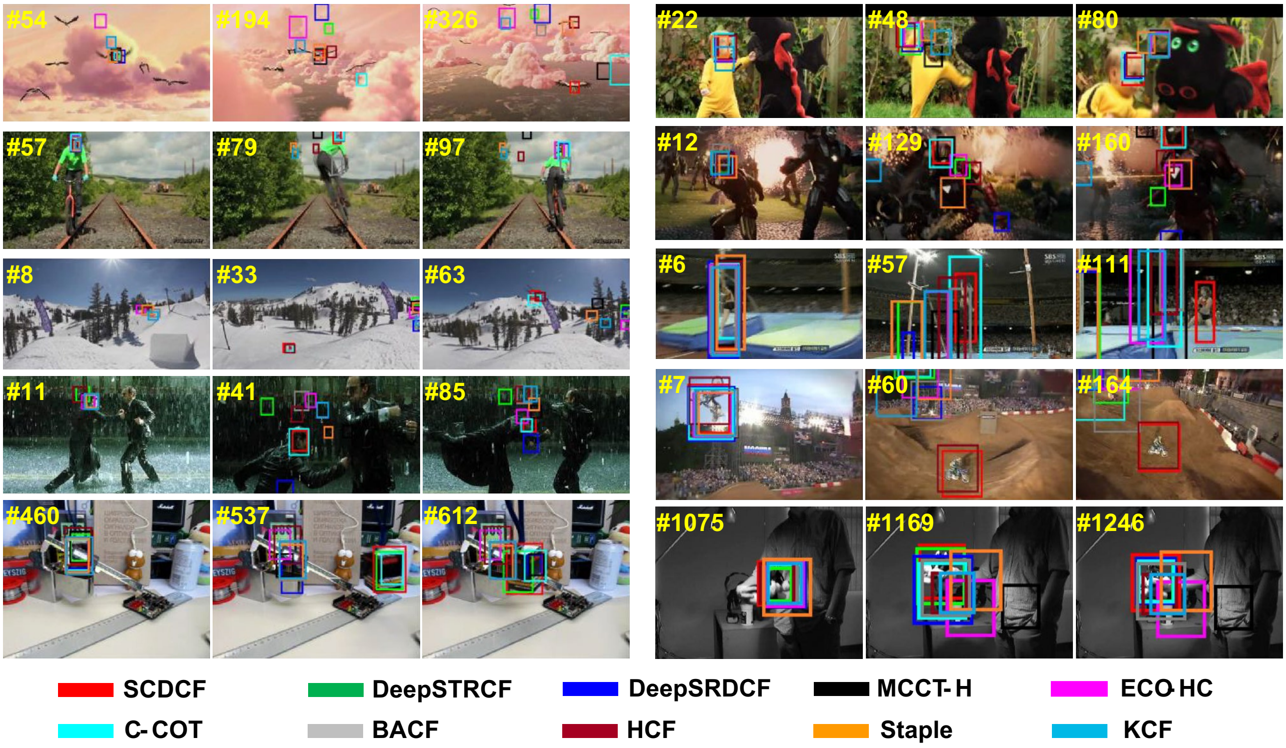


Fig. 5 Qualitative evaluation results of the proposed tracker and other advanced trackers for 10 challenge sequences from the OTB2015 benchmark. From top to bottom and from left to right, these sequences are *Bird1*, *Biker*, *Skiing*, *Matrix*, *Box*, *Dragonbaby*, *Ironman*, *Jump*, *MotorRolling* and *Sylvester*, respectively

algorithms. Similarly, because we use an adaptive update strategy to avoid unnecessary model updates, SCDCF also succeeds when the target is out of view (i.e., *Bird1*, *Biker*). In addition, Our tracker also performs well in terms of fast motion and illumination variation (i.e., *Skiing*, *Ironman*). The qualitative evaluation results show that the proposed SCDCF method is superior to many advanced tracking algorithms in various complex situations.

4.4 Quantitative evaluation

OTB: Fig. 6 shows the precision and success plots of our method and other 16 trackers on the OTB2013 and OTB2015 datasets, including handcrafted features-based DCF trackers (i.e., ECO-HC [8], LADCF-HC [42], BACF [7], CACF [6], SRDCF [33], SAMF [43], KCF [29]), deep features-based DCF trackers (i.e., MCCT [40], DeepSTRCF [41], C-COT [34], DeepSRDCF [9], MCPF [44], HDT [45], HCF [10]), and deep learning-based trackers (i.e., SiamFC [46], SiamRPN [47]). Overall, our SCDCF tracker is superior to many advanced tracking algorithms in terms of DP and AUC scores. On OTB 2013, our method ranks first with a DP of 94.2% and an AUC of 71.6%. On OTB2015, SCDCF has the highest DP and AUC scores of 92.2% and 68.3%, respectively, which are 4.2%/0.8% and 7.1%/4.8% higher than DeepSTRCF and DeepSRDCF, which use multi-channel deep features, and 6.1%/2.3% higher than LADCFHC, which is the best performer among handcrafted feature based trackers. In addition, this work comprehensively compares SCDCF with other 9 deep learning-based trackers, including LUDT+ [48], LUDT [48], PrDiMP-18 [49], ROAM [50], ROAM+ [50], ATOM [51], GradNet [52], DiSiamRPN [53],

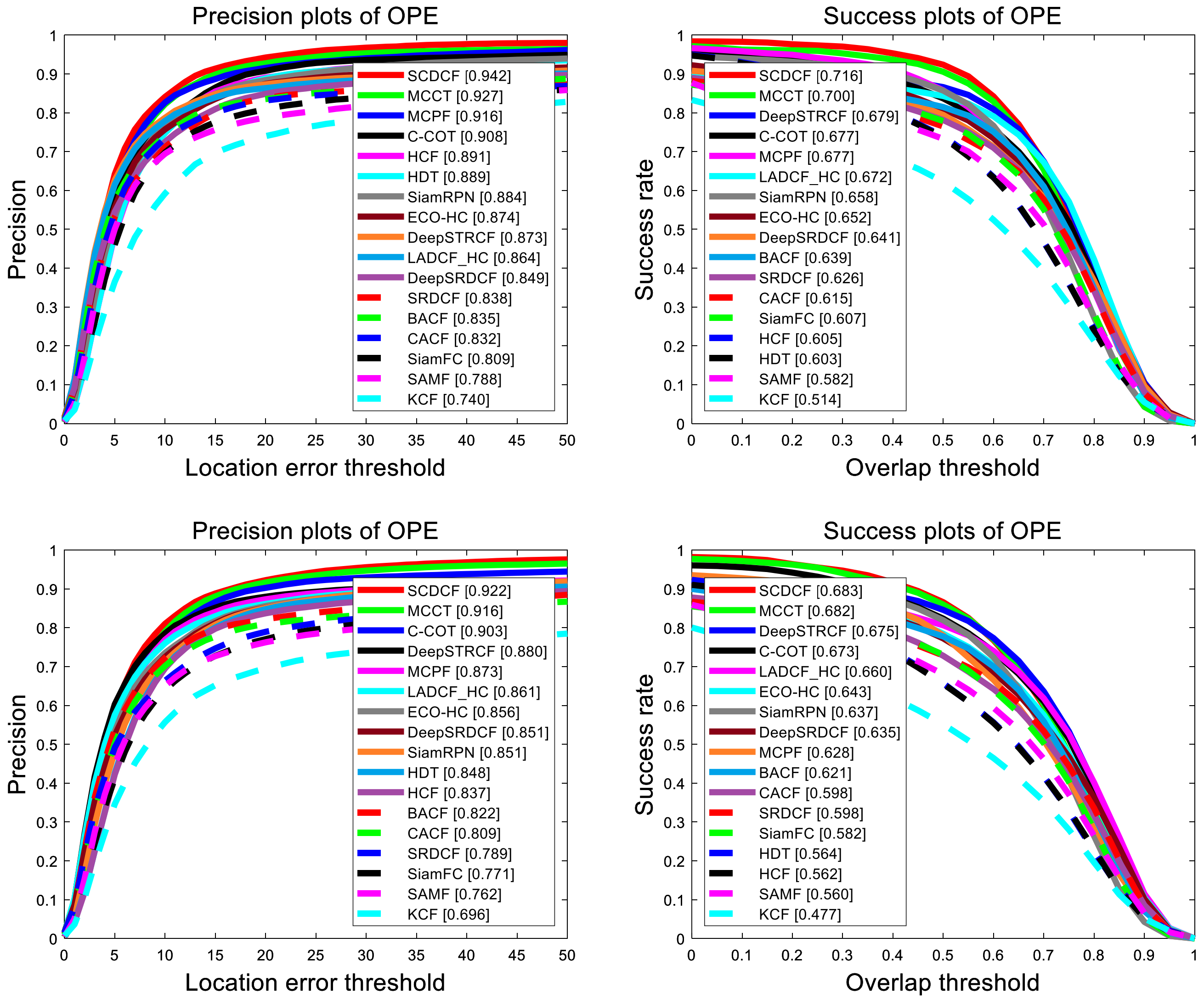


Fig. 6 Precision and success plots of SCDCF and other state-of-the-art trackers on OTB2013 (first row) and OTB2015 (second row), with AUC and DP scores reported in the figure legend

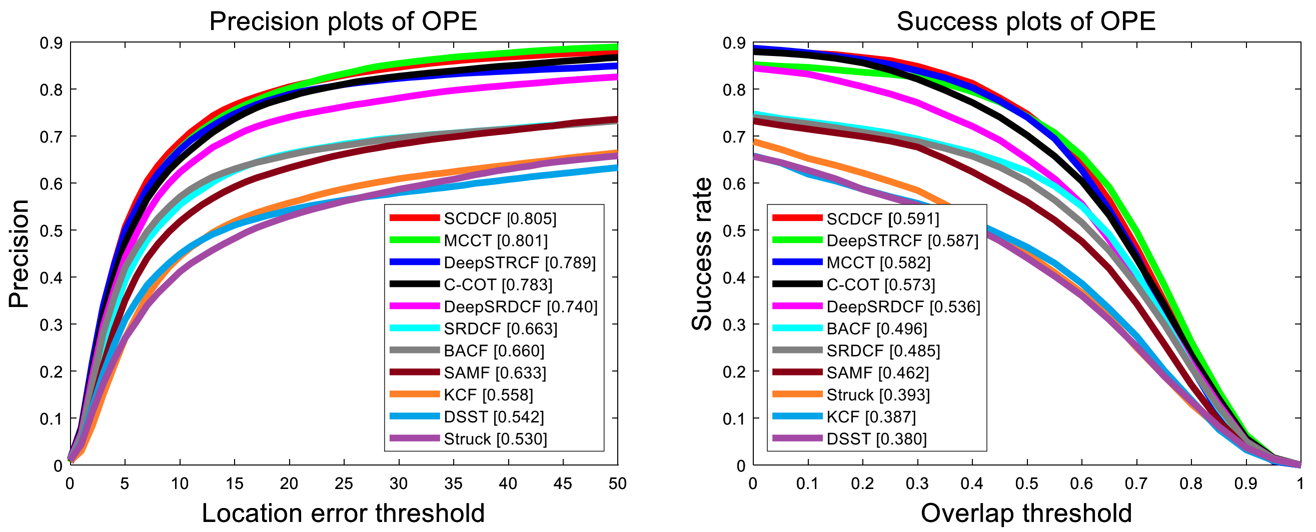


Fig. 7 Precision and success plots of SCDCF and other trackers on TC128, with AUC and DP scores reported in the figure legend

Table 1 Performance comparison with other SOTA trackers on OTB2015

Tracker	Venue	Prec.	Succ.	Tracker	Venue	Prec.	Succ.
SCDCF	This work	0.922	0.683	SiamRPN	2018’CVPR	0.851	0.637
LUDT+	2021’IJCV	0.843	0.639	ECO-HC	2017’CVPR	0.856	0.643
LUDT	2021’IJCV	0.769	0.602	MCPF	2017’CVPR	0.873	0.628
PrDiMP-18	2020’CVPR	0.875	0.681	CACF	2017’CVPR	0.809	0.598
ROAM++	2020’CVPR	0.904	0.680	BACF	2017’ICCV	0.822	0.621
ROAM	2020’CVPR	0.908	0.681	C-COT	2016’ECCV	0.903	0.673
ATOM	2019’CVPR	0.876	0.662	SiamFC	2016’ECCVW	0.771	0.582
GradNet	2019’ICCV	0.861	0.639	HDT	2016’CVPR	0.848	0.564
DiMP-18	2019’ICCV	0.856	0.658	DeepSRDCF	2015’ICCVW	0.851	0.635
LADCF-HC	2019’TIP	0.861	0.660	SRDCF	2015’ICCV	0.789	0.598
DaSiamRPN	2018’ECCV	0.858	0.644	HCF	2015’ICCV	0.837	0.562
MCCT	2018’CVPR	0.916	0.682	KCF	2015’TPAMI	0.696	0.477
DeepSTRCF	2018’CVPR	0.880	0.675	SAMF	2014’ECCV	0.762	0.560

Table 2 A comparison of our SCDCF method with 16 advanced trackers on OTB2015

Tracker	SiamRPN	SiamFC	SAMF	HCF	HDT	SRDCF	CACF	BACF
OP(%)	81.6	73.1	68.0	65.6	65.7	72.9	72.8	77.8
CLE(pixels)	19.6	33.2	33.9	22.8	20.1	38.6	31.5	26.5
Speed(fps)	34.2	84.3	33.7	10.4	5.5	4.3	37.6	37.1
Tracker	MCPF	DeepSRDCF	ECO-HC	LADCF-HC	C-COT	DeepSTRCF	MCCT	**SCDCF**
OP(%)	78.1	77.3	78.5	80.6	82.4	84.6	86.0	86.6
CLE(pixels)	20.9	21.4	22.7	20.6	14.0	17.8	10.7	8.4
Speed(fps)	0.5	0.3	24.6	20.0	0.2	6.4	7.5	14.8

and DiMP-18 [54], on the OTB2015 dataset to present a more comprehensive evaluation. The results are shown in Table 1. Our SCDCF tracker achieves the best precision and success rate, outperforming the recent SOTA trackers.

To analyze the performance of the tracker in more detail, we evaluate the mean CLE, OP, and speed (fps) of SCDCF on OTB2015. Table 2 reports the comparison results of our method with 15 other trackers. In terms of OP, SCDCF achieves the best performance with 0.6%/2.0% improvement over the second and third places (i.e., MCCT and DeepSTRCF). In terms of mean CLE, SCDCF maintains at 8.4 pixels, outperforming many state-of-the-art trackers. In terms of speed, the end-to-end Siamese network-based tracking algorithms (i.e., SiamFC and SiamRPN) are faster, reaching 84.3 fps and 34.2 fps, respectively. Due to their offline learning method, the tracking speed is fast, but the target appearance model cannot be dynamically adjusted by analyzing the context environment. The tracking effect is poor compared with the SCDCF using the online learning method. Among the many deep feature-based correlation filter trackers, SCDCF is the fastest at 14.8fps, 4.4fps faster than the second-place HCF algorithm. This is because SCDCF uses channel selection to remove a large number of interfering feature channels to improve tracking efficiency. Overall, the tracking performance of our method is superior to the other advanced trackers.

TC128: We compare the proposed tracker with 10 advanced trackers on TC128, such as MCCT [40], DeepSRDCF [9], DeepSTRCF [41], C-COT [34], SRDCF [33], BACF [7], SAMF [43], Struck [55], DSST [56], and KCF [29]. The evaluation results are shown in Fig. 7, and our method achieves the best DP/AUC scores. In terms of DP, our tracker scores the highest with 80.5%, which is 0.4% and 1.6% better than the second and third places (i.e., MCCT and DeepSTRCF). In terms of AUC, SCDCF ranks first, outperforming C-COT and DeepSRDCF by 1.8% and 5.5%.

Table 3 A comparison of our SCDCF method with 9 advanced trackers on UAV123

Tracker	Venue	Prec.	Succ.
SCDCF	This work	0.770	0.520
ACSDCF	2021'IJCV	0.765	0.536
AutoTrack	2020'CVPR	0.689	0.472
fECO	2020'TIP	0.737	0.511
fDeepSTRCF	2020'TIP	0.721	0.508
ASRCF	2019'CVPR	0.737	0.508
MCCT	2018'CVPR	0.734	0.507
DeepSTRCF	2018'CVPR	0.728	0.513
ECO	2017'CVPR	0.741	0.525
ECO-HC	2017'CVPR	0.710	0.496

UAV123: The UAV123 is one of the most popular datasets in the field of UAV object tracking. We compare SCDCF with 23 recent trackers on UAV123. As shown in Fig. 8, the overall performance of SCDCF is excellent. For more clarity, we also show the DP/AUC scores of the ten best-performing trackers in Table 3. It can be seen from the

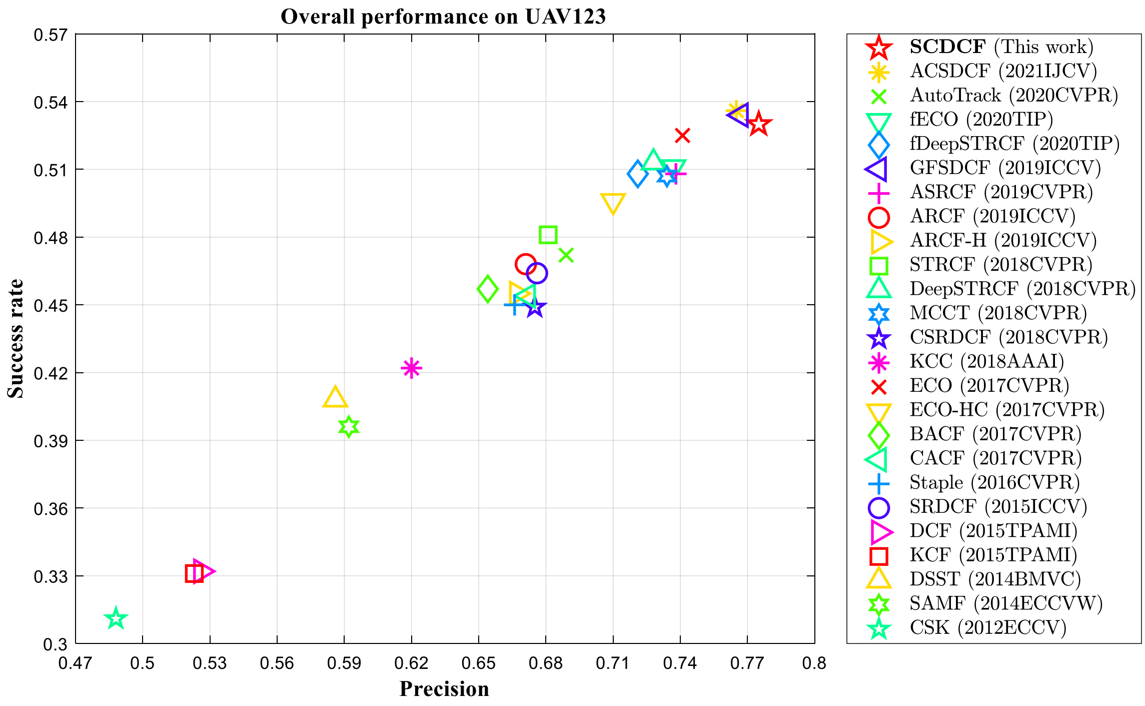


Fig. 8 Precision and success plots of SCDCF and other trackers on UAV123, with AUC and DP scores reported in the figure legend

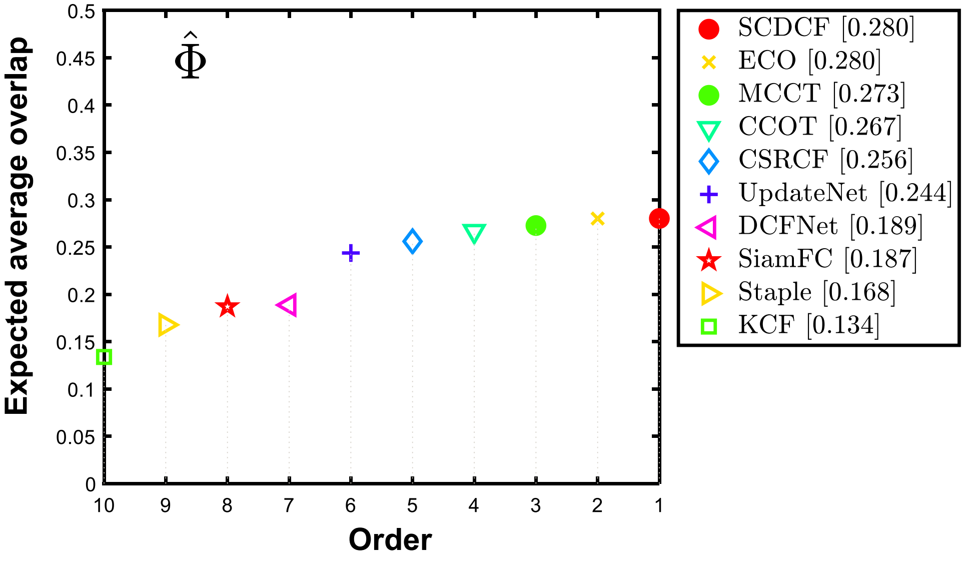


Fig. 9 Expected average overlap (EAO) ranking plots on the VOT2018 dataset

Table 4 A comparison of our SCDCF method with 9 advanced trackers on VOT2018

Tracker	KCF	Staple	SiamFC	DCFNet	UpdateNet	CSRCF	C-COT	ECO	MCCT	SCDCF
EAO	0.134	0.168	0.187	0.189	0.244	0.256	0.267	0.280	0.273	0.280
Accuracy	0.448	0.526	0.500	0.469	0.519	0.491	0.493	0.483	0.530	0.492
Robustness	0.781	0.684	0.585	0.518	0.454	0.356	0.318	0.276	0.318	0.295

table that SCDCF scored 77.0% on the DP index, ranking first among the 23 trackers, leading the second and third (i.e., ACSDCF and ECO) by 0.5% and 2.9%. Likewise, SCDCF scored 52.0% on the AUC index, ranking third. Experimental results show that the SCDCF tracker performs better than most contrast trackers on UAV123 and is not inferior to the recently advanced trackers AutoTrack and ACSDCF, further validating the advantages of the proposed method.

VOT2018: To further evaluate the robustness and accuracy of the tracker, we also compare the SCDCF with 9 advanced trackers on VOT2018, including ECO [8], MCCT [40], C-COT [34], CSRDCF [13], UpdateNet [57], SiamFC [46], DCFNet [58], Staple [5], and KCF [29]. We rank all algorithms according to the EAO score, and the results are shown in Fig. 9. Table 4 reports the scores of all algorithms on the three indicators in detail. From the table, we can see that the EAO score of SCDCF reaches the highest 0.280, which is 1.3% higher than C-COT and 0.7% higher than MCCT. On Robustness, our method ranks second only to ECO. Although the performance of MCCT and ECO tracker based on deep features on VOT2018 is also excellent, the number of deep feature channels used is enormous, and the algorithm runs slowly, especially ECO. Therefore, compared with other advanced tracking algorithms, the overall performance of SCDCF is still in the optimal position.

4.5 Attribute-based evaluation

To fully evaluate the performance of the tracker in various complex scenarios, we perform attribute-based evaluation of SCDCF on OTB2015 and TC128 datasets. These attributes include occlusion (OCC), scale variation (SV), illumination variation (IV), background clutter (BC), fast motion (FM), blur (MB). Low resolution (LR), deformation (DEF), out-of-view (OV), out-of-plane rotation (OPR), and in-plane rotation (IPR). Figure 10 shows the results of the comparative analysis on OTB2015. The SCDCF ranks first in DP for eight attributes: IV, FM, DEF, BC, SV, OV, OPR, and LR. Especially in OV and LR, it is 3.5% and 4.2% higher than the second and third places (i.e., C-COT and MCPF), and the DP score of SCDCF is also among the top in the remaining attributes. Figure 11 shows the results of the attribute analysis of SCDCF on TC128. In terms of DP, our method achieves the best in SV, OCC, FM, and OPR and remains in the top three in the remaining seven challenges. In terms of AUC, the SCDCF tracker ranks first in SV, OCC, and second in six attributes: IV, FM, OPR, IPR, OV, and BC. The evaluation results show that SCDCF can better cope with target variations in a variety of complex scenarios

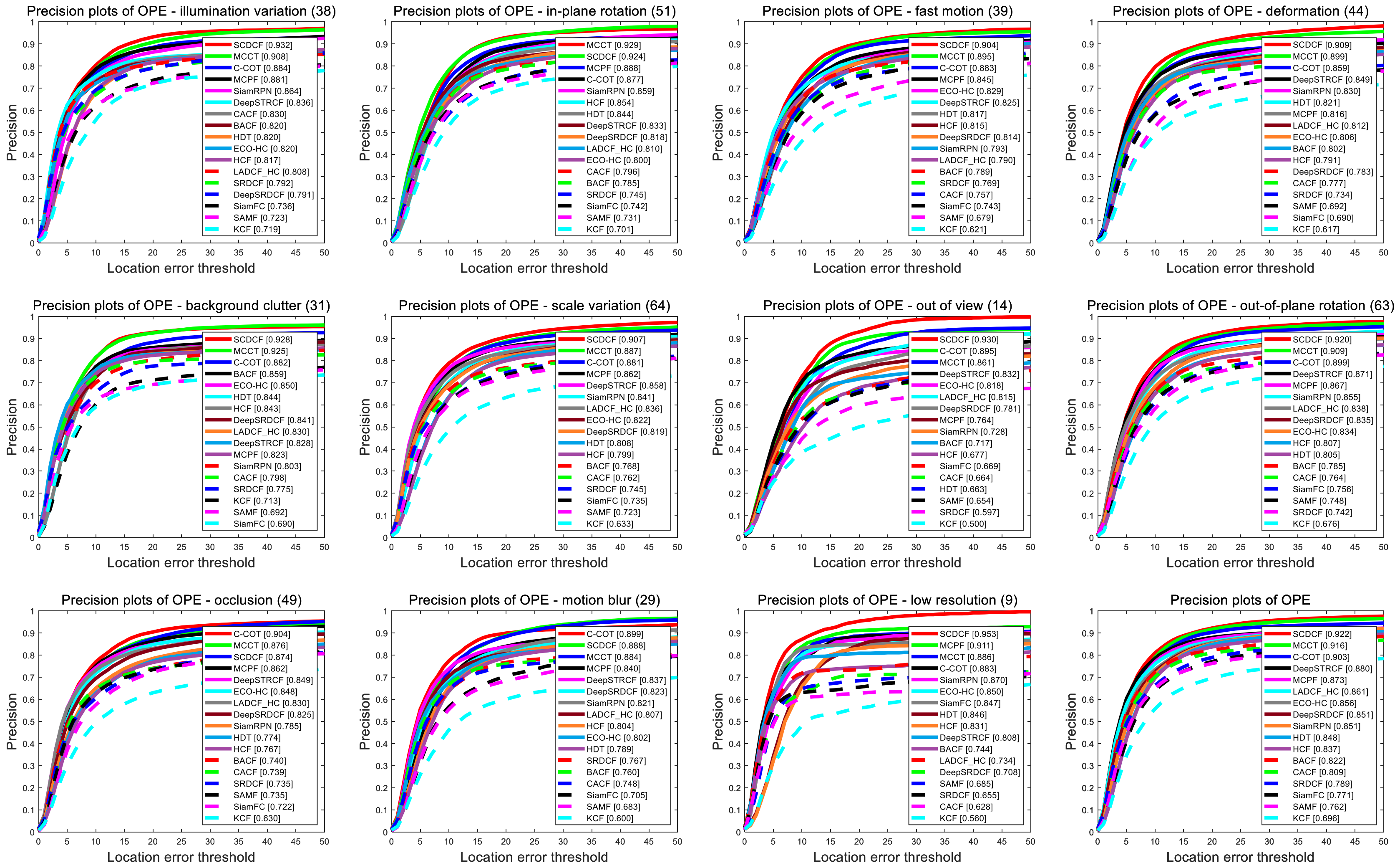


Fig. 10 Precision plots of SCDCF and 16 state-of-the-art trackers under 11 attributes on OTB2015. For completeness, we also show the overall results obtained by these trackers

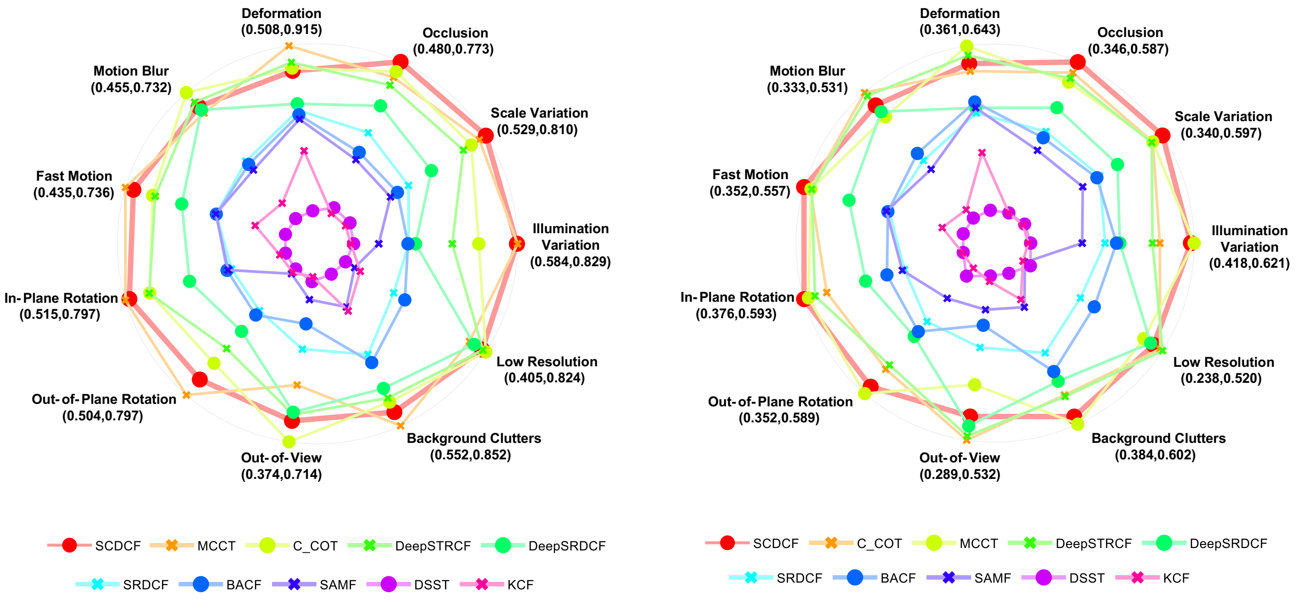


Fig. 11 The 11 attributes-based DP (left) and AUC (right) scores of our tracker and other trackers on TC128

by fully using effective feature channels to represent the target during tracking.

4.6 Ablation study

we further conduct ablation studies on the OTB2013 and OTB2015 datasets to evaluate the contribution of each component in the proposed SCDCF tracker, and the evaluation results are shown in Table 5. 'MU' indicates the proposed adaptive model update strategy we designed. 'SCS' stands for the saliency-aware channel selection strategy. 'CW' represents the channel weight assigned to the selected feature channel according to the SAER score. It can be seen from the table that each component improves the performance of the tracker to certain extent. In particular, after the introduction of salience-aware channel selection, the DP/AUC scores of the tracker improved significantly on OTB2013 and OTB2015, 3.0%/2.3% and 2.8%/1.8%, respectively. Using SAER scores to assign channel weights also enhanced the stability of the tracker. Compared to the baseline, SCDCF combines all component strengths to improve the AUC and DP metrics by 10.7%/7.7% and 10.0%/6.2% on OTB2013 and OTB2015, respectively.

4.7 Discussion

Qualitative and quantitative experiments on several datasets have verified that the proposed saliency-aware channel selection can effectively improve the tracking accuracy of the correlation filter algorithms. Although in most practical scenarios, using the saliency-aware detection mechanism proposed in Sect. 3.3, we can obtain masks that match the target appearance profile, as shown in Fig. 2. However, a few environments will still affect the effectiveness of the saliency-aware detection mechanism, as shown in Fig. 12. In Fig. 12 a, due to the similarity between the target local and

Table 5 DP and AUC scores of various variants of the proposed SCDCF on OTB2013 and OTB2015 datasets

Variant of our tracker	OTB2013		OTB2015	
	DP Score	AUC Score	DP Score	AUC Score
Baseline	0.835	0.639	0.822	0.621
Baseline+Deep	0.889	0.679	0.870	0.653
Baseline+Deep+MU	0.908	0.689	0.884	0.658
Baseline+Deep+SCS	0.919	0.702	0.898	0.671
Baseline+Deep+SCS+CW	0.924	0.707	0.913	0.677
Baseline+Deep+SCS+CW+MU	0.942	0.716	0.922	0.683

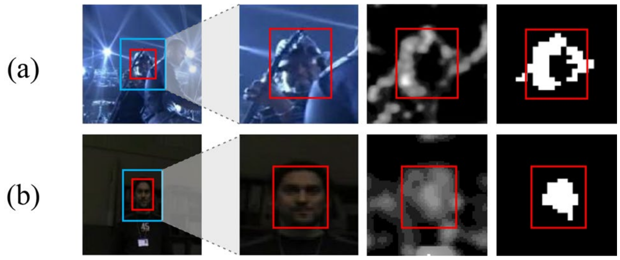


Fig. 12 Visualization of failure cases. The blue box represents the saliency detection region and the red box denotes the object region

the background environment, the mask generated by the saliency-aware detection has some missing regions. In Fig. 12 b, due to the low brightness and resolution of the image and the low color discrimination, the generated target mask is not complete enough. These results reveal the shortcomings of our method. To better improve the tracking performance, we will explore more advanced saliency detection algorithms to alleviate the above problems and study how to integrate saliency information with the DCF model further.

5 Conclusion

In this paper, we research the correlation between multi-channel deep features and target saliency-aware region information and propose a novel DCF-based tracking method via saliency-aware and adaptive channel selection. By comparing the feature energy of the target saliency-aware region and the background region, the more discriminative effective channels in the multi-dimensional convolution features are selected, and high tracking accuracy can be achieved using a small number of feature channels. In addition, the proposed SAER indicator can also be used to determine the importance of channels and realize the adaptive allocation of channel weights. We also introduce the ADMM method to optimize the proposed SCDCF model. Extensive experiments on five well-known datasets validate the effectiveness and robustness of the proposed method.

Acknowledgements This work was supported in part by the National Natural Science Foundation of China (Grant No. 62072370), in part by the Natural Science Foundation of Shaanxi Province (Grant No. 2023-JC-YB-598), and in part by the Science and Technology Project of Xi'an City (Grant No. 22GXFW0125)

Author contributions SM: conceptualization, methodology, validation, writing-review & editing. ZZ: software, writing-original draft, visualization. LP: data curation, investigation. ZH: validation, writing-review & editing. LZ: methodology, visualization. XZ: validation, project administration. All authors reviewed the manuscript.

Data availability Data will be made available on appropriate request.

Declarations

Conflict of interest All authors declare that they have no conflicts of interest affecting the work reported in this article.

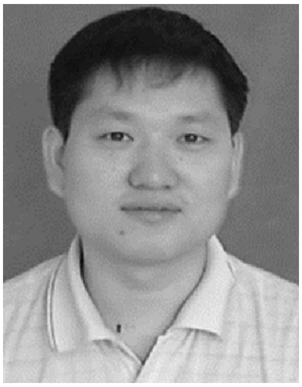
References

1. Fiaz, M., Mahmood, A., Javed, S.: Jung SK Handcrafted and deep trackers: Recent visual object tracking approaches and trends. ACM Comput. Surv. **52**(2), 1–44 (2019)
2. Ahmed, I.: Jeon G A real-time person tracking system based on siammask network for intelligent video surveillance. J. Real-Time Image Process **18**(5), 1803–1814 (2021)
3. Zhang, Z., Zhang, Y., Cheng, X.: Li K Siamese network for real-time tracking with action-selection. J. Real-Time Image Process **17**(5), 1647–1657 (2020)
4. Lin, B., Xue, X., Li, Y.: Shen Q Learning correlation filter with fused feature and reliable response for real-time tracking. J. Real-Time Image Process **19**(2), 417–427 (2022)
5. Bertinetto, L., Valmadre, J., Golodetz, S., Miksik, O., Torr, P.H., Staple: Complementary learners for real-time tracking. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), pp 1401–1409 (2016)
6. Mueller, M., Smith, N., Ghanem, B.: Context-aware correlation filter tracking. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), pp 1387–1395 (2017)
7. Galoogahi, H.K., Fagg, A., Lucey, S.: Learning background-aware correlation filters for visual tracking. In: Proc. IEEE Int. Conf. Comput. Vis. (ICCV), pp 1144–1152 (2017)
8. Danelljan, M., Bhat, G., Khan, F.S., Felsberg, M.: Eco: Efficient convolution operators for tracking. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), pp 6931–6939 (2017)
9. Danelljan, M., Hãger, G., Khan, F.S., Felsberg, M.: Convolutional features for correlation filter based visual tracking. In: Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW), pp 621–629 (2015)
10. Ma, C., Huang, J.B., Yang, X., Yang, M.H.: Hierarchical convolutional features for visual tracking. In: Proc. IEEE Int. Conf. Comput. Vis. (ICCV), pp 3074–3082 (2015)
11. Dai, K., Wang, D., Lu, H., Sun, C., Li, J.: Visual tracking via adaptive spatially-regularized correlation filters. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), pp 4665–4674 (2019)
12. Hu, J., Shen, L., Albanie, S., Sun, G.: Wu E Squeeze-and-excitation networks. IEEE Trans Pattern Anal. Mach. Intell. **42**(8), 2011–2023 (2020)
13. Lukežic A, Vojír T, Zajc LC, Matas J, Kristan M, Discriminative correlation filter with channel and spatial reliability. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), pp 4847–4856 (2017)
14. Du, F., Liu, P., Zhao, W.: Tang X Joint channel reliability and correlation filters learning for visual tracking. IEEE Trans. Circuits Syst. Video Technol. **30**(6), 1625–1638 (2020)
15. Fu, C., Xu, J., Lin, F., Guo, F., Liu, T.: Zhang Z Object saliency-aware dual regularized correlation filter for real-time aerial tracking. IEEE Trans. Geosci. Remote Sensing **58**(12), 8940–8951 (2020)
16. Feng, W., Han, R., Guo, Q., Zhu, J.: Wang S Dynamic saliency-aware regularization for correlation filter-based object tracking. IEEE Trans. Image Process **28**(7), 3232–3245 (2019)
17. Yang, X., Li, S., Ma, J.: yan Yang J, Yan J Co-saliency-regularized correlation filter for object tracking. Signal Process-Image Commun. **103**, 116655 (2022)
18. Zhang, P., Liu, W., Wang, D., Lei, Y., Wang, H.: Lu H Non-rigid object tracking via deep multi-scale spatial-temporal

discriminative saliency maps. Pattern Recognit. **100**, 107130 (2020)
19. Gao, L., Liu, B., Fu, P., Xu, M.: Li J Visual tracking via dynamic saliency discriminative correlation filter. Appl. Intell. **52**(6), 5897–5911 (2022)
20. Liang, Y., Liu, Y., Yan, Y., Zhang, L.: Wang H Robust visual tracking via spatio-temporal adaptive and channel selective correlation filters. Pattern Recognit. **112**, 107738 (2021)
21. Xu, T., Feng, Z., Wu, X.J.: Kittler J Adaptive channel selection for robust visual object tracking with discriminative correlation filters. Int. J. Comput. Vis. **129**(5), 1359–1375 (2021)
22. Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J., et al.: Distributed optimization and statistical learning via the alternating direction method of multipliers. Found Trends Mach. Learn. **3**(1), 1–122 (2011)
23. Wu, Y., Lim, J., Yang, M.H.: Online object tracking: A benchmark. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), pp 2411–2418 (2013)
24. Wu, Y., Lim, J., Yang, M.H.: Object tracking benchmark. IEEE Trans. Pattern Anal. Mach. Intell. **37**(9), 1834–1848 (2015)
25. Liang, P., Blasch, E., Ling, H.: Encoding color information for visual tracking: algorithms and benchmark. IEEE Trans Image Process **24**(12), 5630–5644 (2015)
26. Mueller M, Smith N, Ghanem B, A benchmark and simulator for uav tracking. In: Proc. Eur. Conf. Comput. Vis., Springer, pp 445–461 (2016)
27. Kristan M, Leonardis A, Matas J, Felsberg M, Pflugfelder R, Cehovin Zajc L, Vojir T, Bhat G, Lukezic A, Eldesokey A, et al., The sixth visual object tracking vot2018 challenge results. In: Proc. Eur. Conf. Comput. Vis. Workshops (ECCVW), pp 3–53 (2018)
28. Bolme DS, Beveridge JR, Draper BA, Lui YM, Visual object tracking using adaptive correlation filters. In: Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., pp 2544–2550 (2010)
29. JaF, Henriques, Caseiro, R., Martins, P., Batista, J.: High-speed tracking with kernelized correlation filters. IEEE Trans. Pattern Anal. Mach. Intell. **37**(3), 583–596 (2015)
30. Danelljan, M., Shahbaz Khan, F., Felsberg, M., Van de Weijer, J.: Adaptive color attributes for real-time visual tracking. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), pp 1090–1097 (2014)
31. Huang, Z., Fu, C., Li, Y., Lin, F., Lu, P.: Learning aberrance repressed correlation filters for real-time uav tracking. In: Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), pp 2891–2900 (2019)
32. Li, Y., Fu, C., Ding, F., Huang, Z., Lu, G.: Autotrack: Towards high-performance visual tracking for uav with automatic spatio-temporal regularization. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), pp 11920–11929 (2020)
33. Danelljan, M., Hãger, G., Khan, F.S., Felsberg, M.: Learning spatially regularized correlation filters for visual tracking. In: Proc. IEEE Int. Conf. Comput. Vis. (ICCV), pp 4310–4318 (2015)
34. Danelljan, M., Robinson, A., Khan, F.S., Felsberg, M.: Beyond correlation filters: Learning continuous convolution operators for visual tracking. In: Proc. Eur. Conf. Comput. Vis., Springer, pp 472–488 (2016)
35. Ma, S., Zhang, L., Hou, Z., Yang, X., Pu, L., Zhao, X.: Robust visual tracking via adaptive feature channel selection. Int. J. Intell. Syst. **37**(10), 6951–6977 (2022)
36. Hou, X., Zhang, L.: Saliency detection: A spectral residual approach. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), pp 1–8 (2007)
37. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. Int. J. Comput. Vis. **115**(3), 211–252 (2015)
38. Wang, M., Liu, Y., Huang, Z.: Large margin object tracking with circulant feature maps. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), pp 4800–4808 (2017)
39. Vedaldi, A., Lenc, K.: MatConvNet - Convolutional Neural Networks for MATLAB. arXiv e-prints arXiv:1412.4564, 1412.4564 (2014)
40. Wang, N., Zhou, W., Tian, Q., Hong, R., Wang, M., Li, H.: Multi-cue correlation filters for robust visual tracking. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., pp 4844–4853 (2018)
41. Li, F., Tian, C., Zuo, W., Zhang, L., Yang, M.H.: Learning spatial-temporal regularized correlation filters for visual tracking. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., pp 4904–4913 (2018)
42. Xu, T., Feng, Z.H., Wu, X.J., Kittler, J.: Learning adaptive discriminative correlation filters via temporal consistency preserving spatial feature selection for robust visual object tracking. IEEE Trans. Image Process **28**(11), 5596–5609 (2019)
43. Li, Y., Zhu, J.: A scale adaptive kernel correlation filter tracker with feature integration. In: Proc. Eur. Conf. Comput. Vis., Springer, pp 254–265 (2014)
44. Zhang, T., Xu, C., Yang, M.H.: Multi-task correlation particle filter for robust object tracking. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), pp 4819–4827 (2017)
45. Qi, Y., Zhang, S., Qin, L., Yao, H., Huang, Q., Lim, J., Yang, M.H.: Hedged deep tracking. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), pp 4303–4311 (2016)
46. Bertinetto, L., Valmadre, J., Henriques, J.F., Vedaldi, A., Torr, P.H.: Fully-convolutional siamese networks for object tracking. In: Proc. Eur. Conf. Comput. Vis., Springer, pp 850–865 (2016)
47. Li, B., Yan, J., Wu, W., Zhu, Z., Hu, X.: High performance visual tracking with siamese region proposal network. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., pp 8971–8980 (2018)
48. Wang, N., Zhou, W., Song, Y., Ma, C., Liu, W.: Li H Unsupervised deep representation learning for real-time tracking. Int. J. Comput. Vis. **129**, 400–418 (2021)
49. Danelljan, M., Gool, L.V., Timofte, R.: Probabilistic regression for visual tracking. In: Proc. IEEE/CVF Comput. Vis. Pattern Recognit. (CVPR), pp 7183–7192 (2020)
50. Yang, T., Xu, P., Hu, R., Chai, H., Chan, A.B.: Roam: Recurrently optimizing tracking model. In: Proc. IEEE/CVF Comput. Vis. Pattern Recognit. (CVPR) (2020)
51. Danelljan, M., Bhat, G., Khan, F.S., Felsberg, M.: Atom: Accurate tracking by overlap maximization. In: Proc. IEEE/CVF Comput. Vis. Pattern Recognit. (CVPR), pp 4660–4669 (2019)
52. Li, P., Chen, B., Ouyang, W., Wang, D., Yang, X., Lu, H.: Gradnet: Gradient-guided network for visual object tracking. In: Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), pp 6162–6171 (2019)
53. Zhu, Z., Wang, Q., Li, B., Wu, W., Yan, J., Hu, W.: Distractor-aware siamese networks for visual object tracking. In: Proc. Eur. Conf. Comput. Vis. (ECCV), pp 101–117 (2018)
54. Bhat, G., Danelljan, M., Gool, L.V., Timofte, R.: Learning discriminative model prediction for tracking. In: Proc. IEEE/CVF Comput. Vis. Pattern Recognit. (CVPR), pp 6182–6191 (2019)
55. Hare, S., Golodetz, S., Saffari, A., Vineet, V., Cheng, M.M., Hicks, S.L.: Torr PHS struck: structured output tracking with kernels. IEEE Trans. Pattern Anal. Mach. Intell. **38**(10), 2096–2109 (2016)
56. Danelljan, M., Häger, G., Khan, F., Felsberg, M.: Accurate scale estimation for robust visual tracking. In: British Machine Vision Conference, Nottingham, September 1-5, 2014, Bmva Press (2014)

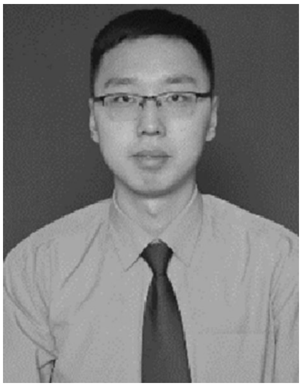
57. Zhang, L., Gonzalez-Garcia, A., Weijer, J.V.D., Danelljan, M., Khan, F.S.: Learning the model update for siamese trackers. In: Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), pp 4009–4018 (2019)
58. Wang, Q., Gao, J., Xing, J., Zhang, M., Hu, W.: DCFNet: Discriminant Correlation Filters Network for Visual Tracking. arXiv e-prints arXiv:1704.04057, 1704.04057 (2017)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



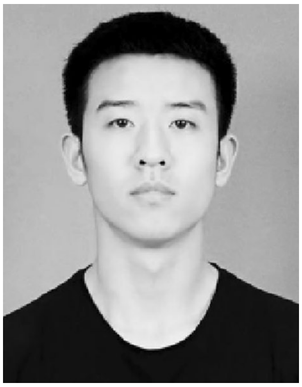
Sugang Ma is currently pursuing the Ph.D. degree in computer science with Chang'an University. He received his Master's degree in XIDIAN University in 2010. He is currently a Senior Engineer with the Xi'an University of Posts and Telecommunications and a Senior Member of the China Communications Society. His current research interests include Computer Vision and Machine Learning.

Zhixian Zhao received the B.E. degree from the Anyang Institute of Technology, Anyang, Henan, China, in 2019. He is currently pursuing the M.S. degree with the School of Computer Science and Technology, Xi'an University of Posts and Telecommunications, Xi'an, Shaanxi, China. His research interests include visual tracking and computer vision.



Lei Pu received the Ph.D. degree from Air Force Engineering University, Xi'an, China, in 2020. He is currently working in Rocket Force Engineering University and his main research interests include visual tracking, pattern recognition, and computer vision.

Zhiqiang Hou received the Ph.D. degree from Xi'an Jiaotong University, in 2005. He was a Visiting Scholar with the University of Bristol, U.K., in 2009. He is currently a Professor with Xi'an University of Posts and Telecommunications. His research interests include pattern recognition, computer vision, image processing, and information fusion.



Lei Zhang received the M.E. degree from the School of Computer Science and Technology, Xi'an University of Posts and Telecommunications, Shaanxi, China, in 2022. He is currently pursuing the Ph.D. degree in School of Automation, Northwestern Polytechnical University, Shaanxi. His current research interests are computer vision and autonomous driving.

Xiangmo Zhao is currently the Vice President of Chang' an University and the Director of the Science and Technology Innovation Team of Multi-sources Traffic Information Sensing and Fusion, Ministry of Education, and also a Professor with the School of Information Engineering. His research interests include the Internet of Vehicles, testing of intelligent vehicles, intelligent transportation systems, and nondestructive testing for road infrastructures.