Complex & Intelligent Systems (2022) 8:3765–3777
https://doi.org/10.1007/s40747-021-00544-1

ORIGINAL ARTICLE

Spatio-temporal joint aberrance suppressed correlation filter for visual tracking

Libin Xu[1] · Pyoungwon Kim[2] · Mengjie Wang[1] · Jinfeng Pan[1] · Xiaomin Yang[3] · Mingliang Gao[1]

Received: 31 July 2021 / Accepted: 14 September 2021 / Published online: 25 September 2021

Abstract

The discriminative correlation filter (DCF)-based tracking methods have achieved remarkable performance in visual tracking. However, the existing DCF paradigm still suffers from dilemmas such as boundary effect, filter degradation, and aberrance. To address these problems, we propose a spatio-temporal joint aberrance suppressed regularization (STAR) correlation filter tracker under a unified framework of response map. Specifically, a dynamic spatio-temporal regularizer is introduced into the DCF to alleviate the boundary effect and filter degradation, simultaneously. Meanwhile, an aberrance suppressed regularizer is exploited to reduce the interference of background clutter. The proposed STAR model is effectively optimized using the alternating direction method of multipliers (ADMM). Finally, comprehensive experiments on TC128, OTB2013, OTB2015 and UAV123 benchmarks demonstrate that the STAR tracker achieves compelling performance compared with the state-of-the-art (SOTA) trackers.

Keywords Visual tracking · Correlation filter · Spatio-temporal constraint · Aberrance suppression

Introduction

Visual tracking aims to estimate the state of the target in image sequences, given its initial state. It plays a crucial role in computer vision-based applications, e.g., vehicle navigation, video surveillance and robotic perception [2,16,26,31]. In recent years, the DCF-based methods have attracted extensive attention due to the high efficiency. However, DCF-based tracking remains a challenging problem due to many intricate issues, such as boundary effect, filter degradation, and aberrance.

Boundary effect. The efficiency of DCF-based methods relies on the periodic assumption at the stage of training and detection. However, this assumption induces the filters to be trained and performed on partially unreal samples and subsequently results in the unexpected boundary effect. The boundary effect mainly impedes the performance of the DCF in two aspects [13]. (i) The inaccurate negative training samples reduce the discriminative power of the learned filters. (ii) The detection scores are reliable only around the center of the region, while the remaining scores are heavily influenced by the periodic repetitions of the detection samples. To address this issue, several competitive DCF-based trackers utilize the constant spatial regularizer to penalize the filter coefficients outside the bounding box [13,18,25]. However, these constant spatial constraints are usually fixed at the stage of tracking, and the diverse information (e.g., the appearance variation of the target and the confidence of the tracking results) is not fully utilized. To address this problem, in this paper, we propose a dynamic spatial regularizer based on response variation rate, which enables the filter to learn more reliable filter coefficients.

Filter degradation. Generally, the DCF-based methods adopt the model update mechanism based on fixed rate, which ignores the variation between different frames [45]. Once the appearance of the target varies dramatically, the filter learned from the previous frame cannot adjust to appearance changes, resulting in the filter degradation. To cope with the filter degradation, several DCF-based trackers adopt the temporal regularizer into filter training [25,28,45]. Nevertheless, the temporal regularizer is based on the assumption

✉ Mingliang Gao
mlgao@sdut.edu.cn

[1] School of Electrical and Electronic Engineering, Shandong University of Technology, Zibo 255000, China

[2] College of Education, Incheon National University, Incheon 22012, South Korea

[3] School of Electronics and Information, Sichuan University, Chengdu 610065, China

that filters between consecutive frames should be coherent. The filter training may be interfered with severe occlusion, background clutter, etc., resulting in a corrupted filter and breaking this assumption. To solve this issue, in this paper, we propose a dynamic temporal regularizer based on average peak-to-correlation energy (APCE) [39] to suppress the filter degradation.

Aberrance. Due to the spatial regularization, the correlation filter can be learned on larger image regions [13]. Nevertheless, with the expansion of the learning regions, more background clutter will be introduced, leading to aberrance at the detection stage, which is manifested as the abrupt variation in response maps. To reduce the effect of aberrance, Wang et al. [39] proposed the Large Margin Object Tracking (LMCF) method, in which the quality of response maps is verified during the filter learning and used to carry out the model updating in high confidence. Choi et al. [7] proposed the Attentional Correlation Filter Network (ACFN) tracker that integrates multiple correlation filters into a network. The verified scores which are generated based on response maps are utilized to select the suitable filter. However, these trackers deal with the aberrance at the stage of detection, and thus the tracking performance is decreased inevitably. Unlike these trackers, in this paper, we integrate an aberrance suppressed regularizer into the DCF schema to suppress the aberrance at the stage of filter training.

In this work, we address the above issues simultaneously under a unified framework of response map by learning a spatio-temporal joint aberrance suppressed regularization correlation filter. The main contributions are summarized as follows.

1. A novel tracking method by learning spatio-temporal joint aberrance suppressed regularization correlation filter (STAR) is proposed under a unified framework of response map.
2. A dynamic spatio-temporal regularizer is introduced to alleviate the boundary effect and filter degradation, simultaneously.
3. An aberrance suppressed strategy is introduced into the filter learning to minimize the interference by the background cluster.
4. Extensive evaluations are conducted on four challenging tracking benchmarks, and the experimental results demonstrate the competitive performance of the proposed tracker compared with the state-of-the-art (SOTA) tracking methods.

The rest of this paper is organized as follows. In “Related work”, we present an overview of the prior work most relevant to the proposed method. In “Proposed method”, the proposed STAR model is introduced, and the ADMM algorithm is developed to solve the STAR efficiently. In “Experimental results”, quantitative and qualitative evaluations of the proposed tracker with the SOTA trackers are presented. Conclusions are presented in “Conclusion”.

Related work

The visual tracking methods can be classified into generative tracking methods and discriminative tracking methods [31,40]. Among the discriminative-based trackers, the DCF promote the visual tracking to a new level.

Generative tracking

The generative tracking attempts to build models to represent the appearance of the target and search the most similar candidate region with minimal reconstruction error. Comaniciu et al. [8] proposed the mean-shift tracking method with iterative histogram matching for visual tracking. Adam et al. [1] proposed the fragments-based tracker, which utilizes multiple image fragments to represent the object. Subsequently, Ross et al. [35] proposed the subspace-based tracking method to learn and update the low-dimensional subspace representation of the target. Although generative tracking has achieved considerable success in constrained scenarios, they are vulnerable to complicated appearance variations of the target. Therefore, more attention is shifted to discriminative tracking, due to it is less susceptible to background clutter during the tracking process.

Discriminative tracking

The discriminative tracking trains a classifier to discriminate the target from the background. Grabner et al. [19] proposed an online boosting tracker by fusing multiple weak classifiers. Kalal et al. [24] proposed the Tracking–Learning–Detection (TLD) tracker that decomposes the long-term tracking into three sub-tasks, namely tracking, learning, and detection. More recently, many deep neural network (DNN) based trackers under the framework of “end-to-end learning” and “offline-learning and online-tracking” are proposed. For example, Bertinetto et al. [4] proposed the Fully Convolutional Siamese Networks (SiamFC) tracker that trains a fully convolutional siamese network by cross-correlating two inputs of the bilinear layer. Valmadre et al. [37] put forward the CFNet tracker that considers the correlation filter as a differentiable layer of the deep neural network. In general, discriminative tracking is relatively more effective than generative tracking in preventing the negative effects of complex background clutter or target appearance variations [40].

DCF-based tracking

Recently, DCF has received considerable attention due to its efficiency and scalability. Bolme et al. [5] first proposed the correlation filter tracker, termed minimum output sum of squared error (MOSSE), to learn a filter between multiple training image patches and a template of user-specified ideal correlation response. Henriques et al. [21] proposed the circulant structure of Tracking-by-Detection with Kernels (CSK) tracker, which exploits the circulant structure of the local image patch to learn a kernel regularized least squares classifier.

To further improve the tracking performance, the follow-up improvements are mainly carried out around two aspects, namely feature representation and scale estimation. In feature representation, Danelljan et al. [11] proposed the color attributes tracker by investigating the color names (CN) [38] feature in the tracking-by-detection framework. Henriques et al. [22] proposed the kernelized correlation filters (KCF) method by utilizing the histogram of oriented gradient (HOG) [9] feature. In addition, Bertinetto et al. [3] proposed the Sum of Template And Pixel-wise LEarners (STAPLE) tracker using the HOG and colour features to improve the tracking credibility. Moreover, Convolutional Neural Network (CNN) features have been used to further improve the feature representation [12,14,25,45]. In scale estimation, Danelljan et al. [10] proposed the Discriminative Scale Space Tracking (DSST) method, which learns a separate scale filter to address the scale variation. Li et al. [27] proposed the Scale Adaptive with Multiple Features (SAMF) tracker by employing a bilinear interpolation to generate image representations in multiple scales.

Proposed method

Revisit the standard DCF

In the standard DCF [22], $\mathbf{x} \in \mathbb{R}^{M\times N\times C}$ denotes the training sample with $M \times N$ feature size and C channels. $\mathbf{y} \in \mathbb{R}^{M\times N}$ is the corresponding Gaussian-shaped label (desired output). The filter $\mathbf{f} \in \mathbb{R}^{M\times N\times C}$ is trained by regressing the samples, which is defined as follows,

$$\arg\min_{\mathbf{f}} \frac{1}{2}\left\|\sum_{c=1}^{C}\mathbf{x}^{c} * \mathbf{f}^{c} - \mathbf{y}\right\|_F^2 + \alpha\sum_{c=1}^{C}\left\|\mathbf{f}^{c}\right\|_F^2, \tag{1}$$

where $*$ stands for the circular convolution operator, and α is the regularization parameter to prevent overfitting.

In the standard DCF model, there are several problems need to be further addressed. (i) It suffers from periodic repetitions on boundary positions caused by circulant shifted training sample. (ii) It does not tackle the problem of filter degradation, since the model is updated based on fixed rate. (iii) There is no response mechanism to copy with the aberrance, and the target will be easily lost when aberrance occurs.

The proposed model STAR

To address the problems mentioned above, we propose a novel spatio-temporal joint aberrance suppressed regularization (STAR) correlation filter for robust visual tracking. The tracking framework of the proposed STAR model is shown in Fig. 1. The spatial regularizer, temporal regularizer and aberrance suppressed regularizer are exploited to the standard DCF to tackle the boundary effect, filter degradation and aberrance suppression, simultaneously.

We assume that the learning of the correlation filter $\mathbf{f}$ is conducted for the t-th frame. The filter is learned by minimizing the following objective function,

$$\arg\min_{\mathbf{f}} \frac{1}{2}\left\|\sum_{c=1}^{C}\mathbf{x}^{c} * \mathbf{f}^{c} - \mathbf{y}\right\|_F^2 + \frac{\lambda}{2}\mathcal{R}_{\mathrm{s}} + \frac{\mu}{2}\mathcal{R}_{\mathrm{t}} + \frac{\eta}{2}\mathcal{R}_{\mathrm{a}}, \tag{2}$$

where $\left\|\sum_{c=1}^{C}\mathbf{x}^{c} * \mathbf{f}^{c} - \mathbf{y}\right\|_F^2$ denotes the regression loss parameterized by $\mathbf{f}$. The $\mathcal{R}_{\mathrm{s}}$, $\mathcal{R}_{\mathrm{t}}$ and $\mathcal{R}_{\mathrm{a}}$ refer to the spatial, temporal and aberrance suppressed regularizer, respectively. The parameters λ, μ and η are the corresponding coefficients to the regularizers.

Dynamic spatial regularizer

The constant spatial regularizer in the SOTA trackers (e.g., SRDCF [13], BACF [18] and STRCF [25]) does not fully exploit the diversity information of the target. The filter coefficients will be unreliable, leading to tracking failures, when the target suffers from interferences, e.g., severe occlusion, background clutter. To solve this problem, we design a dynamic spatial regularizer based on the response variation rate.

The response variation rate is defined as $\mathbf{\Pi} = \left\|\Pi^{1}, \Pi^{2}, \ldots, \Pi^{MN}\right\|$, and the i-th element Π^{i} is defined as,

$$\Pi^{i} = \frac{\mathrm{R}_{t}^{i} - (\mathrm{R}_{t-1}[\psi\triangle])^{i}}{(\mathrm{R}_{t-1}[\psi\triangle])^{i}}, \tag{3}$$

where $[\psi\triangle]$ is the shift operator. It enables the peaks of response R_{t} and R_{t-1} to coincide with each other to eliminate the motion influence [23]. Considering that the response variation rate $\mathbf{\Pi}$ reveals the confidence level of each pixel in the search area, we introduce $\mathbf{\Pi}$ into the spatial weight $\mathbf{w}$,

$$\mathbf{w} = \delta \log \mathbf{\Pi} + \tilde{\mathbf{w}}, \tag{4}$$

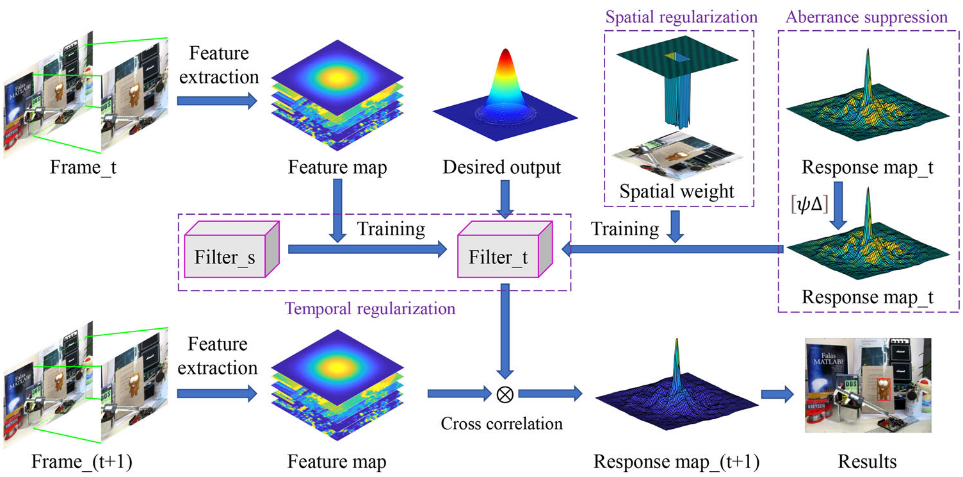

Fig. 1 Tracking framework of the proposed STAR model

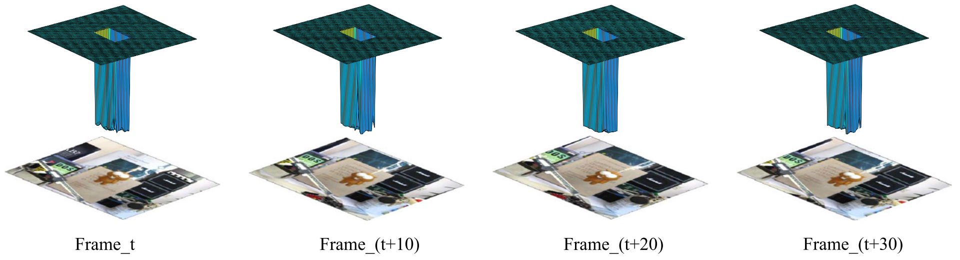

Fig. 2 Visualization of the dynamic variation of the spatial regularization weight in the process of tracking

where δ is a hyperparameter for adjusting the weight of $\mathbf{\Pi}$, and $\tilde{\mathbf{w}}$ is a matrix for initializing spatial regularization weight $\mathbf{w}$. The dynamic spatial regularizer of STAR model is defined as,

$$\mathcal{R}_{\mathrm{s}} = \sum_{c=1}^{C} \left\| \mathbf{w}_t \odot \mathbf{f}_t^c \right\|_F^2 , \tag{5}$$

where $\odot$ is the Hadamard product. The visualization of the dynamic variation of the spatial regularization is shown in Fig. 2. It shows that the dynamic spatial regularizer can impose different penalties on the spatial position according to the value of the response variation rate. Specifically, it imposes a higher penalty on the larger part of the response variation rate while a lower penalty on the smaller part. Thus, it achieves more reliable filter coefficients at the detection state.

Dynamic temporal regularizer

The existing temporal regularizer $\sum_{c=1}^{C} \left\| \mathbf{f}_t^c - \mathbf{f}_{t-1}^c \right\|_F^2$ is constructed using the previous filter $\mathbf{f}_{t-1}$ (e.g., STRCF [25], LADCF [45] and AutoTrack [28]). The filter learned at frame t is affected to a large extent by the filter $\mathbf{f}_{t-1}$. However, $\mathbf{f}_{t-1}$ may be corrupted by occlusion or background clutter; thus, it will break the assumption that the filters between consecutive frames should be coherent. To tackle this issue, we propose to learn a dynamic temporal regularizer based on APCE measure. The APCE measure is defined as,

$$\mathrm{APCE} = \frac{|\mathrm{R}_{\max} - \mathrm{R}_{\min}|^2}{\mathrm{mean}\left[\sum_{w,h} \left(\mathrm{R}_{w,h} - \mathrm{R}_{\min}\right)^2\right]}, \tag{6}$$

where $\mathrm{R}_{\max}$, $\mathrm{R}_{\min}$ and $\mathrm{R}_{w,h}$ denote the maximum, minimum and the wth row hth column elements of the response R, respectively. The visualization of the value of APCE with its corresponding threshold in a typical tracking sample is shown in Fig. 3. At the stage of training, the filter may be corrupted by occlusion, background clutter, etc., then, the response map with interference is generated by the convolution of the corrupted filter and the feature map. As a consequence, the value of APCE obtained by Eq. (6) will drop significantly. This specialty of APCE can be adopted to judge whether the filter is corrupted or not. Subsequently, the uncorrupted filter $\mathbf{f}_{\mathrm{s}}$ is selected for temporal regularizer instead of $\mathbf{f}_{t-1}$, as follows,

$$\begin{aligned} \mathbf{f}_{\mathrm{s}} &= \begin{cases} \mathbf{f}_{t-1} & \text{if } \mathrm{APCE}_t > \zeta \mathrm{APCE}_{\mathrm{hm}} \\ \mathbf{f}_{t-i} & \text{otherwise} \end{cases} \\ \text{s.t.,}\ \ i &= \begin{cases} i \in \mathbb{N} \\ i > 1 \\ \arg\min\limits_{i} \left(\mathrm{APCE}_{t-i+1} > \zeta \mathrm{APCE}_{\mathrm{hm}}\right) \end{cases}, \end{aligned} \tag{7}$$

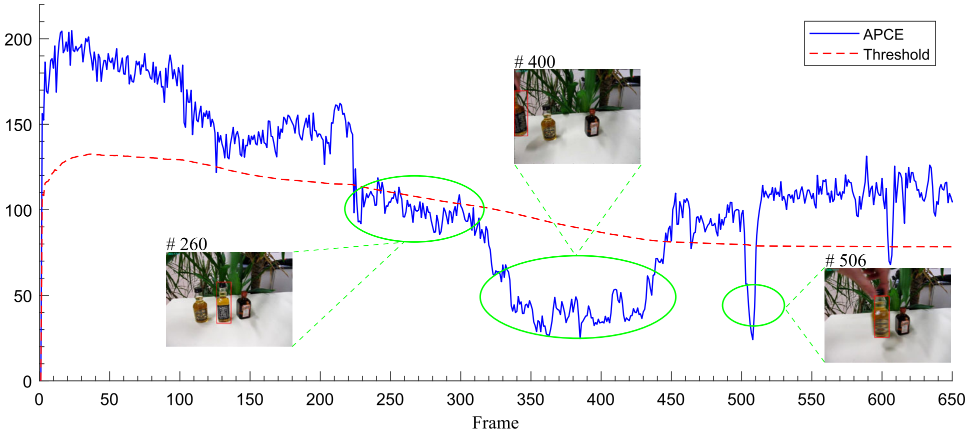

Fig. 3 Visualization of the variation of APCE value with its corresponding threshold in a typical tracking sample. When the target encounters motion blur (Frame 260), out-of-view (Frame 400) or full occlusion (Frame 506), the value of APCE drops significantly, and it is lower than the threshold ($\zeta\,\mathrm{APCE_{hm}}$)

where $\mathbf{f}_{t-1}$ and $\mathbf{f}_{t-i}$ denote the filter at the $(t-1)$-th and $(t-i)$-th frame, respectively. ζ is hyperparameter, and $\mathrm{APCE_{hm}}$ stands for the historical mean value of APCE.

The uncorrupted filter $\mathbf{f}_{\mathrm{s}}$ is selected to construct the dynamic temporal regularizer for the STAR model as follows,

$$\mathcal{R}_{\mathrm{t}} = \sum_{c=1}^{C} \left\| \mathbf{f}_t^c - \mathbf{f}_{\mathrm{s}}^c \right\|_F^2 . \tag{8}$$

Compared with the existing temporal regularization methods [25,28,45], the STAR model takes the full advantage of the video continuity natures by exploiting $\|\mathbf{f}_t - \mathbf{f}_{\mathrm{s}}\|_F^2$ to penalize the difference between the current filter $\mathbf{f}_t$ and the uncorrupted filter $\mathbf{f}_{\mathrm{s}}$. Thus, the proposed STAR gains a more robust appearance model, and alleviate the filter degradation effectively.

Aberrance suppressed regularizer

The response map can reveal the confidence degree about the tracking results to a large extent [39]. The aberrance caused by background clutter occurs at the detection stage, and it will result in an abrupt variation in response maps. The aberrance can be effectively repressed by restricting the response variation. As a result, an aberrance suppressed regularizer is introduced to handle the aberrance at the stage of training. The aberrance suppressed regularizer is formulated as,

$$\mathcal{R}_{\mathrm{a}} = \left\| \mathrm{R}_t - \mathrm{R}_{t-1}[\psi\triangle] \right\|_F^2 , \tag{9}$$

where all the variables have been explained in the Eq. (3).

Optimization of STAR

After all the regularization defined, optimization of the Eq. (2) is one of the key to solve the tracking. The Eq. (2) can be minimized using ADMM [6] to achieve the optimal solution benefitting from its convexity. Specifically, we introduce the auxiliary variable $\mathbf{g} = \mathbf{f}$ and the step size parameter γ to construct the following augmented Lagrange function,

$$\begin{aligned}\mathcal{L} &= \frac{1}{2}\left\| \sum_{c=1}^{C} \mathbf{x}_t^c * \mathbf{f}_t^c - \mathbf{y} \right\|_F^2 + \frac{\lambda}{2}\sum_{c=1}^{C} \left\| \mathbf{w}_t \odot \mathbf{g}_t^c \right\|_F^2 \\ &+ \frac{\mu}{2}\sum_{c=1}^{C} \left\| \mathbf{f}_t^c - \mathbf{f}_s^c \right\|_F^2 + \frac{\eta}{2}\left\| \sum_{c=1}^{C} \mathbf{x}_t^c * \mathbf{f}_t^c - \mathbf{r} \right\|_F^2 \\ &+ \sum_{c=1}^{C} \left(\mathbf{f}_t^c - \mathbf{g}_t^c\right)^{\mathrm{T}} \mathbf{s}_t^c + \frac{\gamma}{2}\sum_{c=1}^{C} \left\| \mathbf{f}_t^c - \mathbf{g}_t^c \right\|_F^2 ,\end{aligned} \tag{10}$$

where $\mathbf{r} = \mathbf{R}_{t-1}[\psi\triangle]$, and $\mathbf{s}$ refers to the Lagrange multiplier. By introducing $\mathbf{h} = \frac{1}{\gamma}\mathbf{s}$, Eq. (10) can be reformulated as,

$$\begin{aligned}\mathcal{L} &= \frac{1}{2}\left\| \sum_{c=1}^{C} \mathbf{x}_t^c * \mathbf{f}_t^c - \mathbf{y} \right\|_F^2 + \frac{\lambda}{2}\sum_{c=1}^{C} \left\| \mathbf{w}_t \odot \mathbf{g}_t^c \right\|_F^2 \\ &+ \frac{\mu}{2}\sum_{c=1}^{C} \left\| \mathbf{f}_t^c - \mathbf{f}_s^c \right\|_F^2 + \frac{\eta}{2}\left\| \sum_{c=1}^{C} \mathbf{x}_t^c * \mathbf{f}_t^c - \mathbf{r} \right\|_F^2 \\ &+ \frac{\gamma}{2}\sum_{c=1}^{C} \left\| \mathbf{f}_t^c - \mathbf{g}_t^c + \mathbf{h}_t^c \right\|_F^2 .\end{aligned} \tag{11}$$

Then, the following subproblems are alternately optimized via ADMM formulation.

$$\begin{cases} \mathbf{f}^{i+1} = \arg\min\limits_{\mathbf{f}} \frac{1}{2} \left\| \sum_{c=1}^{C} \mathbf{x}_t^c * \mathbf{f}_t^c - \mathbf{y} \right\|_F^2 + \frac{\mu}{2} \sum_{c=1}^{C} \left\| \mathbf{f}_t^c - \mathbf{f}_s^c \right\|_F^2 \\ \qquad + \frac{\eta}{2} \left\| \sum_{c=1}^{C} \mathbf{x}_t^c * \mathbf{f}_t^c - \mathbf{r} \right\|_F^2 + \frac{\gamma}{2} \sum_{c=1}^{C} \left\| \mathbf{f}_t^c - \mathbf{g}_t^c + \mathbf{h}_t^c \right\|_F^2 \\ \mathbf{g}^{i+1} = \arg\min\limits_{\mathbf{g}} \frac{\lambda}{2} \sum_{c=1}^{C} \left\| \mathbf{w}_t \odot \mathbf{g}_t^c \right\|_F^2 + \frac{\gamma}{2} \sum_{c=1}^{C} \left\| \mathbf{f}_t^c - \mathbf{g}_t^c + \mathbf{h}_t^c \right\|_F^2 \\ \mathbf{h}^{i+1} = \mathbf{h}^i + \mathbf{f}^{i+1} - \mathbf{g}^{i+1} \end{cases} . \tag{12}$$

Subproblem f: For the first subproblem of Eq. (12), it can be transformed into the frequency domain using Parseval's formulation as,

$$\begin{aligned} \widehat{\mathbf{f}}^* = \arg\min_{\widehat{\mathbf{f}}} \frac{1}{2} \left\| \sum_{c=1}^{C} \widehat{\mathbf{x}}_t^c \odot \widehat{\mathbf{f}}_t^c - \widehat{\mathbf{y}} \right\|_F^2 + \frac{\mu}{2} \sum_{c=1}^{C} \left\| \widehat{\mathbf{f}}_t^c - \widehat{\mathbf{f}}_s^c \right\|_F^2 \\ + \frac{\eta}{2} \left\| \sum_{c=1}^{C} \widehat{\mathbf{x}}_t^c \odot \widehat{\mathbf{f}}_t^c - \widehat{\mathbf{r}} \right\|_F^2 + \frac{\gamma}{2} \sum_{c=1}^{C} \left\| \widehat{\mathbf{f}}_t^c - \widehat{\mathbf{g}}_t^c + \widehat{\mathbf{h}}_t^c \right\|_F^2 , \end{aligned} \tag{13}$$

where $\hat{}$ denotes the discrete Fourier transform (DFT). The j-th element of the label $\widehat{\mathbf{y}}$ relies on the j-th element of the sample $\widehat{\mathbf{x}}_t$ and the filter $\widehat{\mathbf{f}}_t$ across all C channels. $\mathcal{V}(\mathbf{f}) \in \mathbb{R}^C$ is the vector consisting of the j-th element of $\mathbf{f}$ along the channels. Equation (13) can be further decomposed into $M \times N$ subproblems, where each subproblem is defined as,

$$\begin{aligned} \mathcal{V}_j(\widehat{\mathbf{f}}^*) = \arg\min_{\mathcal{V}_j(\widehat{\mathbf{f}})} \frac{1}{2} \left\| \mathcal{V}_j(\widehat{\mathbf{x}}_t)^{\mathrm{T}} \mathcal{V}_j(\widehat{\mathbf{f}}_t) - \widehat{\mathbf{y}}_j \right\|_F^2 \\ + \frac{\mu}{2} \left\| \mathcal{V}_j(\widehat{\mathbf{f}}_t) - \mathcal{V}_j(\widehat{\mathbf{f}}_s) \right\|_F^2 \\ + \frac{\eta}{2} \left\| \mathcal{V}_j(\widehat{\mathbf{x}}_t)^{\mathrm{T}} \mathcal{V}_j(\widehat{\mathbf{f}}_t) - \widehat{\mathbf{r}}_j \right\|_F^2 + \frac{\gamma}{2} \left\| \mathcal{V}_j(\widehat{\mathbf{f}}_t) \right. \\ \left. - \mathcal{V}_j(\widehat{\mathbf{g}}_t) + \mathcal{V}_j(\widehat{\mathbf{h}}_t) \right\|_F^2 , \end{aligned} \tag{14}$$

where superscript $^{\mathrm{T}}$ on a complex vector or matrix indicates conjugate transpose operation. Taking the derivative of Eq. (14) as zero, the closed-form solution of $\mathcal{V}_j(\widehat{\mathbf{f}}^*)$ can be denoted as,

$$\mathcal{V}_j(\widehat{\mathbf{f}}^*) = \left[(1+\eta)\, \mathcal{V}_j(\widehat{\mathbf{x}}_t) \mathcal{V}_j(\widehat{\mathbf{x}}_t)^{\mathrm{T}} + (\mu+\gamma) \right]^{-1} \mathbf{q}, \tag{15}$$

where the vector $\mathbf{q} = \mathcal{V}_j(\widehat{\mathbf{x}}_t)\widehat{\mathbf{y}}_j + \eta \mathcal{V}_j(\widehat{\mathbf{x}}_t)\widehat{\mathbf{r}}_j + \gamma \mathcal{V}_j(\widehat{\mathbf{g}}_t) - \gamma \mathcal{V}_j(\widehat{\mathbf{h}}_t) + \mu \mathcal{V}_j(\widehat{\mathbf{f}}_s)$. Since $\mathcal{V}_j(\widehat{\mathbf{x}}_t)\mathcal{V}_j(\widehat{\mathbf{x}}_t)^{\mathrm{T}}$ is a rank-1 matrix, Eq. (15) can be further rewritten via the Sherman–Morrsion formulation [32] as,

$$\mathcal{V}_j(\widehat{\mathbf{f}})^* = \frac{1}{\mu+\gamma} \left[\mathbf{I} - \frac{\mathcal{V}_j(\widehat{\mathbf{x}}) \mathcal{V}_j(\widehat{\mathbf{x}})^{\mathrm{T}}}{\frac{\mu+\gamma}{1+\eta} + \mathcal{V}_j(\widehat{\mathbf{x}})^{\mathrm{T}} \mathcal{V}_j(\widehat{\mathbf{x}})} \right] \mathbf{q}. \tag{16}$$

Note that Eq. (16) only contains vector multiply–add operation, thus it can be computed efficiently. $\mathbf{f}$ can be further obtained by the IDFT of $\widehat{\mathbf{f}}$.

Subproblem g: For the second subproblem of Eq. (12), each element of $\mathbf{g}$ can be computed independently as,

$$\mathbf{g}^* = \frac{\gamma\,(\mathbf{f}+\mathbf{h})}{\lambda\,(\mathbf{w} \odot \mathbf{w}) + \gamma \mathbf{I}}. \tag{17}$$

Lagrangian multiplier update: The Lagrange multiplier is updated as,

$$\mathbf{h}^{i+1} = \mathbf{h}^i + \mathbf{f}^{*(i+1)} - \mathbf{g}^{*(i+1)}, \tag{18}$$

where the subscript i represents the i-th iteration. $\mathbf{f}^*$ and $\mathbf{g}^*$ are the solution of subproblem $\mathbf{f}$ and $\mathbf{g}$, respectively.

By solving the aforementioned subproblems iteratively, the optimal filter $\mathbf{f}^*$ of the t-th frame can be obtained and then used for tracking at $(t+1)$-th frame.

Target localization

The response map $\mathbf{R}_t$ at the t-th frame in Fourier domain can be calculated as,

$$\widehat{\mathbf{R}}_t = \sum_{c=1}^{C} \widehat{\mathbf{x}}_t^c \odot \widehat{\mathbf{f}}_{t-1}^{*c}. \tag{19}$$

After computing the IDFT on $\widehat{\mathbf{R}}$ to obtain the response map $\mathbf{R}_t$, the location can be predicted based on the maximum value of the response map. The overall tracking algorithm of the STAR model is summarized in Algorithm 1.

Algorithm 1: Overall tracking algorithm of the STAR model.

```
Input: Initial the target state (i.e., position $p_1$ and scale $s_1$) at the
       first frame.
Output: Target state at frame $t$.
Initialize the hyperparameters in STAR.
for $t = 1$ : end do
    Training
    1. Extract multi-channel feature map $\mathbf{x}_t$.
    2. Construct the spatial, temporal and aberrance suppressed
    regularizer using Eq. (5), Eq. (8) and Eq. (9), respectively.
    3. Optimize the filter model $\mathbf{f}_t$ at the $t$-th frame via Eq. (16),
    Eq. (17) and Eq. (18) for $N$ iterations.
    Detecting
    1. Crop multi-scale search regions centered at $p_t$ with $S$
    scales based on the bounding box at frame $t$.
    2. Extract multi-channel feature map $\mathbf{x}_{t+1}$.
    3. Compute response maps $\mathbf{R}_r$, $(r = 1, 2, \ldots, S)$ using
    Eq. (19).
    4. Estimate the target bounding box with the center position
    $p_{t+1}$ and scale $s_{t+1}$, based on the maximum value of
    response maps.
end
```

Experimental results

Evaluation metrics

Quantitative and qualitative experiments are conducted on four tracking benchmarks, i.e., TC128 [29], OTB2013 [43], OTB2015 [44] and UAV123 [33]. For these benchmarks, success rate and precision are utilized under the rule of one pass evaluation (OPE) [43,44]. The AREA UNDER CURVE (AUC) in the success rate and the distance precision (DP) at a threshold of 20 pixels in the precision are adopted as the evaluation metrics to measure the tracking accuracy. Meanwhile, the speed is measured in frames per second (FPS). For the sake of fair comparison, the compared trackers are based on publicly available code or results reported in the original paper.

Experimental setup

The experiments are conducted on a PC equipped with i7-9700K CPU and NVIDIA GTX 1080Ti GPU using MATLAB R2017a and MatConvNet toolbox.[1] We combine the output of Conv-3 layer from VGG-M network [36] with HOG+CN features for target representation. The values of spatial, temporal and aberrance suppressed regularizer are set as $\lambda = 1$, $\mu = 10$ and $\eta = 0.1$, respectively. The step size parameter γ is initialized to 1 and updated by $\gamma^{i+1} = \min\left(\gamma_{\max},\ \rho\gamma^{i}\right)$, (where $\rho = 10$, $\gamma_{\max} = 1000$). Other hyper-parameters are set to $\delta = 0.1$ and $\zeta = 0.7$, and the ADMM iteration is set to $N = 3$. To make a fair comparison, the parameters of the STAR tracker are fixed throughout the experiments.

Quantitative evaluation

Evaluation on TC128

The TC128 benchmark [29] contains 128 challenging color sequences. We compare the proposed STAR tracker with some SOTA DCF-based trackers, e.g., MCCT [41], LADCF-HC [45], MCCT-HC [41], STRCF [25], ECO-HC [14], CFWCR [20], MCPF [46], UDT+ [42], ARCF [23], UDT [42], AutoTrack [28], STRAPLE_CA [34], ARCF-H [23], DR2Track [17], BACF [18], TB-BiCF [30], RSST [47] and fDSST [15]. The success and precision plots of the evaluated trackers are depicted in Fig. 4 and the comparative results of the evaluated trackers in accuracy and speed are shown in Table 1. It shows that the STAR obtains the scores of 0.582 and 0.780 in AUC and DP, which outperform all the compared trackers. Specifically, compared with STRCF [41] which only adopts spatio-temporal regularization, the STAR increases the AUC and DP by 3.4 and 3.6%. Compared with ARCF [42] which only applies the aberrance suppressed strategy, the STAR gains an increase of 6.3 and 7.7% in AUC and DP. The performance improvement can be attributed to the effect of the dynamic spatio-temporal and the aberrance suppressed regularizer. In addition, the STAR runs at a speed of 10.6 fps, which is competitive compared with other deep-based trackers, i.e., RSST (1.5 fps), UDT+ (19.8 fps), MCPF (0.5 fps), CFWCR (10.2 fps) and MCCT (2.7 fps).

Evaluation on OTB2013 and OTB2015

The OTB2013 and OTB2015 are two popular tracking benchmarks, which consist of 50 and 100 video sequences, respectively. We compare the proposed STAR with several representative trackers, including ECO [14], DeepSTRCF [25], STRCF [25], LADCF-HC [45], CFWCR [20], MCCT-HC [41], BACF [18], ECO-HC [14], UDT [42], ARCF [23], ARCF-H [23], UDT+ [42], AutoTrack [28], STAPLE_CA [34], TB-BiCF [30], fDSST [15], RSST [47] and DR2Track [17]. The overall comparison results on OTB2013 [43] and OTB2015 [44] are presented in Fig. 5.

On the OTB2013 benchmark, the proposed STAR archives the best AUC (0.688) and the second-best DP (0.892). Compared with the feature selection-based tracker, i.e., LADCF-HC, the STAR improves the AUC and DP by 1.6 and 2.8%, respectively. Compared with UDT, which is trained in an unsupervised manner, the STAR improves by 6.1 and 6.6% in AUC and DP, respectively.

On the OTB2015 benchmark, the proposed STAR achieves the score of 0.672 and 0.875 in AUC and DP, respectively. Compared with the BACF tracker that uses the constant spatial regularizer, the STAR improves the AUC by 5.7% and the DP by 5.9%. This is mainly benefited from the dynamic spatial regularizer, which can impose different penalties on the spatial position based on the value of response variation rate, and produces more reliable filter coefficients at the tracking stage.

Evaluation on UAV123

Compared with the generic object tracking, UAV-based tracking is to locate a certain target from a low-altitude aerial perspective, which poses new challenges, e.g., rapid changes in scale and perspective, limited pixels in the target region, and multiple similar disruptors [48]. The compared trackers include CFWCR [20], DeepSTRCF [25], UDT+ [42], ECO-HC [14], LADCF-HC [45], STRCF [25], UDT [42], AutoTrack [28], TB-BiCF [30], ARCF [23], RSST [47], DR2Track [17], BACF [18], MCCT-H [41], ARCF-H [23], STAPLE_CA [34] and fDSST [15]. The comparative results are presented in Fig. 6. It shows that the STAR ranks first and third place in AUC (0.516) and DP (0.723), respectively.

[1] https://www.vlfeat.org/matconvnet/.

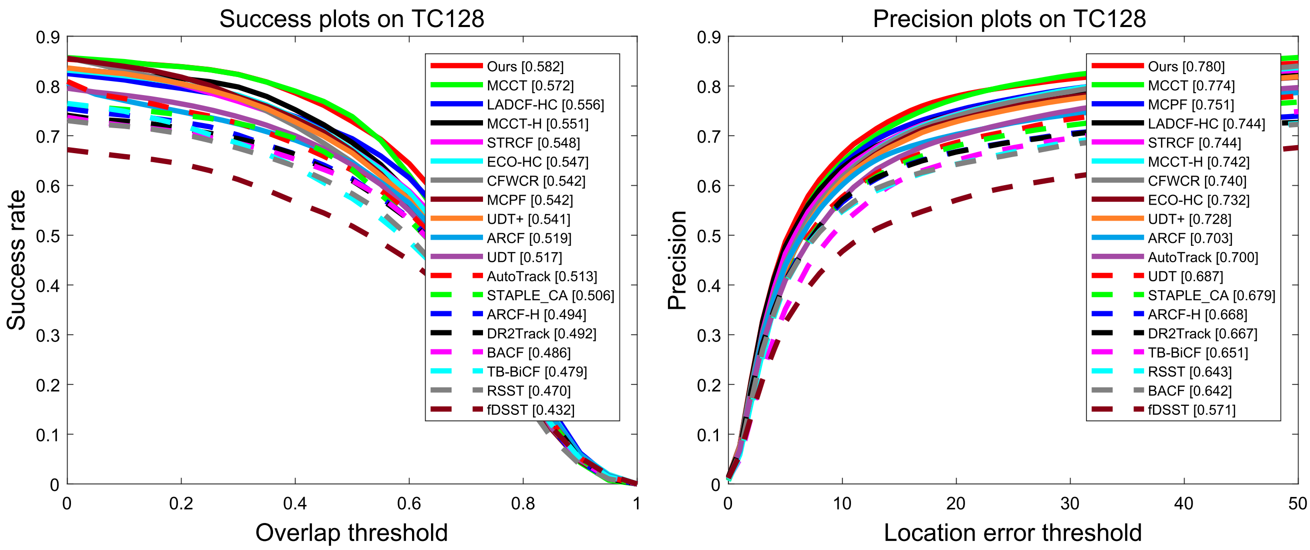

Fig. 4 Success and precision plots of the evaluated trackers on TC128

Table 1 Comparative results of the evaluated trackers on TC128 in accuracy and speed

Trackers	fDSST	RSST	TB-BiCF	BACF	DR2Track	ARCF-H	STAPLE_CA	AutoTrack	UDT	ARCF
AUC	0.432	0.470	0.479	0.486	0.492	0.494	0.506	0.513	0.517	0.519
DP	0.571	0.643	0.651	0.642	0.667	0.668	0.679	0.700	0.687	0.703
FPS	130.0	1.5*	49.2	36.4	55.8	51.4	52.4	34.1	14.9	32.0

Trackers	UDT+	MCPF	CFWCR	ECO-HC	STRCF	MCCT-H	LADCF-HC	MCCT	Ours
AUC	0.541	0.542	0.542	0.547	0.548	0.551	0.556	0.572	0.582
DP	0.728	0.751	0.740	0.732	0.744	0.742	0.744	0.774	0.780
FPS	19.8*	0.50*	10.2*	60.5	20.6	43.2	21.6	2.7*	10.6*

Note that the number with * indicates the speed of running on the GPU

Compared with other DCF-based trackers, e.g., ECO-HC, AutoTrack and DR2Track, STAR increases by 2.3, 4.0 and 5.7% in AUC, and 1.5, 3.3, and 6.1% in DP, respectively. Compared with DeepSTRCF that adopts the spatio-temporal regularization and multi-features (CNN+HOG+CN), STAR increases the AUC and DP by 0.8 and 1.8%, respectively. This can be attributed to the dynamic spatio-temporal regularizer, which can effectively alleviate the boundary effect and filter degradation, and provide a robust appearance model.

Table 2 Ablation studies of the critical components in STAR on OTB2013

Trackers	AUC	DP	FPS
Baseline	0.642	0.841	**8.27**
Baseline + DSR	0.670	0.873	6.82
Baseline + DTR	0.656	0.857	7.68
Baseline + AR	0.663	0.865	7.14
Baseline + DSR + DTR + AR	**0.688**	**0.892**	5.55

The best results are shown in bold

Attribute evaluation

To analyze the abilities of handling different challenges, attribute-based evaluations are performed. There are 12 attributes on UAV123 benchmark, i.e., occlusion (POC), full occlusion (FOC), fast motion (FM), illumination variation (IV), aspect ratio change (ARC), similar object (SOB), scale variation (SV), out-of-view (OV), background clutter (BC), viewpoint change (VC), camera motion (CM) and low resolution (LR). The success and precision plots of the evaluated trackers under these challenging attributes are presented in Figs. 7 and 8, respectively. It can be seen that the proposed STAR achieves the best AUC on several attributes, including POC (0.444), CM (0.511), ARC (0.454), VC (0.483), OV (0.436) and FM (0.419). Meanwhile, the proposed tracker achieves the best DP of 0.667, 0.626 and 0.654 in terms of OC, OV and FM, respectively.

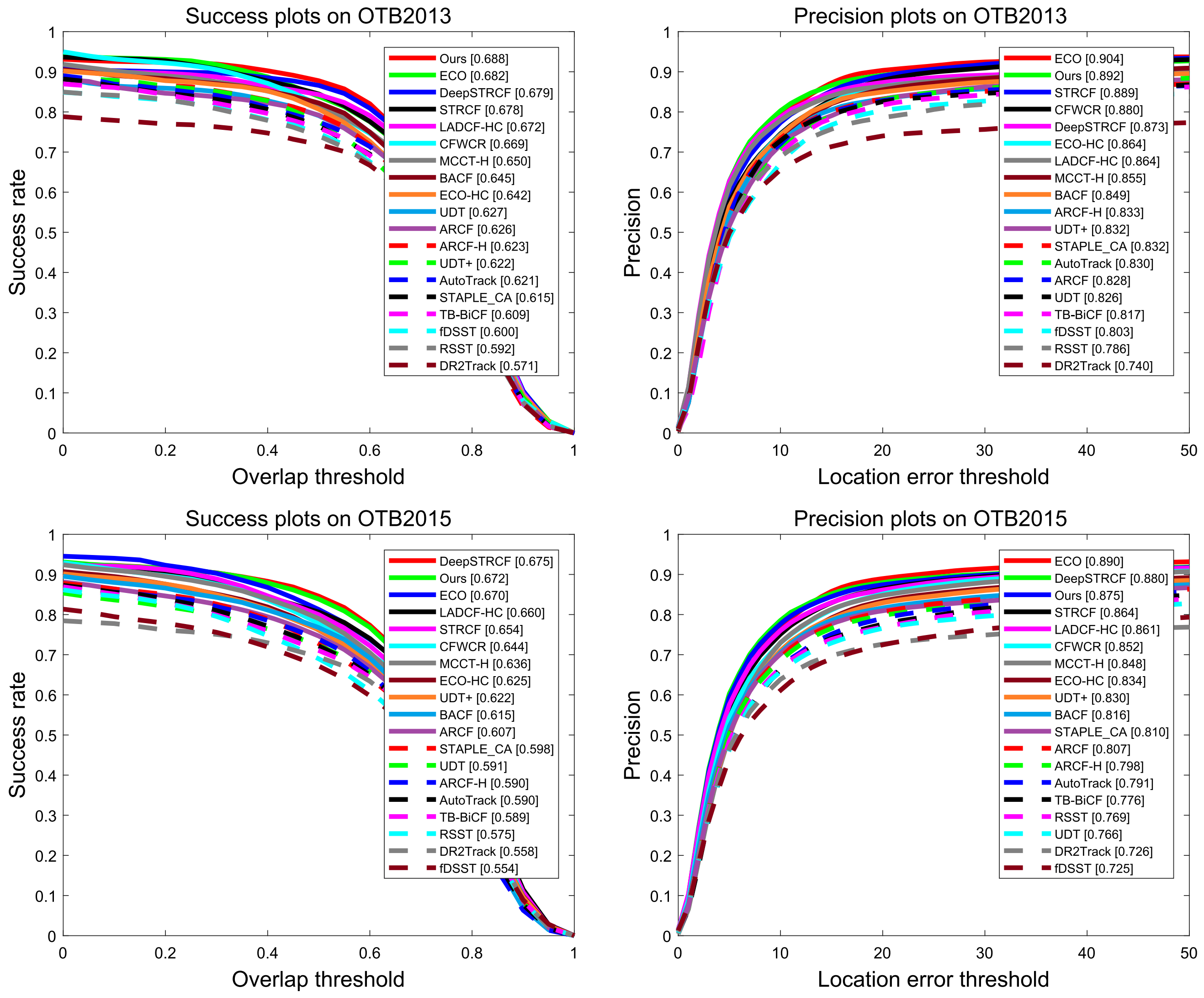

Fig. 5 Success and precision plots of the evaluated trackers on OTB2013 and OTB2015

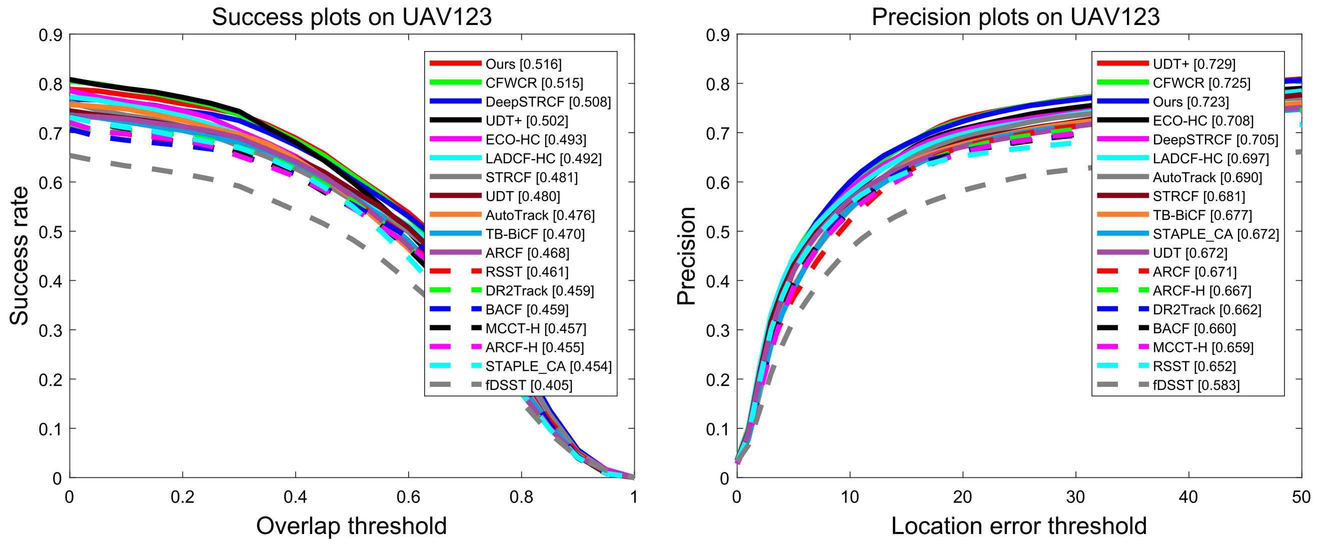

Fig. 6 Success and precision plots of the evaluated trackers on UAV123

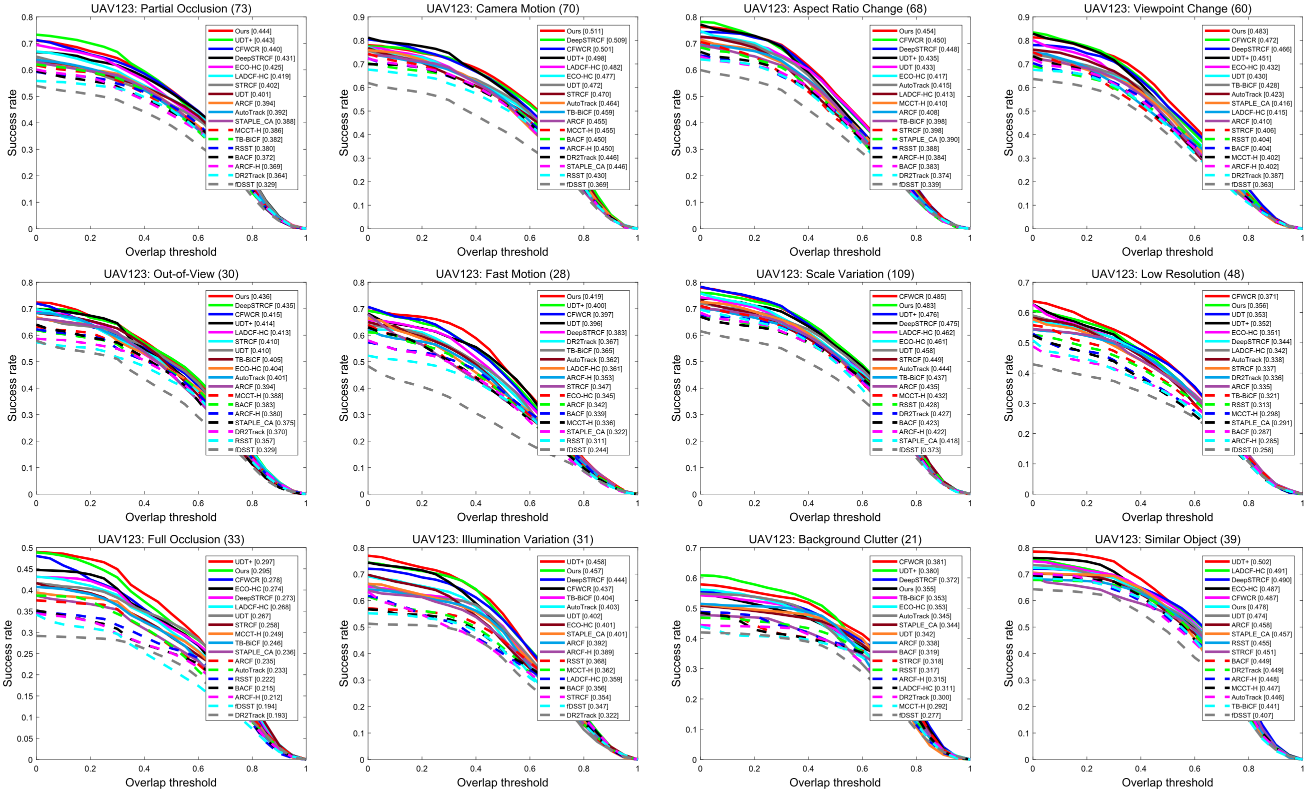

Fig. 7 Success plots of the evaluated trackers under different attributes on UAV123

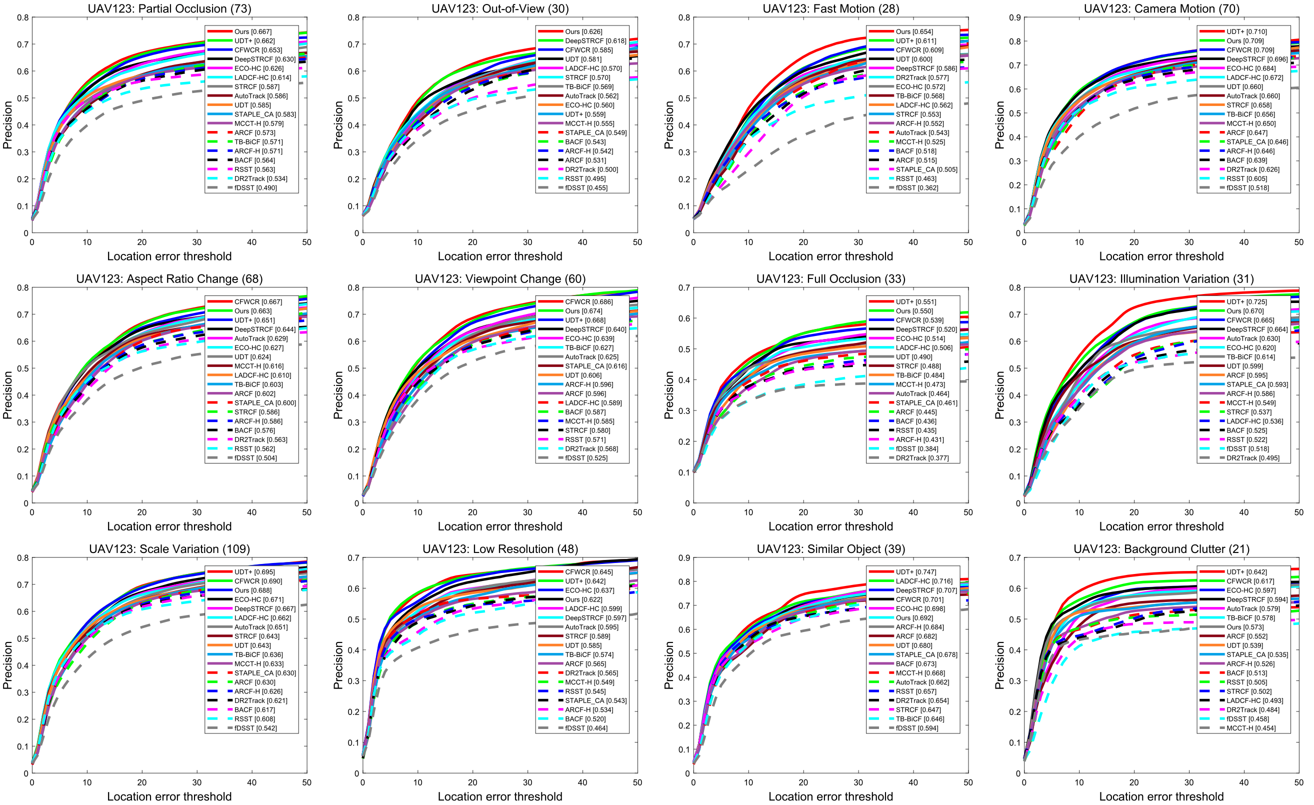

Fig. 8 Precision plots of the evaluated trackers under different attributes on UAV123

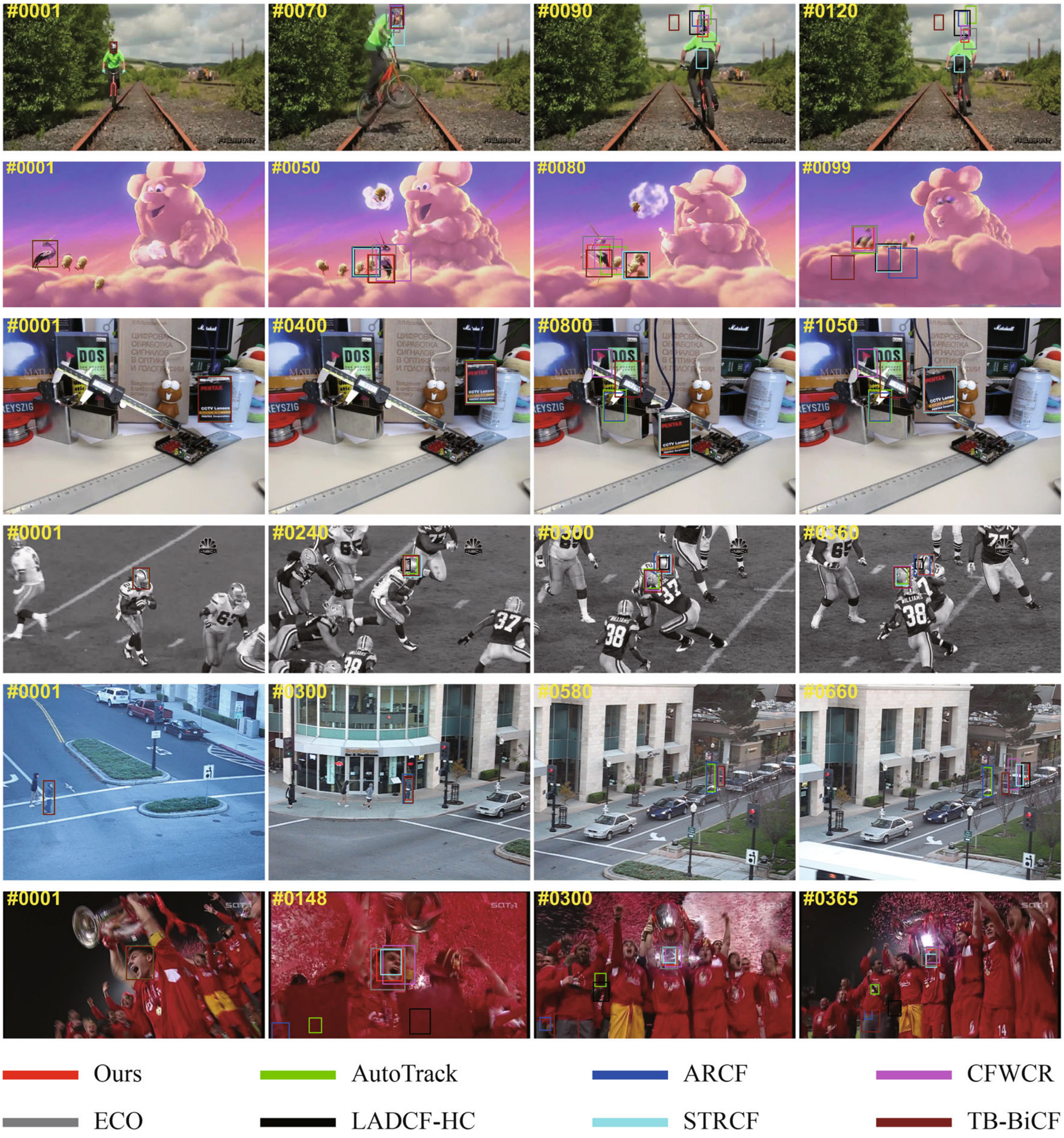

Fig. 9 Qualitative evaluations of the trackers on 6 image sequences from OTB2015. (Note, from top to bottom is biker, bird2, box, football, human4 and soccer. The indices of the frames are shown in the top-left of sub-figures)

Ablation studies

Ablation studies on OTB2013 [43] are conducted to demonstrate the effectiveness of the key components in the proposed STAR tracker. The key components include the dynamic spatial regularizer (DSR), dynamic temporal regularizer (DTR) and aberrance suppressed regularizer (AR). We compare the baseline with four variants, i.e., "Baseline" (the standard DCF tracker in "Revisit the standard DCF" which adopts the same feature representation as in STAR), "Baseline+DSR", "Baseline+DTR", "Baseline+AR" and "Baseline+DSR+DTR+AR" (i.e., the final STAR tracker). The ablation results are reported in Table 2. It shows that the baseline tracker achieves the score of 0.642 and 0.841 in AUC and DP. When the components of "DTR", "AR" and "DSR" are introduced into the "Baseline", they can improve the tracking performance gradually. Finally, the proposed STAR which integrates all the key components surpasses the "Baseline" by 4.6 and 5.1% in AUC and DP, respectively.

Qualitative evaluations

To intuitively exhibit the superiority of the STAR tracker, six sets of screenshots of the tracking results from OTB2015, i.e., biker, bird2, box, football, human4 and soccer(from top to bottom) are shown in Fig. 9. The target in these sequences undergoes challenging attributes such as rotation, scale variation, occlusion, motion blur, and fast motion. The compared trackers include AutoTrack [23], ARCF [23], CFWCR [20], ECO [14], LADCF-HC [45], STRCF [25] and TB-BiCF [30]. It shows that the proposed STAR (in red box) achieves much better tracking precision compared with other SOTA trackers. Specifically, in the "biker" sequence in which the target suffers from fast motion and motion blur, most of the compared trackers fail at frame 70. The attributes of "soccer" sequences include occlusion and background cluster, causing most compared trackers to fail at frame 365. In contrast, the proposed STAR achieves satisfying performance in these sequences.

Conclusion

In this paper, we propose a novel spatio-temporal joint aberrance suppressed regularization (STAR) correlation filter for robust visual tracking. The STAR tracker takes full advantage of spatio-temporal information and employs aberrance suppressed strategy. The dynamic spatio-temporal regularizer can effectively alleviate boundary effect and filter degradation, while the aberrance suppressed strategy reduces the interference caused by background cluster. Besides, the STAR tracker is efficiently optimized based on the ADMM formulation. Comprehensive experiments on four tracking benchmarks demonstrate the superiority of the proposed method against the SOTA trackers.

Acknowledgements This work is partially supported by the National Natural Science Foundation of China (No. 61801272).

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Adam A, Rivlin E, Shimshoni I (2006) Robust fragments-based tracking using the integral histogram. Proc IEEE Conf Comput Vis Pattern Recognit 1:798–805. https://doi.org/10.1109/CVPR.2006.256
2. Ben X, Gong C, Zhang P, Jia X, Wu Q, Meng W (2019) Coupled patch alignment for matching cross-view gaits. IEEE Trans Image Process 28:3142–3157. https://doi.org/10.1109/TIP.2019.2894362
3. Bertinetto L, Valmadre J, Golodetz S, Miksik O, Torr PHS (2016) Staple: complementary learners for real-time tracking. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1401–1409. https://doi.org/10.1109/CVPR.2016.156
4. Bertinetto L, Valmadre J, Henriques JF, Vedaldi A, Torr PHS (2016) Fully-convolutional siamese networks for object tracking. In: Proceedings of the European conference on computer vision workshops, pp 850–865. https://doi.org/10.1007/978-3-319-48881-3_56
5. Bolme DS, Beveridge JR, Draper BA, Lui YM (2010) Visual object tracking using adaptive correlation filters. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2544–2550. https://doi.org/10.1109/CVPR.2010.5539960
6. Boyd S, Parikh N, Chu E, Peleato B, Eckstein J (2010) Distributed optimization and statistical learning via the alternating direction method of multipliers. Found Trends Mach Learn 3:1–122
7. Choi J, Chang HJ, Yun S, Fischer T, Demiris Y, Choi JY (2017) Attentional correlation filter network for adaptive visual tracking. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4828–4837. https://doi.org/10.1109/CVPR.2017.513
8. Comaniciu D, Ramesh V, Meer P (2000) Real-time tracking of non-rigid objects using mean shift. Proc IEEE Conf Comput Vis Pattern Recognit 2:142–149. https://doi.org/10.1109/CVPR.2000.854761
9. Dalal N, Triggs B (2005) Histograms of oriented gradients for human detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition, vol 1, pp 886–893. https://doi.org/10.1109/CVPR.2005.177
10. Danelljan M, Häger G, Khan F, Felsberg M (2014) Accurate scale estimation for robust visual tracking. In: Proceedings of the British machine vision conference. https://doi.org/10.5244/C.28.65
11. Danelljan M, Khan FS, Felsberg M, Van De Weijer J (2014) Adaptive color attributes for real-time visual tracking. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1090–1097. https://doi.org/10.1109/CVPR.2014.143
12. Danelljan M, Häger G, Khan FS, Felsberg M (2015a) Convolutional features for correlation filter based visual tracking. In: Proceedings of the international conference on computer vision workshops, pp 621–629. https://doi.org/10.1109/ICCVW.2015.84
13. Danelljan M, Häger G, Khan FS, Felsberg M (2015b) Learning spatially regularized correlation filters for visual tracking. In: Proceedings of the international conference on computer vision, pp 4310–4318. https://doi.org/10.1109/ICCV.2015.490
14. Danelljan M, Bhat G, Khan FS, Felsberg M (2017a) Eco: Efficient convolution operators for tracking. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 6931–6939. https://doi.org/10.1109/CVPR.2017.733
15. Danelljan M, Häger G, Khan FS, Felsberg M (2017b) Discriminative scale space tracking. IEEE Trans Pattern Anal Mach Intell 39(8):1561–1575. https://doi.org/10.1109/TPAMI.2016.2609928
16. Fiaz M, Mahmood A, Javed S, Jung SK (2019) Handcrafted and deep trackers: recent visual object tracking approaches and trends. ACM Comput Surv. https://doi.org/10.1145/3309665

17. Fu C, Ding F, Li Y, Jin J, Feng C (2021) Learning dynamic regression with automatic distractor repression for real-time UAV tracking. Eng Appl Artif Intell 98:104116. https://doi.org/10.1016/j.engappai.2020.104116
18. Galoogahi HK, Fagg A, Lucey S (2017) Learning background-aware correlation filters for visual tracking. In: Proceedings of the international conference on computer vision, pp 1144–1152. https://doi.org/10.1109/ICCV.2017.129
19. Grabner H, Grabner M, Bischof H (2006) Real-time tracking via on-line boosting. In: Proceedings of the British machine vision conference
20. He Z, Fan Y, Zhuang J, Dong Y, Bai H (2017) Correlation filters with weighted convolution responses. In: Proceedings of the international conference on computer vision workshops, pp 1992–2000. https://doi.org/10.1109/ICCVW.2017.233
21. Henriques JF, Caseiro R, Martins P, Batista J (2012) Exploiting the circulant structure of tracking-by-detection with kernels. In: Proceedings of the European conference on computer vision, pp 702–715. https://doi.org/10.1007/978-3-642-33765-9_50
22. Henriques JF, Caseiro R, Martins P, Batista J (2015) High-speed tracking with kernelized correlation filters. IEEE Trans Pattern Anal Mach Intell 37(3):583–596. https://doi.org/10.1109/TPAMI.2014.2345390
23. Huang Z, Fu C, Li Y, Lin F, Lu P (2019) Learning aberrance repressed correlation filters for real-time UAV tracking. In: Proceedings of the international conference on computer vision, pp 2891–2900. https://doi.org/10.1109/ICCV.2019.00298
24. Kalal Z, Mikolajczyk K, Matas J (2012) Tracking-learning-detection. IEEE Trans Pattern Anal Mach Intell 34(7):1409–1422. https://doi.org/10.1109/TPAMI.2011.239
25. Li F, Tian C, Zuo W, Zhang L, Yang M (2018) Learning spatial-temporal regularized correlation filters for visual tracking. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4904–4913. https://doi.org/10.1109/CVPR.2018.00515
26. Li P, Wang D, Wang L, Lu H (2018) Deep visual tracking: review and experimental comparison. Pattern Recognit 76:323–338. https://doi.org/10.1016/j.patcog.2017.11.007
27. Li Y, Zhu J (2015) A scale adaptive kernel correlation filter tracker with feature integration. In: Proceedings of the European conference on computer vision workshops, pp 254–265. https://doi.org/10.1007/978-3-319-16181-5_18
28. Li Y, Fu C, Ding F, Huang Z, Lu G (2020) Autotrack: towards high-performance visual tracking for UAV with automatic spatio-temporal regularization. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 11920–11929. https://doi.org/10.1109/CVPR42600.2020.01194
29. Liang P, Blasch E, Ling H (2015) Encoding color information for visual tracking: algorithms and benchmark. IEEE Trans Image Process 24(12):5630–5644. https://doi.org/10.1109/TIP.2015.2482905
30. Lin F, Fu C, He Y, Guo F, Tang Q (2021) Learning temporary block-based bidirectional incongruity-aware correlation filters for efficient UAV object tracking. IEEE Trans Circuits Syst Video Technol 31(6):2160–2174. https://doi.org/10.1109/TCSVT.2020.3023440
31. Marvasti-Zadeh SM, Cheng L, Ghanei-Yakhdan H, Kasaei S (2021) Deep learning for visual tracking: a comprehensive survey. IEEE Trans Intell Transp Syst. https://doi.org/10.1109/TITS.2020.3046478
32. Morrison SWJ (1950) Adjustment of an inverse matrix corresponding to a change in one element of a given matrix. Ann Math Stat 21(1):124–127. https://doi.org/10.2307/2236561
33. Mueller M, Smith N, Ghanem B (2016) A benchmark and simulator for UAV tracking. In: Proceedings of the European conference on computer vision, pp 445–461. https://doi.org/10.1007/978-3-319-46448-0_27
34. Mueller M, Smith N, Ghanem B (2017) Context-aware correlation filter tracking. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1387–1395. https://doi.org/10.1109/CVPR.2017.152
35. Ross DA, Lim J, Lin RS, Yang MH (2008) Incremental learning for robust visual tracking. Int J Comput Vis 77(1):125–141. https://doi.org/10.1007/s11263-007-0075-7
36. Simonyan K, Zisserman A (2015) Very deep convolutional networks for large-scale image recognition. In: Proceedings of the international conference on learning representations
37. Valmadre J, Bertinetto L, Henriques J, Vedaldi A, Torr PHS (2017) End-to-end representation learning for correlation filter based tracking. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 5000–5008. https://doi.org/10.1109/CVPR.2017.531
38. van de Weijer J, Schmid C, Verbeek J, Larlus D (2009) Learning color names for real-world applications. IEEE Trans Image Process 18(7):1512–1523. https://doi.org/10.1109/TIP.2009.2019809
39. Wang M, Liu Y, Huang Z (2017) Large margin object tracking with circulant feature maps. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4800–4808. https://doi.org/10.1109/CVPR.2017.510
40. Wang N, Shi J, Yeung D, Jia J (2015) Understanding and diagnosing visual tracking systems. In: Proceedings of the international conference on computer vision, pp 3101–3109. https://doi.org/10.1109/ICCV.2015.355
41. Wang N, Zhou W, Tian Q, Hong R, Wang M, Li H (2018) Multi-cue correlation filters for robust visual tracking. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4844–4853. https://doi.org/10.1109/CVPR.2018.00509
42. Wang N, Song Y, Ma C, Zhou W, Liu W, Li H (2019) Unsupervised deep tracking. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1308–1317. https://doi.org/10.1109/CVPR.2019.00140
43. Wu Y, Lim J, Yang M (2013) Online object tracking: a benchmark. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2411–2418. https://doi.org/10.1109/CVPR.2013.312
44. Wu Y, Lim J, Yang M (2015) Object tracking benchmark. IEEE Trans Pattern Anal Mach Intell 37(9):1834–1848. https://doi.org/10.1109/TPAMI.2014.2388226
45. Xu T, Feng ZH, Wu XJ, Kittler J (2019) 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. https://doi.org/10.1109/TIP.2019.2919201
46. Zhang T, Xu C, Yang M (2017) Multi-task correlation particle filter for robust object tracking. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4819–4827. https://doi.org/10.1109/CVPR.2017.512
47. Zhang T, Xu C, Yang M (2019) Robust structural sparse tracking. IEEE Trans Pattern Anal Mach Intell 41(2):473–486. https://doi.org/10.1109/TPAMI.2018.2797082
48. Zhu P, Wen L, Du D, Bian X, Hu Q, Ling H (2020) Vision meets drones: past, present and future. arXiv:2001.06303