Signal Processing 195 (2022) 108463

Contents lists available at ScienceDirect

Signal Processing

journal homepage: www.elsevier.com/locate/sigpro

Robust DCF object tracking with adaptive spatial and temporal regularization based on target appearance variation

Lin Zhou[a,b], Yong Jin[a,b,*], Han Wang[a,b], Zhentao Hu[a,b], Shuaipeng Zhao[a,b]

[a] School of Artificial Intelligence, Henan University, Zhengzhou, 450046, China
[b] School of Computer and Information Engineering, Henan University, Kaifeng 475004, China

ARTICLE INFO

Article history:
Received 8 May 2021
Revised 1 January 2022
Accepted 16 January 2022
Available online 17 January 2022

Keywords:
Correlation filter
Object tracking
Adaptive spatial-temporal regularization
Appearance variation

ABSTRACT

To alleviate the boundary effect and constrain the update of the tracking model, current object tracking methods based on Discriminative Correlation Filter (DCF) usually introduce spatial and temporal regularization constraints in the filter training objective function. However, these regularization constraints with fixed coefficients greatly limit the adaptability of the tracker with respect to target appearance variation. This paper proposes a spatial-temporal regularization model based on the real-time target appearance variation for the filter training, improving the adaptability of the filter related to target appearance variation. Moreover, the filter training objective function with the adaptive spatial-temporal regularization is proposed to enhance the robustness of the filter. Finally, an iterative optimization method based on the alternating direction method of multipliers (ADMM) is proposed to update the filter, and the convergence proof of the optimization method is also presented. Comparison experiments with some representative trackers including ASRCF,ARCF, CSR_CF, DSAR_CF and SSR_CF etc. on OTB2015, UAV123 and LaSOT databases show that the proposed algorithm effectively improves the tracking accuracy.

1. Introduction

In today's Internet of things and the "Internet +" era, services such as scene monitoring, autonomous driving, drone navigation, and human-computer interaction based on computer vision object tracking technology are booming. Therefore, video object tracking methods have attracted extensive attention and in-depth research from academia and industry. The task of video object tracking is to predict the state of a single target in a video sequence according to the initial position and scale information of the target in the first frame. However, some factors including deformation, occlusion, illumination variation and background cluttered etc., caused by complex environment make visual object tracking challenging. Therefore, improving the tracker's robustness to the interference and the adaptability to target variation is of great significance.

In recent years, many trackers have been proposed to improve the tracking performance. Existing approaches include: template matching [1], statistical learning [2], particle filter [3], subspace learning [4], discriminant correlation filter [5], tracking method based on deep convolutional neural networks (CNNs) [6] and Siamese networks [7–9], etc. Due to the outstanding tracking accuracy and higher tracking speed, the discriminative correlation filter (DCF) method attracts the attention of many researchers [10–13]. The DCF method was firstly applied in the field of visual tracking by Bolme et al. in the minimum output sum of squares error (MOSSE) tracker [5]. The CSK (Circulant Structure of Tracking-by-detection with Kernels)tracker [14] introduced the kernel trick and the circulant matrix on the basis of MOSSE. By performing cyclic shift operation on the search area, the CSK tracker greatly enriched the training samples of the filter and improved the tracking accuracy. However, CSK only used gray information as the sample feature, leading to poor tracking robustness.

To improve the robustness of tracking model, some works introduced high-dimensional features including histogram of oriented gradient(HOG) [15] and Color Names [16] as the representation of the target appearance [17,18]. The CSR_DCF [19] adaptively fused each channel in the HOG feature according the channel filter response to improve the tracking robustness. With the development of deep network technology [20,21], deep features and networks have been widely used in DCF-based tracking methods [22–26], too. For example, the ATOM [24] and DiMP [25] constructed a powerful scale estimation model to deal with target rotation and viewpoint change using an offline-trained IOU-Net [27]. Some trackers [28–30] are also proposed by using offline-trained network to tackled great deformation induced by long-time span videos. Since deep features are invariant to small changes of tar-

* Corresponding author at: School of Artificial Intelligence, Henan University, Zhengzhou, 450046, China
E-mail address: jy@henu.edu.cn (Y. Jin).

https://doi.org/10.1016/j.sigpro.2022.108463
0165-1684/

get appearance, they can greatly improve the tracking robustness. However, the offline-trained network is difficult to be adjusted according to the real-time target variation, lacking the flexibility to adapt the dramatic target appearance variation.

1.1. Related work

DCF methods efficiently update the filter in frequency domain based on the periodic assumption of samples. But it will produce boundary effect on the edge of the sample images, which results in a large number of unreal training samples, damaging the tracking performance. On the other hand, temporal smoothness of the successive filter model is helpful to improve the robustness and generalization ability of the filter, thus enhancing overfitting resisting ability of the tracker. However, traditional temporal smoothness framework will persistently increase the computational complexity of the tracker. Meanwhile, the learning rate in the temporal smoothness framework is usually fixed, which severely restricts the tracker's adaptability following the target appearance changing.

To alleviate the boundary effect and enhance the adaptability of the tracking model, researchers have implemented the spatial and temporal regularization into the filter training objective function.

The spatial regularization

Traditional DCF trackers used the cosine window to alleviate the boundary effect, limiting the search area of the tracking model [5,17,31–33]. The CSR_DCF [19] introduced a spatial reliability map to identify pixels that should be ignored in filter learning. However, this spatial reliability map set to be fixed in the whole tracking process, thus cannot adapt the target appearance variation. Referencing human's visual attention mechanism, Danelljan et al. introduced a quadratic polynomial spatial regularization matrix in the filter training objective function of the spatially regularized discriminative correlation filter (SRDCF) [34]. This spatial regularization matrix punishs the filter model coefficients corresponding to the boundary region, effectively enhancing the discrimination of the tracker. Galoogahi et al. proposed the background-aware correlation filter (BACF), which not only employed a binary matrix into eliminate unreal negative samples, but also used the real positive and negative samples for filter training [35]. The selective spatial regularization(SSR-CF) [36] tracker constructed three different spatial weight maps and used a selector to determine which spatial weight map to use. However, above-mentioned trackers all use spatial constraint matrixes with fixed coefficients or several spatial weight maps, which are less adaptable to the target whose appearance is constantly changing. To eliminate this defect, the adaptive spatially-regularized correlation filter (ASRCF) [37] introduced the regularization constraint of the coefficients of the spatial constraint matrix. Unfortunately, the penalty coefficient of this regularization is fixed, and the adaptability of the spatial constraint matrix with respect to the target deformation is still lacking, leading target missing in the scene of target severe deformation. The DSARCF [38] tracker dynamically varied the spatial regularization weight map by considering both the saliency map and the response map. However, the saliency map merely captures object shape and size variation, while ignores other variation of the target appearance including the color, texture, and illumination variation that are essential for target tracking.

The temporal regularization

Most DCF trackers implement temporal smoothness by updating the filter and target appearance model in a moving average scheme [17,22,34,37,39]. However, this scheme will persistently increase the training sample space and the computational complexity of the tracker. For this reason, the deep spatial-temporal regularized correlation filter (DeepSTRCF) instead implemented the temporal smoothness by introducing the temporal regularization into the objective function [40]. The temporal regularization restricts the filter training into a single frame, effectively avoiding the sample space expansion. Meanwhile, the aberrance repressed correlation filter (ARCF) introduced the temporal regular term of the filter response into the objective function to enhance tracking robustness [41]. However, the regularization coefficients of the trackers above are fixed in the whole tracking process. Although it allows the tracker to make good use of the historical information of the target appearance and ensure its robustness, it limits the tracker's adaptability with respect to target appearance changing.

To sum up, spatial and temporal regularization constraints with fixed coefficients are often used in the objective function of current tracking methods. The regularization constraints can reduce the influence of boundary effect and enhance the robustness of the tracking model. However, they greatly limit the adaptability of tracker with respect to target appearance changing. In situations where the appearance of the target changes greatly, the tracker is prone to lose the target.

1.2. Contributions

To maintain the tracking robustness and improve the tracker's adaptability as far as possible, this paper embeds an adaptive spatial-temporal regularization constraint based on target appearance variation into the filter training objective function. The key innovations of the proposed method are listed as follows,

- This paper constructs an adaptive spatial-temporal regularization model based on the target appearance variation. Different from the regularization with fixed coefficients in traditional trackers, the proposed method can adaptively adjust the regularization coefficients according to the real-time variation of target appearance, thus can significantly improve the adaptability and the robustness of the tracker.
- This paper proposes a novel filter training objective function under the proposed adaptive spatial-temporal regularization. And an iterative objective function solving algorithm based on the alternating direction method (ADMM) is also proposed to improve the tracking accuracy while guaranteeing tracking speed.
- This paper presents extensive experiments on OTB2015, UAV123 and LaSOT datasets to verify the superiority of our method over state-of-the-art competitors that employ the spatial regularization or temporal regularization technology, such as ASRCF, ARCF, STRCF and so on.

1.3. Paper organization

The rest of this paper is organized as follows: Section 2 briefly gives the filter training constraint function in traditional DCF methods and its shortcomings; Section 3 proposes the adaptive spatial-temporal regularization model based on real-time target appearance variation, and proposes the filter training objective function with this regularization model. Furthermore, the iterative solution process of the objective function is derived in detail; Section 4 provides an overview of the proposed tracking algorithm; The analysis and experimental comparison results with some competing trackers are presented in Sections 5 and 6 gives some conclusions of this paper.

2. Traditional spatial-temporal regularization of filter training

Among the multi-channel feature DCF trackers, the earliest method introducing the spatial and temporal regularization constraints into filter training function is DeepSTRCF [40]. The filter

training objective function in this tracker is defined as,

$$\mathbf{f}_t = \arg\min_{\mathbf{f}_t} \frac{1}{2}\left\|\sum_{d=1}^{D} x_t^d * f_t^d - y_t\right\|_{l^2}^2 + \frac{1}{2}\sum_{d=1}^{D}\left\|w \cdot f_t^d\right\|_{l^2}^2 + \frac{\mu}{2}\|\mathbf{f}_t - \mathbf{f}_{t-1}\|_{l^2}^2 \quad (1)$$

The "$*$" represents the circular convolution operation, and the "$\cdot$" denotes the Hadamard product. The l^2-norm of the matrix $\mathbf{x}$ is defined as $\|\mathbf{x}\|_{l^2}^2 = \mathbf{x}^T\mathbf{x}$. The training sample $\mathbf{x}_t = \left(x_t^1, \ldots, x_t^D\right) \in \mathbb{R}^{W\times H\times D}$ is obtained by extracting sample features from the sample image patch centered at the estimated target position p_t in frame t. The size of the sample feature $\mathbf{x}_t$ is $W \times H$, and there are D number of feature channels in $\mathbf{x}_t$. The $y_t \in \mathbb{R}^{W\times H}$ is the label of the training sample $\mathbf{x}_t$, and it is a 2-D Gaussian function centered at the position of the target. The matrix $w \in \mathbb{R}^{W\times H}$ is the spatial regularization constraint, and the μ is a constant temporal regularization constraint, and the $\mathbf{f}_{t-1}$ is the filter obtained in frame $t-1$. The multi-channel filter $\mathbf{f}_t = \left(f_t^1, \ldots, f_t^D\right) \in \mathbb{R}^{W\times H\times D}$is updated in frame t by solving the above objective function. In frame t, the response score of the tracking sample $\mathbf{z}_t = \left(z_t^1, \ldots, z_t^D\right)$ on the filter is expressed as,

$$S(\mathbf{z}_t) = \sum_{d=1}^{D} z_t^d * f_{t-1}^d \quad (2)$$

where the tracking sample $\mathbf{z}_t$ is extracted in frame t centered at the previous target position p_{t-1}. The target in frame t is located at the position with the maximum response. It can be seen that the objective function in Eq. (1) consists of three parts, namely the least squares loss $\left\|\sum_{d=1}^{D} x_t^d * f_t^d - y_t\right\|_{l^2}^2$, the spatial regularization $\frac{1}{2}\sum_{d=1}^{D}\left\|w \cdot f_t^d\right\|_{l^2}^2$ and the temporal regularization constraint $\frac{\mu}{2}\|\mathbf{f}_t - \mathbf{f}_{t-1}\|_{l^2}^2$ weighted by a constant μ.

Given the training sample size $W \times H$, the spatial regularization $w \in \mathbb{R}^{W\times H}$ in $\frac{1}{2}\sum_{d=1}^{D}\left\|w \cdot f_t^d\right\|_{l^2}^2$ gets the minimum value at the center of the target and the larger value at the boundary of the sample patch. Therefore, it can suppress the filter coefficients corresponding to image boundary and make the filter pay more attention to the central area of the target, thus effectively alleviating the boundary effect. However, the spatial regularization w is fixed in the whole tracking process. Therefore, it cannot adapt to the persistent target appearance variation, and the filter $\mathbf{f}_t$ cannot fully capture the diversity of the target appearance change during the updating process.

On the other hand, in the temporal regularization $\frac{\mu}{2}\|\mathbf{f}_t - \mathbf{f}_{t-1}\|_{l^2}^2$ of the filter, the updated filter $\mathbf{f}_t$ in frame t is constrained to have some similarity with the filter model $\mathbf{f}_{t-1}$ in the previous frame $t-1$. The temporal regularization allows the tracker to make fully use of the historical information of the target appearance, so as to improve the tracking robustness and avoid filter degradation. However, the temporal regularization of the filter in Eq. (1) does not consider the real-time change rate of the target appearance, and its regularization coefficient remains fixed during the whole tracking process. This conservative time regularization constraint limits the adaptive ability of the filter with respect to the target appearance changing, and the trained filter $\mathbf{f}_t$ usually has less sensitivity with respect to the current state of the target when the target appearance changing violently, which leads to tracking failure.

3. Adaptive spatial-temporal regularization based on target appearance variation

In order to solve the problems in the traditional spatial and temporal regularization described in Section 2 and improve the tracker's adaptability, this section proposes an adaptive spatial-temporal regularization model based on the real-time target appearance variation, and constructs the filter training objective function related to this constraint model. Based on the ADMM, the iterative solution process of the filter model is derived. Fig. 1 shows the diagram of the filter training method with the proposed adaptive spatial-temporal regularization model. More details about the filter training method in Fig. 1 are discussed in the rest of this section.

3.1. Adaptive spatial-temporal regularization

Introducing the proposed adaptive spatial-temporal regularization constraint into the traditional filter training objective function, we adjust the Eq. (1) as,

$$\mathbf{f}_t = \arg\min_{\mathbf{f}_t} \frac{1}{2}\left\|\sum_{d=1}^{D} x_t^d * f_t^d - y_t\right\|_{l^2}^2 + \frac{\lambda_1}{2}\sum_{d=1}^{D}\left\|f_t^d\right\|_{l^2}^2 + \frac{\lambda_2}{2}R_{ST}(\mathbf{f}_t) \quad (3)$$

where the $R_{ST}(\mathbf{f}_t)$ represents the adaptive spatial-temporal regularization model in frame t based on the real-time target appearance variation. The $R_{ST}(\mathbf{f}_t)$ is expressed as,

$$R_{ST}(\mathbf{f}) = \sum_{d=1}^{D}\left\|(1+\psi(t)) \cdot w \cdot \left(f_t^d - f_{t-1}^d\right)\right\|_{l^2}^2 \quad (4)$$

where the spatial regularization constraint w defined on the training sample $\mathbf{x} \in \mathbb{R}^{W\times H\times D}$ is expressed as,

$$w(m,n) = \mu + \eta\left[(m/W)^2 + (n/H)^2\right] \quad (5)$$

where $m \in \left[-\frac{W}{2}, \frac{W}{2}-1\right]$, and $n \in \left[-\frac{H}{2}, \frac{H}{2}-1\right]$.

In Eq. (4), the $1+\psi(t) \in \mathbb{R}^{W\times H}$ is the temporal regularization constraint based on the real-time target appearance variation. And the proposed adaptive spatial-temporal regularization is constructed as the element-wise multiplication of the above spatial and temporal regularization matrixes.

For the training sample $\mathbf{x}_t$ obtained in frame t, its jth cross-channel element is denoted as $V_j(\mathbf{x}_t) = \left(x_t^1(j), \ldots, x_t^D(j)\right) \in \mathbb{R}^{D\times 1}$, thus there $N = W \times H$ number of cross-channel elements in $\mathbf{x}_t$ that correspond to different regions of the sample image. The jth element of the proposed adaptive temporal regularization $1+\psi(t)$ based on the real-time target appearance variation is defined as,

$$1+\psi(t)\{j\} = 1 + a \cdot e^{-b\left\|V_j(\mathbf{x}_t) - V_j\left(\mathbf{x}_{t-1}^*\right)\right\|_{l^2}^2}, j = 1, \ldots, N \quad (6)$$

where the a and b are constants, and the reference training sample $\mathbf{x}_{t-1}^*$ is obtained by the weighted average of samples $(\mathbf{x}_1, \ldots, \mathbf{x}_{t-1})$ in frame 1 to t. Therefore, the value of the $1+\psi(t)\{j\} \in \mathbb{R}$ is calculated by evaluating the change rate $\left\|V_j(\mathbf{x}_t) - V_j\left(\mathbf{x}_{t-1}^*\right)\right\|_{l^2}^2$ of the jth cross-channel element $V_j(\mathbf{x}_t)$ of $\mathbf{x}_t$, and it is the temporal regularization constraint of the jth cross-channel element $V_j(\mathbf{f}_t) = \left(f_t^1(j), \ldots, f_t^D(j)\right)$ of the filter. The values of $1+\psi(t)\{j\}$ and $\left\|V_j(\mathbf{x}_t) - V_j\left(\mathbf{x}_{t-1}^*\right)\right\|_{l^2}^2$ are negatively correlated. That is, when the value of $\left\|V_j(\mathbf{x}_t) - V_j\left(\mathbf{x}_{t-1}^*\right)\right\|_{l^2}^2$ is small, indicating that the target appearance in the jth region is relatively stable, the constraint value $1+\psi(t)\{j\}$ will be relatively large to let the filter update stably with the historical information of the target appearance. On the contrary, when the $\left\|V_j(\mathbf{x}_t) - V_j\left(\mathbf{x}_{t-1}^*\right)\right\|_{l^2}^2$ is considerably large, which reflects that the target appearance in the jth region changes violently, the value of $1+\psi(t)\{j\}$ will adaptively decrease to allow $V_j(\mathbf{f}_t)$ to update more significantly, ensuring the adaptability with respect to target appearance changing.

In conclusion, firstly, the spatial regularization w assigns different weight to each $V_j(\mathbf{f}_t)$ according to the position of $V_j(\mathbf{x}_t)$ relative to the target center, endowing the learned filter with more attention to the central region of the target. Secondly, the temporal

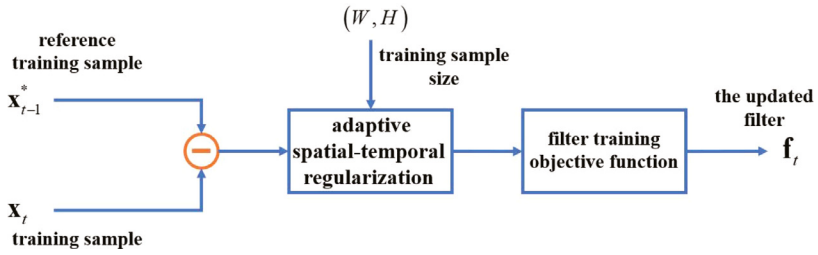


Fig. 1. Filter training with the adaptive spatial-temporal regularization model.

regularization $(1+\psi(t))$ further adjusts the weight to the $V_j(\mathbf{f}_t)$ based on the real-time change $\left\|V_j(\mathbf{x}_t)-V_j\left(\mathbf{x}_{t-1}^*\right)\right\|_{l^2}^2$ of $V_j(\mathbf{x}_t)$, enhancing the tracker's adaptability.

3.2. Filter training objective function with the proposed adaptive spatial-temporal regularization

Substituting Eq. (4) into Eq. (3), the filter training objective function with the proposed adaptive spatial-temporal regularization is described as,

$$\begin{aligned}\mathbf{f}_t = \arg\min_{\mathbf{f}_t} \frac{1}{2}\left\|\sum_{d=1}^{D} x_t^d * f_t^d - y_t\right\|_{l^2}^2 + \frac{\lambda_1}{2}\sum_{d=1}^{D}\left\|f_t^d\right\|_{l^2}^2 \\ + \frac{\lambda_2}{2}\sum_{d=1}^{D}\left\|(1+\psi(t))\cdot w\cdot\left(f_t^d - f_{t-1}^d\right)\right\|_{l^2}^2 \end{aligned} \tag{7}$$

It can be found in Eq. (7) that the proposed adaptive spatial-temporal regularization $(1+\psi(t))\cdot w$ adaptively adjusts the constraint of the training and updating of the filter $\mathbf{f}_t=\left(f_t^1,\ldots,f_t^D\right)$ according to the target appearance variation. Thus it can effectively improve the adaptability of filter with respect to the current target appearance, while reducing the boundary effect and ensuring the tracker's stability.

Using fast fourier transform (FFT), the objective function above can be effectively solved in the Fourier domain. By applying Parseval's theorem to Eq. (7), the filter training objective function in Fourier domain is equivalent to,

$$\begin{aligned}\hat{\mathbf{f}}_t = \arg\min_{\hat{\mathbf{f}}_t} \frac{1}{2}\left\|\sum_{d=1}^{D} \hat{x}_t^d \cdot \hat{f}_t^d - \hat{y}_t\right\|_{l^2}^2 + \frac{\lambda_1}{2}\sum_{d=1}^{D}\left\|\hat{f}_t^d\right\|_{l^2}^2 \\ + \frac{\lambda_2}{2}\sum_{d=1}^{D}\left\|\left(1+\hat{\psi}(t)\right)*\hat{w}*\left(\hat{f}_t^d - \hat{f}_{t-1}^d\right)\right\|_{l^2}^2 \end{aligned} \tag{8}$$

3.3. Iterative optimization of the filter based on ADMM

To solve the Eq. (8), we employ the ADMM. Let the $\hat{\mathbf{g}}=\hat{\mathbf{f}}_t$ be the introduced auxiliary variable, the required augmented Lagrangian form of Eq. (8) is written as,

$$\begin{aligned}L\left(\hat{\mathbf{f}}_t,\hat{\mathbf{g}},\hat{\mathbf{h}}\right) = \frac{1}{2}\left\|\sum_{d=1}^{D} \hat{x}_t^d \cdot \hat{f}_t^d - \hat{y}_t\right\|_{l^2}^2 + \frac{\lambda_1}{2}\sum_{d=1}^{D}\left\|\hat{g}^d\right\|_{l^2}^2 \\ + \frac{\lambda_2}{2}\sum_{d=1}^{D}\left\|\left(1+\hat{\psi}(t)\right)*\hat{w}*\left(\hat{f}_t^d - \hat{f}_{t-1}^d\right)\right\|_{l^2}^2 \\ + \frac{\eta}{2}\sum_{d=1}^{D}\left\|\hat{f}_t^d - \hat{g}^d + \hat{h}^d\right\|_{l^2}^2 \end{aligned} \tag{9}$$

where the $\hat{\mathbf{h}}=\left(\hat{h}^1,\ldots,\hat{h}^D\right)$ is the Lagrange multiplier, and the is η a penalty factor. The ADMM algorithm is adopted by alternately solving the three subproblems in Eq. (10), and each subproblem in can be solved as follows,

$$\begin{cases} Q1: \hat{\mathbf{f}}_t^{(i+1)} = \arg\min_{\hat{\mathbf{f}}_t} \frac{1}{2}\left\|\sum_{d=1}^{D} \hat{x}_t^d \cdot \hat{f}_t^d - \hat{y}_t\right\|_{l^2}^2 \\ \quad + \frac{\lambda_2}{2}\sum_{d=1}^{D}\left\|\left(1+\hat{\psi}(t)\right)*\hat{w}*\left(\hat{f}_t^d - \hat{f}_{t-1}^d\right)\right\|_{l^2}^2 + \frac{\eta}{2}\sum_{d=1}^{D}\left\|\hat{f}_t^d - \hat{g}^d + \hat{h}^d\right\|_{l^2}^2 \\ Q2: \hat{\mathbf{g}}^{(i+1)} = \arg\min_{\hat{\mathbf{g}}} \frac{\lambda_1}{2}\sum_{d=1}^{D}\left\|\hat{g}^d\right\|_{l^2}^2 + \frac{\eta}{2}\sum_{d=1}^{D}\left\|\hat{f}_t^d - \hat{g}^d + \hat{h}^d\right\|_{l^2}^2 \\ Q3: \hat{\mathbf{h}}^{(i+1)} = \hat{\mathbf{h}}^{(i)} + \hat{\mathbf{f}}^{(i+1)} - \hat{\mathbf{g}}^{(i+1)} \end{cases} \tag{10}$$

- Q1. Solution to the filter $\hat{\mathbf{f}}_t$

When updating the filter $\hat{\mathbf{f}}_t$ in frame t, the $\left(1+\hat{\psi}(t)\right)*\hat{w}$ of Q1 in Eq. (10) is known and therefore can be denoted as $\beta(t)=\left(1+\hat{\psi}(t)\right)*\hat{w}$, which means the $\beta(t)\in\mathbb{R}^{W\times H}$. However, it is difficult to optimize the subproblem Q1 due to its high computation complexity caused be the high-dimensional features. In order to improve computational efficiency, it is necessary to consider solving the filter on all cross-channel elements of each feature.

Denoting the jth cross-channel element of the filter $\hat{\mathbf{f}}_t=\left(\hat{f}_t^1,\ldots,\hat{f}_t^D\right)\in\mathbb{R}^{W\times H\times D}$ as $V_j\left(\hat{\mathbf{f}}_t\right)=\left(\hat{f}_t^1(j),\ldots,\hat{f}_t^D(j)\right)\in\mathbb{R}^{D\times 1}$, the optimization problem of the jth cross-channel element of the filter is derived as,

$$\begin{aligned}V_j\left(\hat{\mathbf{f}}_t\right)^* = \arg\min_{V_j\left(\hat{\mathbf{f}}_t\right)} \frac{1}{2}\left\|V_j(\hat{\mathbf{x}}_t)^T V_j\left(\hat{\mathbf{f}}_t\right) - \hat{y}_t(j)\right\|_{l^2}^2 \\ + \frac{\lambda_2}{2}\left\|\beta(t)\{j\}\left(V_j\left(\hat{\mathbf{f}}_t\right) - V_j\left(\hat{\mathbf{f}}_{t-1}\right)\right)\right\|_{l^2}^2 \\ + \frac{\eta}{2}\left\|V_j\left(\hat{\mathbf{f}}_t\right) - V_j(\hat{\mathbf{g}}) + V_j\left(\hat{\mathbf{h}}\right)\right\|_{l^2}^2 \end{aligned} \tag{11}$$

where the $\beta(t)(j)\in\mathbb{R}$ is the jth element of $\beta(t)$. Taking the derivative of Eq. (11) with respect to the $V_j\left(\hat{\mathbf{f}}\right)$ be zero, we can get the closed-form solution for jth cross-channel element of the filter with the Sherman–Morrsion formula as,

$$V_j\left(\hat{\mathbf{f}}_t\right)^* = \frac{1}{\lambda_2\beta(t)\{j\}+\eta}\left(I - \frac{V_j(\hat{\mathbf{x}}_t)V_j(\hat{\mathbf{x}}_t)^T}{\lambda_2\beta(t)\{j\}+\eta+V_j(\hat{\mathbf{x}}_t)^T V_j(\hat{\mathbf{x}}_t)}\right)\mathbf{p} \tag{12}$$

where the $\mathbf{p}$ is expressed as,

$$\mathbf{p}=V_j(\hat{\mathbf{x}}_t)\hat{y}_t(j)+\lambda_2\beta(t)\{j\}\left[V_j\left(\hat{\mathbf{f}}_{t-1}\right)-V_j\left(\hat{\mathbf{h}}\right)\right]+\eta V_j(\hat{\mathbf{g}}) \tag{13}$$

Note that Eq. (12) only contains vector multiply-add operation and thus the $V_j\left(\hat{\mathbf{f}}\right)$ can be computed efficiently. And the filter $\mathbf{f}_t$ in the spatial domain then can be further obtained by taking the inverse DFT of the $\hat{\mathbf{f}}_t$ in Eq. (12).

- Q2. Solution to the auxiliary variable $\hat{\mathbf{g}}$

The vectorization of the Q2 in Eq. (10) is expressed as,

$$\hat{\mathbf{g}}^* = \arg\min_{\hat{\mathbf{g}}} \frac{\lambda_1}{2}\|\hat{\mathbf{g}}\|_{l^2}^2 + \frac{\eta}{2}\left\|\hat{\mathbf{f}}_t - \hat{\mathbf{g}} + \hat{\mathbf{h}}\right\|_{l^2}^2 \tag{14}$$

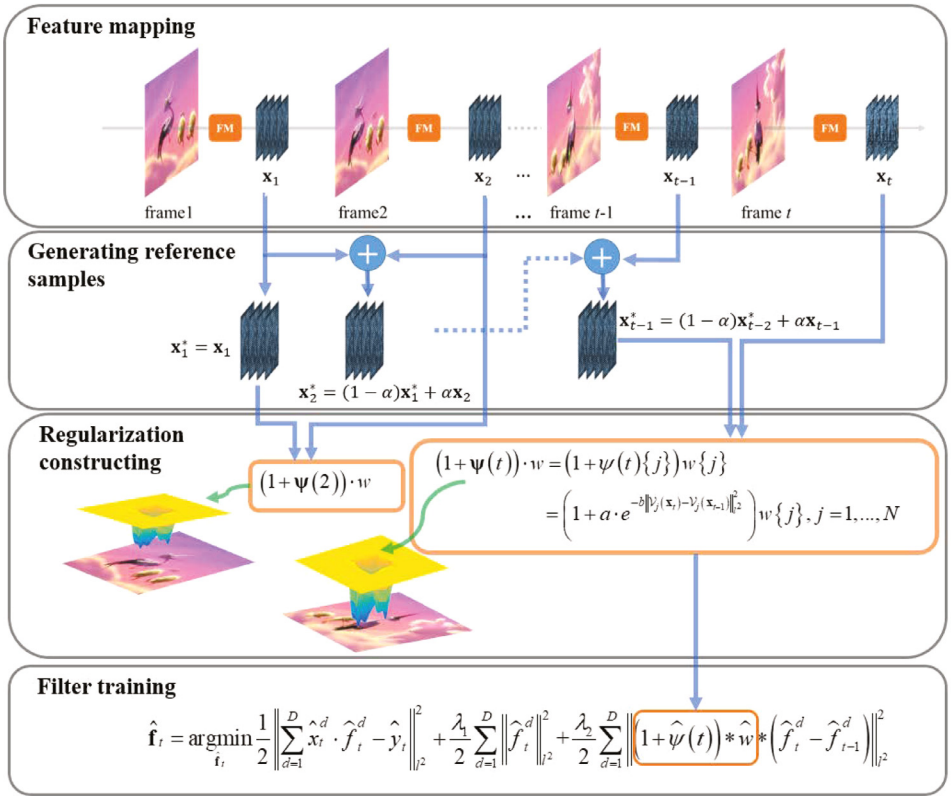


Fig. 2. Filter training with the adaptive spatial-temporal regularization model.

If other variables in Eq. (14) are fixed, the $\hat{\mathbf{g}}$ can be obtained by taking the derivative of Eq. (14) with respect to the $\hat{\mathbf{g}}$ be zero. The solution of $\hat{\mathbf{g}}$ can be expressed as,

$$\hat{\mathbf{g}}^{*} = \frac{\eta}{\lambda_1 + \eta}\left(\hat{\mathbf{f}}_t + \hat{\mathbf{h}}\right) \tag{15}$$

- Q3. Solution to the Lagrange multiplier $\hat{\mathbf{h}}$

The update formula of the Lagrange multiplier $\hat{\mathbf{h}}$ in Q3 of Eq. (10) is,

$$\hat{\mathbf{h}}^{(i+1)} = \hat{\mathbf{h}}^{(i)} + \hat{\mathbf{f}}_t^{(i+1)} - \hat{\mathbf{g}}^{(i+1)} \tag{16}$$

where the $\hat{\mathbf{f}}_t^{(i+1)}$ and $\hat{\mathbf{g}}^{(i+1)}$ are obtained by Eqs. (12) and (15).

It should be noted that the convergence of the proposed filter training method is proved in the Appendix.

4. The implementation framework of the proposed method

The filter training and updating framework with the proposed adaptive spatial-temporal regularization based on the target appearance variation model is shown in Fig. 2.

As can be seen, there are four modules contained in Fig. 2, which going forward one by one from top to bottom are the Feature mapping, the Generating reference samples, the Regularization constructing and the Filter training. Firstly, the sample features $(\mathbf{x}_1, \ldots, \mathbf{x}_t)$ of the training sample patches in frame 1 to t are extracted in the module of Feature mapping. Secondly, the reference sample $\mathbf{x}^*_{t-1} = (1-\alpha)\mathbf{x}^*_{t-2} + \alpha\mathbf{x}_{t-1}$ in frame t is constructed in the module called Generating reference samples by taking weighted average of the training samples $(\mathbf{x}_1, \ldots, \mathbf{x}_{t-1})$. Then we use the training sample $\mathbf{x}_t$ and the reference sample $\mathbf{x}^*_{t-1}$ to adaptively adjust the filter training spatial-temporal regularization constraint $(1+\psi(t))\cdot w$. Finally the filter updating in frame t will be implemented in the Filter training module with the adjusted spatial-temporal regularization constraint $(1+\psi(t))\cdot w$.

4.1. Feature mapping

The rectangles with the abbreviation FM in this module of Fig. 2 represent the process of sample feature extracting. When updating the filter in frame t, the Feature mapping module will firstly crop out the sample patch centered at the target position in frame t, and then extract its multi-channel sample feature $\mathbf{x}_t$ to construct the target appearance model in frame t. The obtained $\mathbf{x}_t$ is used as the training sample of the filter.

4.2. Generating reference samplesto the Lagrange multiplier

In order to evaluate the variation of the target appearance more accurately, it is necessary to establish a reference sample recording the historical target appearance information. The reference sample $\mathbf{x}^*_{t-1}$ in this paper is generated by taking weighted average of the samples $(\mathbf{x}_1, \ldots, \mathbf{x}_{t-1})$ extracted in frame 1 to t, which means,

$$\mathbf{x}^*_{t-1} = (1-\alpha)\mathbf{x}^*_{t-2} + \alpha\mathbf{x}_{t-1} \tag{17}$$

where the α is the learning rate. It should be noted that the reference sample in the second frame is defined as $\mathbf{x}^*_1 = \mathbf{x}_1$.

4.3. Regularization constructing

In the frame t, with the obtained reference sample $\mathbf{x}^*_{t-1}$ and the current target appearance model $\mathbf{x}_t$, the proposed method can evaluate the real-time variation of every parts of target appearance model $\left\|V_j(\mathbf{x}_t) - V_j\left(\mathbf{x}^*_{t-1}\right)\right\|^2_{l^2}$, $j = 1, \ldots, N$, and then adaptively adjust the filter training spatial-temporal regularization $(1+\psi(t))\cdot w$ by Eq. (5).

4.4. Filter training

In the filter training module, the filter can be efficiently learned by solving the objective function described in Eq. (8). The iterative

Input: The target central position p_1 and scale s_1 in frame 1;
Output: The filter model $\mathbf{f}_t$ updated in frame t.
Initialization:

Crop out the training sample patch centered at p_1 in frame 1 and extract the sample feature $\mathbf{x}_1$;
Initialize the w and the $\psi(1) = 0$ in frame 1;
Generate the reference sample in frame 1 as $\mathbf{x}^*{}_1 = \mathbf{x}_1$;
`// Filter training`
Solve the objective function in the Filter training module to obtain $\mathbf{f}_1$; `// Eq.(8)`
for $t = 2, t = t + 1, t \leqslant Num$ **do**
  Crop out the tracking sample patch centered at p_{t-1} in frame t and extract the sample feature $\mathbf{z}_t$;
  `// Translation estimation`
  Estimate the target position p_t in frame t using Eq.(2);
  `// Filter training and updating`
  `//Feature mapping`
  Crop out the training sample patch centered at p_t in frame t and extract the sample feature $\mathbf{x}_t$;
  `// Regularization constructing`
  Adaptively adjust the $1 + \psi(t)$ by Eq.(6);
  `// Filter training`
  Update the filter $\mathbf{f}_t$ by iteratively solving Eq.(8) based on ADMM;
  `// Generating reference samples`
  Update the reference sample $\mathbf{x}^*{}_t$ in frame t by Eq. (17);
end for

Algorithm 1. The filter training and updating process in frame t.

solving method of the objective function with FFT transform based on the ADMM has been described in Section 3.3.

More details about the filter training process on a video that contains *Num* numbers of image frames are shown in the following Algorithm 1.

5. Experiments

To verify the tracking performance related to the proposed method (OURS), comparison experiments are implemented with some state-of-the-art trackers including DeepSTRCF [40], ASRCF [37], ARCF [41], BACF [35], ECO [22], SRDCF [34], CSR_DCF [19], DSST [39] SSR_CF [36] and DSAR_CF [38]. The three datasets including OTB2015 [10], UAV123 [11] and the testset of the LaSOT [13] are used to evaluate the tracking performance of the trackers. The OTB2015 dataset is the most popular tracking benchmark with 100 video sequences. These videos are fully annotated with 11 different attributes, which can effectively evaluate the comprehensive tracking accuracy of the tracker. The UAV123 dataset contains a total of 123 video sequences taken from an aerial viewpoint. These videos have a long-time span, and the targets as well as the viewpoints in these videos also go through more changes. And the LaSOT dataset consists of 1400 sequences with more than 3.5 M frames in total, which comprise various challenges deriving from the wild where target objects may disappear and re-appear again in the view. Thus, these four datasets are appropriate to evaluate the adaptability and robustness of the tracker in a variety of different tracking conditions. It should be noted that only single-object tracking task is considered in the comparison of this paper. Besides, the performances of the trackers above are compared under the same environment conditions using MATLAB2016b equipped with Windows 10-64bit on Intel(R) Core (TM) i5-9300H CPU and 8 GB RAM.

Table 1
Parameters of experimental.

Parameter name	Value
Parameter a in Eq. (6)	1/12
Parameter b in Eq. (6)	0.5
The learning rate α in Eq. (17)	0.15
The regularization parameter λ_1 in Eq. (7)	10
The regularization parameter λ_2 in Eq. (7)	Deep: 12; hand-crafted: 16
Spatial regularization parameters μ in Eq. (5)	0.1
Spatial regularization parameters η in Eq. (5)	3

5.1. Experimental parameters

The proposed method uses the hand-crafted features including Color Names and HOG, and deep CNN features extracted using the 24th and 74th layers of the ResNet-50 as the sample features to construct the target appearance model and train the filter. The experimental parameters related to the proposed method are described in Table 1.

5.2. Evaluation indicators

The one-pass evaluation (OPE) criterion is used to measure the tracking performance. The success plot, precision plot and four numerical values, i.e. mean distance precision (Mean DP), mean overlap precision (Mean OP), tracking speed (FPS) and area-under-curve (AUC) are used as the expressions of the experiment results. The tracking distance precision (DP) in a video is defined as the ratio of frames where the Euclidean distance between the tracking output and ground truth is smaller than a threshold dp, here, $dp = 20(pixel)$. The tracking overlap precision (OP) in a video is defined as the ratio of frames which the intersection-over-union (IOU) is greater than a certain threshold op, here, $op = 0.5$. And given an estimated bounding box ROI_e and the ground-truth bounding box ROI_g of the target, the IOU is defined as,

$$IOU = \frac{area(ROI_e \cap ROI_g)}{area(ROI_e \cup ROI_g)} \tag{18}$$

5.3. Comparisons and analysis

5.3.1. The performance of the proposed method

In this section, we firstly discuss the proposed tracker's adaptability to target appearance variation and the robustness to the interference.

Adaptability To demonstrate the ability of the proposed method to adapt to the target appearance changing, Fig. 3 shows the temporal filter variation in the tracking process on the video named car16_2_1. The value of the filter variation curve in frame t is calculated as $\Delta_{\mathbf{f}}(t) = \xi \|\mathbf{f}_t - \mathbf{f}_{t-1}\|^2$, where the ξ is the normalization factor.

It can be seen from the Fig. 3 that compared with other trackers, the red curve of the filter variation corresponding to the proposed tracker has a floating range of $0.5 \times 10^{-3} \sim 1.3 \times 10^{-3}$, which is relatively flat. This shows that the filter model in our method does not change drastically during tracking process, and has better robustness. Furthermore, as shown by Subsequence1 in Fig. 3, in the frame 1 to 600 frames of the video, the appearance of the target is relatively stable, with only small deformation and scale changes. In this case, the filter model change of our method has a relatively small floating range between $0.5 \times 10^{-3} \sim 1.0 \times 10^{-3}$. This indicates that our tracker can update the filter steadily and smoothly by historical information of the target appearance fully using, leading to tracking robustness enhancing. On the contrary, when the target appearance changes visibly in the frame 700 to 800, which is shown in Subsequence2 of Fig. 3,

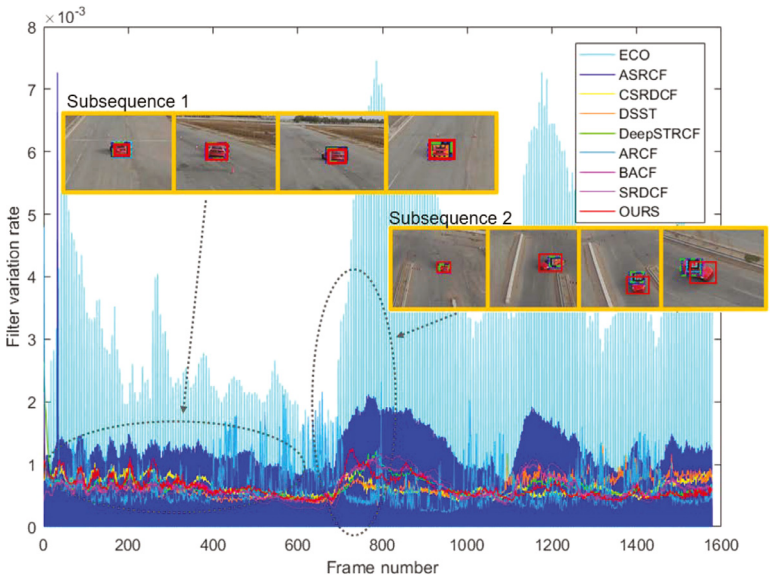


Fig. 3. Comparison of the temporal filter variation on car16_2_1.

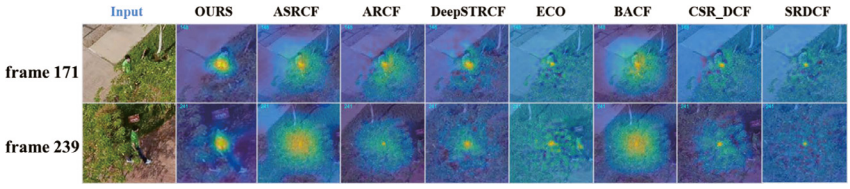


Fig. 4. Comparison of the filter responses on video person16_1.

The filter variation curve of our method firstly reaches a higher peak about 1.3×10^{-3} at about the 750th frame, earlier than other trackers. The results in Fig. 3 can demonstrate that in situations where the target appearance changes greatly, our method can update the filter more significantly, ensuring its ability to adapt to the current target appearance. Therefore, our tracker enhances the adaptability to the target appearance changes.

Robustness In order to further prove the robustness of the proposed tracker to the interference, Fig. 4 shows the comparison of the filter response heat-maps on frame 171 (first row) and frame 239 (second row) of the video person16_1. The comparison counterparts are 7 trackers equipped with temporal or spatial regularization, including ASRCF, ARCF, DeepSTRCF, ECO, BACF, CSR_DCF and SRDCF.

As can be seen in Fig. 4, in frame 171 where the target is partially occluded, the maximum filter response of OURS is accurately located at the center of the target, and the most energy of our filter response concentrates in the target region, without being greatly affected by the occlusion. While, the filter response of other methods is more scattered. For example, the ASRCF, ARCF, DeepSTRCF, and BACF algorithms all produce larger filter responses on obstructions. Moreover, the maximum filter responses of the other trackers including ECO, CSR_DCF and SRDCF are severely attenuated because of the noise caused by the obstructions. In frame 239, it can be seen that the proposed tracker can accurately track the target which is out of the occlusion, and the filter response of OURS is still concentrated in the target area. While, other trackers are greatly affected by the occlusion, their maximum filter response are stuck on the obstructions and cannot continue to effectively track the target.

Therefore, the results in Fig. 4 above show that the proposed method has strong robustness to interference, such as occlusion, in tracking process. In fact, the proposed tracker is also robust against many other interference factors including target in-plane/out-of-plane rotation, motion blur, illumination variation and background clutter. These can be verified in the following experimental results.

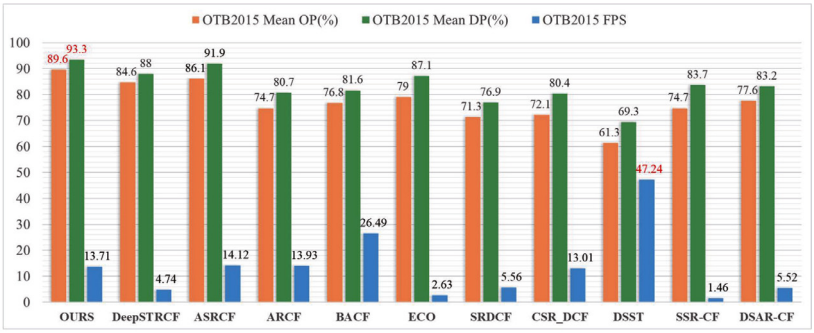
5.3.2. Baseline comparison

In this section, a comprehensive comparative analysis of the tracking performance of the proposed method with 10 comparison trackers is given. Fig. 5(a)–(c) show the comparison of Mean OP, Mean DP and FPS of the trackers on OTB2015, UAV123 and LaSOT datasets. And the best results are highlighted in red fonts.

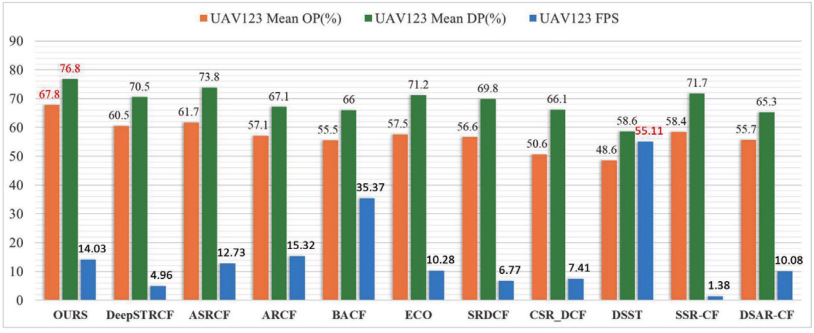
It can be seen that in the comparison on the OTB2015 dataset shown in Fig. 5(a), the OURS achieves the highest Mean OP of 89.6%, the highest Mean DP of 93.3%. Compared with the second-best ASRCF, the proposed method achieves a gain of 3.5% in Mean OP and 6.1% in Mean DP respectively, which is shown in Fig. 5(a). In the comparison on the UAV123 dataset shown in Fig. 5(b), the OURS achieves the highest Mean OP of 67.8% and the highest Mean DP of 76.8%, which are better than that of the second-best ASRCF tracker by 6.3% and 6.3% respectively.

Furthermore, In the comparison on the LaSOT dataset shown in Fig. 5(c), the OURS also achieves the highest Mean OP of 41% and the highest Mean DP of 46%, which are better than that of the second-best DeepSTRCF tracker by 3.5% and 6.9% respectively.

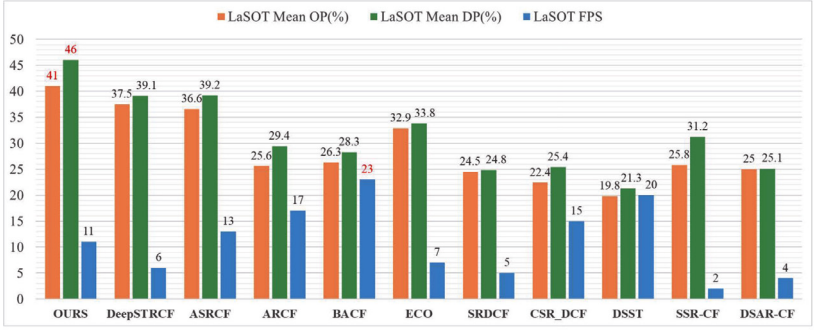
In the comparison of the FPS on three datasets showed in Fig. 5(a)–(c), OURS achieves acceptable speeds as 13.71, 14.03 and 11 fps on OTB2015, UAV123 and LASOT datasets respectively. Benefited from the proposed adaptive spatial-temporal regularization, OURS can update the filter more stably by fully using of the historical target information, to prevent the tracker overfitting. Moreover, OURS can also adaptively adjust the degree of the filter updating according to the current change in target appearance to accurately track the constantly changing target. Therefore, although the tracking speeds of OURS are slightly inferior than that of the



(a) OTB2015



(b) UAV123



(c) LaSOT

Fig. 5. Comparison of Mean OP, Mean DP and FPS on OTB2015(a), UAV123(b) and LaSOT(c).

fastest trackers such as DSST and BACF, the tracking performance of the proposed method is significantly improved.

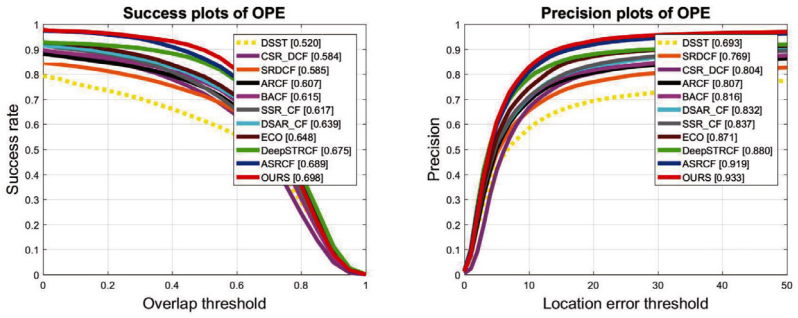
Fig. 6 (a)–(c) show the success plots and the precision plots on OTB2015, UAV123 and LaSOT datasets, respectively. The AUC scores of each tracker are displayed behind its tracker name in the Success plots, and the mean distance precision scores of are displayed behind the tracker name in the Precision plots. It can be seen in Fig. 6(a)–(c), the red curve corresponding to OURS is always at the top of the success plots and the precision plots, indicating that the proposed method outperforms other trackers.

In the comparison results on OTB2015 dataset which is shown in Fig. 6(a), OURS achieves a maximum AUC of 69.6% and a maximum mean distance precision of 93.3%, while the second best ASRCF achieves the tracking performance very close to the proposed tracker, reaching an AUC of 68.9% and a mean distance precision of 91.9%, respectively.

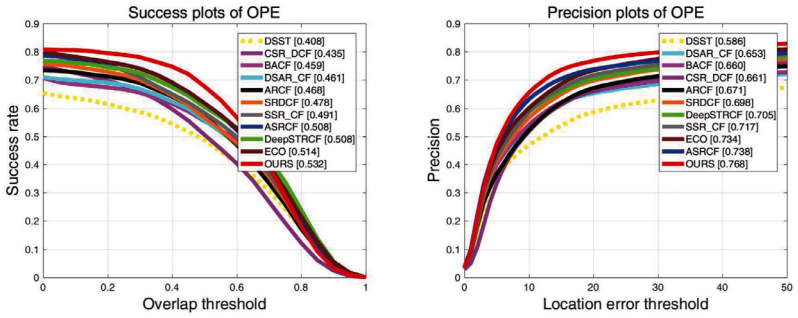
Fig. 6 (b) shows the success plots and the precision plots on the UAV123 dataset. It is obvious that the red curve corresponding to OURS is still at the top. The proposed tracker achieves the highest AUC score of 53.2% and a maximum mean distance precision of 76.8%. Compared with the ASRCF and DeepSTRCF, the tracking performance of our method is significantly improved. The second-best ASRCF tracker achieves an AUC score of 50.8% and a mean distance precision of 73.8%. And the DeepSTRCF tracker achieves an AUC score of 50.8% and a mean distance precision of 70.5%.

The comparison results on LaSOT dataset are shown in Fig. 6(c). It can be seen that the red curve corresponding to OURS is still at the top. The OURS achieves the highest AUC score of 37.3% and a maximum mean distance precision of 46%, which are significantly improved compared with that of the second-best DeepSTRCF tracker, which achieves an AUC score of 34.6% and a mean distance precision of 39.1%.

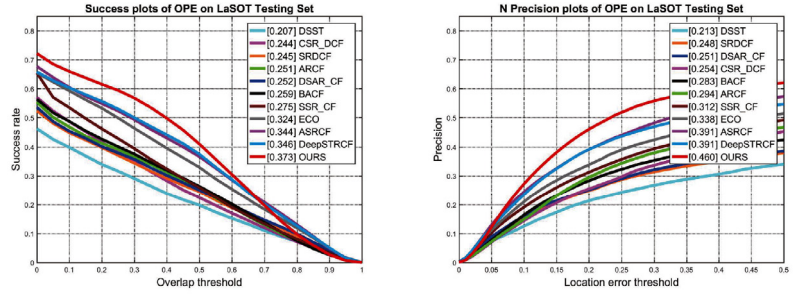
These results above demonstrate that the proposed method has a better adaptability to the rapid changes of the target appearance while ensuring the robustness of the tracker, and therefore can achieve more robust and accurate target tracking.



(a) OTB2015



(b) UAV123



(c) LaSOT

Fig. 6. Success plots and the Precision plots on OTB2015(a), UAV123(b) and LaSOT(c).

5.3.3. Attribute based comparison

In this section, we further perform an attribute-based analysis of all the methods respectively on the OTB2015, UAV 123 and LaSOT datasets.

OTB2015 In OTB2015 dataset, all the sequences are annotated with 11 different attributes, which correspond to 11 common difficulties and challenges that may exist in the tracking process. These 11 attributes are: Fast Motion, Background Clutter, Motion Blur, Deformation, Illumination Variation, In-Plane Rotation, Low Resolution, Occlusion, Out-of-Plane Rotation, Out of View, and Scale Variation.

Fig. 7 shows the comparison of the AUC scores of the trackers on all 11 attributes in OTB2015. It can be found that there are 11 coordinate axes originating from the center of the Fig. 7, and each coordinate axis represents one of the above-mentioned attributes. On each visual attribute axe, the AUC scores of trackers are arranged from the center of the figure to the edge in order from small to large, and the AUC scores on all attributes of the same tracker are connected to generate a polygon. Therefore, as far as a single attribute is concerned, the tracker with a high AUC score is arranged near the edge of the figure. And the tracker has a stronger comprehensive ability to deal with the above 11 tracking problems if its polygon is larger.

As can be seen from Fig. 7, the polygon corresponding to OURS is the largest, indicating that tracker OURS has the strongest comprehensive ability to deal with 11 kinds of interference factors. On the attribute axis of the In-plane Rotation, the AUC score of OURS is about 68%, the highest, indicating that the tracking accuracy of the proposed tracker is the highest when the target has in-plane rotation. The following second-best ASRCF tracker achieves an AUC score of about 61%. As can be see, the tracker OURS has superior AUC scores on all above attributes except the Background Clutter. Moreover, the tracking performance of OURS is significantly improved on four attributes, namely Fast Motion, Motion Blur, Low Resolution and Out of View. In order to explain the tracking performance of the above four attributes in more detail, success plots and precision plots on these four attributes are presented respectively in Fig. 8(a)–(d).

It can be clearly seen in Fig. 8, the red curve corresponding to tracker OURS is always higher than that corresponding to other trackers. For example, in the Low-Resolution attribute shown in Fig. 8(c), the OURS tracker achieves the highest AUC score of 65.7%

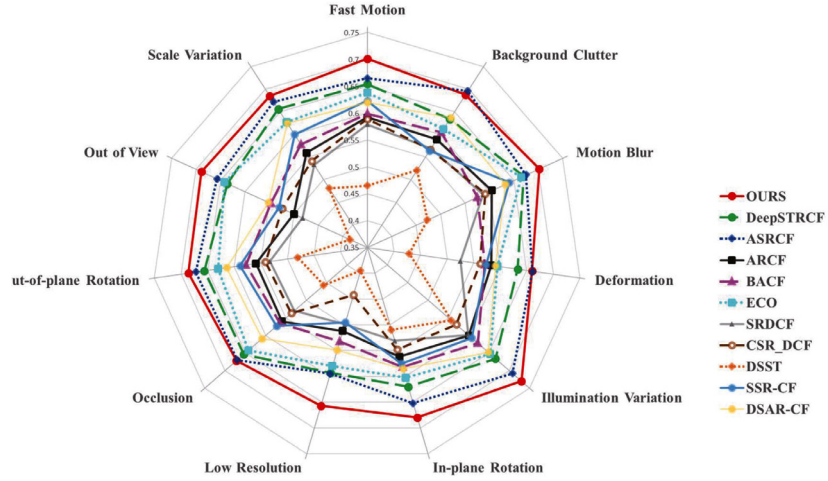
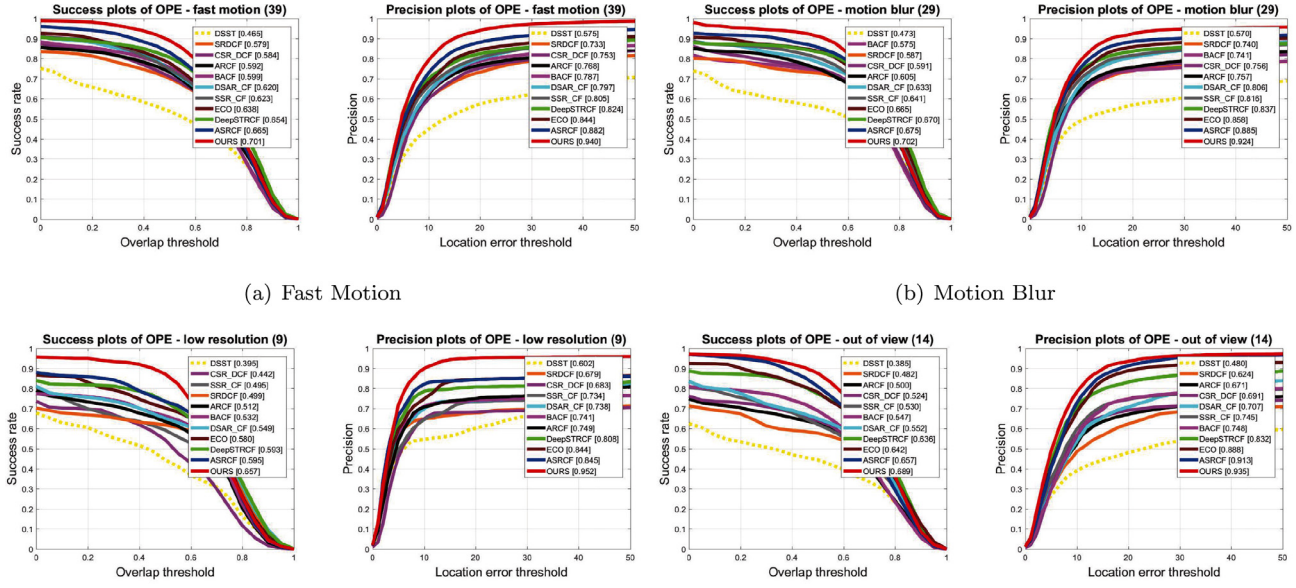


Fig. 7. Comparison of the AUC scores on all visual attributes (OTB2015).



(a) Fast Motion (b) Motion Blur

(c) Low Resolution (d) Out of View

Fig. 8. Success plots and Precision plots on four attributes (OTB2015).

and the highest mean distance precision of 95.2%, which has enhanced by 6.4% and 14.4% respectively than those of the ASRCF tracker.

UAV123 The UAV123 dataset contains 12 attributes corresponding to the common challenges in unmanned aerial object tracking. These 12 attributes are: Illumination Variation, Scale Variation, Partial Occlusion, Full Occlusion, Out of View, Fast Motion, Camera Motion, Background Clutter, Similar Object, Aspect Ratio Change, Viewpoint Change and Low Resolution. Similar to the Figs. 7 and 9 shows comparison of the AUC scores of the trackers on all the 12 attributes in UAV123.

It can be observed from Fig. 9 that compared with other trackers, the OURS tracker achieves higher AUC scores on attributes including Aspect Ratio Change, Camera Motion, Fast Motion and Out of View. For example, on the Aspect Ratio Change axis, the OURS tracker reaches the highest AUC score of about 47%, while the AUC scores of the other comparison trackers on this attribute are all below 45%. This shows that the proposed spatial-temporal regularization can adaptively adjust the filter training constraint according to the target appearance change. Therefore, the OURS tracker has a better adaptability to severe appearance variation such as target position and scale changes, and achieves more accurate and stable tracking.

Furthermore, the OURS tracker outperforms the other trackers significantly on four attributes, namely Background Clutter, Illumination Variation, Partial Occlusion and Viewpoint Change. Fig. 10(a)–(d) presents the success plots and precision plots on these four attributes respectively.

It can be found in Fig. 10 that the red curve corresponding to the OURS tracker is always higher than that corresponding to other trackers and significantly improved. For example, in the Illumination Variation shown in Fig. 10(b), the OURS tracker achieves the highest AUC score of 51.7% and the highest mean distance precision of 78.2%, which is better than those of the second-performing

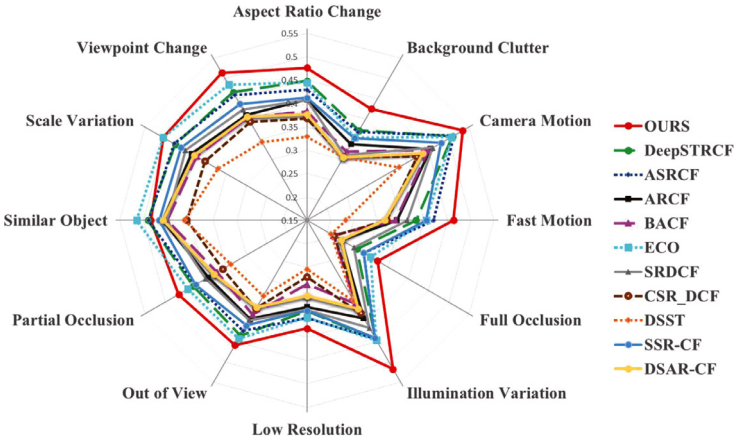
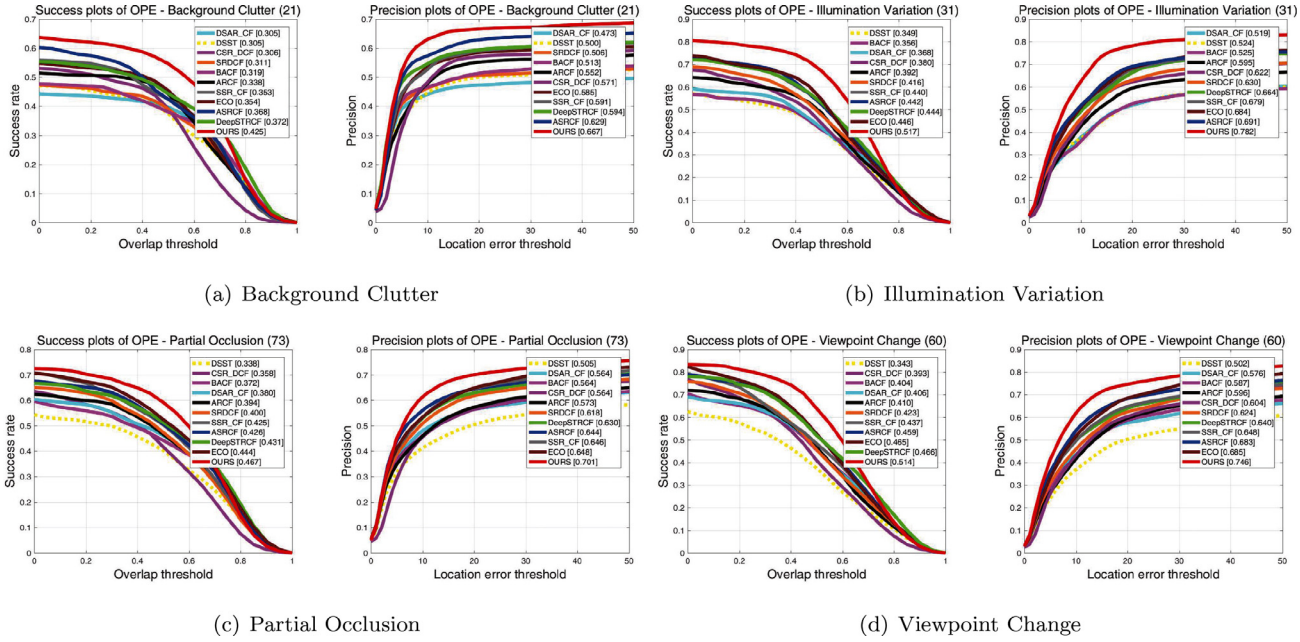


Fig. 9. Comparison of the AUC scores on all visual attributes (UAV123).



(a) Background Clutter (b) Illumination Variation

(c) Partial Occlusion (d) Viewpoint Change

Fig. 10. Success plots and Precision plots on four attributes (UAV123).

tracker by 7.1% and 9.1%. This shows that the proposed method has strong robustness to the changes of illumination in the tracking scene.

LaSOT We further perform attribute-based analysis on the LaSOT dataset, which contains 14 attributes corresponding to the common challenges in object tracking. Similar to the Figs. 10 and 11(a)–(d) presents the success plots and precision plots on four attributes including Camera Motion, Deformation, ROtation and Scale Variation respectively.

It can be found in Fig. 11 that the red curve corresponding to the proposed tracker is always higher than that corresponding to other trackers. For example, in the Rotation shown in Fig. 11(c), the OURS tracker achieves the highest AUC score of 34.9% and the highest mean distance precision of 43.4%, which is better than those of the second-performing tracker by 3.8% and 8.5%. This shows that the proposed method has strong robustness against the target appearance deformation in the tracking scene.

5.3.4. Tracking performance

In order to compare the tracking performance of the trackers more intuitively, Fig. 12 shows the comparison of the tracking bounding boxes in some frames of 6 videos including MotorRolling, Matrix, DragonBaby, Person16, Car3_s and Car16_2.

As can be seen from the Fig. 12, compared with other methods, the red bounding box representing the OURS tracker can track the position and scale change of the targets more stably and accurately. The OURS tracker has better robustness to various interference factors. For example, in the video sequence named MotorRolling, it can be found that all the trackers can accurately capture the target in frame #23. However, in frame #31, the DSR_DCF, DeepSTRCF and ECO trackers cannot adapt to the drastic changes of the target appearance caused by the in-plane and out-of-plane, and have a large tracking deviation, but the OURS tracker can still track the target accurately, After the frame #68, no trackers except the OURS tracker can accurately locate the target whose appearance is constantly changing. The comparison results on motorRolling demonstrate that the OURS tracker has good adaptability to the in-plane rotation, out-of-plane rotation and scale change of the target. In addition, it can be seen from other videos in Fig. 12 that the tracking accuracy and robustness of the OURS7 tracker outperform to that of other trackers in the case of fast motion, out-of-plane rotation, partial occlusion and camera viewpoint change, etc.

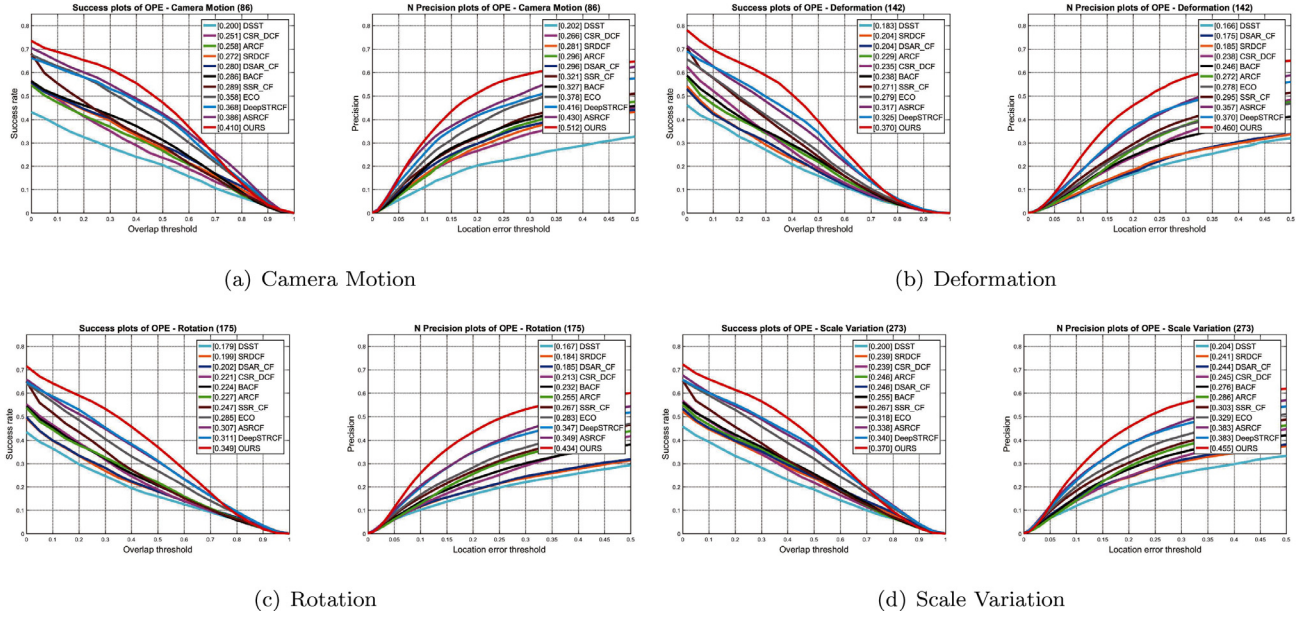


Fig. 11. Success plots and Precision plots on four attributes (LaSOT).

Fig. 12. Ccomparison of the tracking bounding boxes.

6. Conclusion

This paper proposes an adaptive spatial-temporal regularization model based on the target appearance variation in filter training. Moreover, a novel filter training objective function based on the proposed adaptive regularization model and its solution algorithm are also presented. The proposed method enhances the tracker's adaptability to target appearance variation and robustness to interference. The tracking accuracy of the proposed tracker in various tracking scenes such Rotation, Deformance, Illumination Variation and Mast Motion is significantly improved. However, in the long-term tracking task, the object may be blocked or fade out of view for a long time. In these situations, the proposed tracking method is easy to overfit the interference factors that exist for a long time, thus cannot identify the target accurately. Therefore, it is necessary to design a tracker for long-term tracking scenarios in the feature work.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT authorship contribution statement

Lin Zhou: Conceptualization, Methodology, Software. **Yong Jin:** Writing – review & editing. **Han Wang:** Data curation, Writing – original draft. **Zhentao Hu:** Investigation, Supervision. **Shuaipeng Zhao:** Software, Validation.

Acknowledgments

This work was supported by the National Science Foundation Council of China (61771006, 61976080, U1804149, 61701170). Key Research Projects of University in Henan Province of China (21A413002, 19A413006, 20B510001), the Programs for Science and Technology Development of Henan Province (192102210254).

Appendix A

The filter training objective function in Eq. (8) is defined as,

$$\widehat{\mathbf{f}}_t = \arg\min_{\widehat{\mathbf{f}}_t} \frac{1}{2}\left\|\sum_{d=1}^{D}\widehat{x}_t^d\cdot\widehat{f}_t^d-\widehat{y}_t\right\|_{l^2}^2+\frac{\lambda_1}{2}\sum_{d=1}^{D}\left\|\widehat{f}_t^d\right\|_{l^2}^2 +\frac{\lambda_2}{2}\sum_{d=1}^{D}\left\|\left(1+\widehat{\psi}(t)\right)*\widehat{w}*\left(\widehat{f}_t^d-\widehat{f}_{t-1}^d\right)\right\|_{l^2}^2 \tag{A.1}$$

Let the $\widehat{\mathbf{g}}=\widehat{\mathbf{f}}_t$ be the introduced auxiliary variable, the Eq. (A.1) can be rearranged as,

$$\widehat{\mathbf{f}}_t = \arg\min_{\widehat{\mathbf{f}}_t} \frac{1}{2}\left\|\sum_{d=1}^{D}\widehat{x}_t^d\cdot\widehat{f}_t^d-\widehat{y}_t\right\|_{l^2}^2+\frac{\lambda_1}{2}\sum_{d=1}^{D}\left\|\widehat{g}_t^d\right\|_{l^2}^2 +\frac{\lambda_2}{2}\sum_{d=1}^{D}\left\|\left(1+\widehat{\psi}(t)\right)*\widehat{w}*\left(\widehat{f}_t^d-\widehat{f}_{t-1}^d\right)\right\|_{l^2}^2 \tag{A.2}$$

We define,

$$p\left(\widehat{\mathbf{f}}_t\right) = \frac{1}{2}\left\|\sum_{d=1}^{D}\widehat{x}_t^d\cdot\widehat{f}_t^d-\widehat{y}_t\right\|_{l^2}^2 +\frac{\lambda_2}{2}\sum_{d=1}^{D}\left\|\left(1+\widehat{\psi}(t)\right)*\widehat{w}*\left(\widehat{f}_t^d-\widehat{f}_{t-1}^d\right)\right\|_{l^2}^2 \tag{A.3}$$

and

$$q(\widehat{\mathbf{g}}_t)=\frac{\lambda_1}{2}\sum_{d=1}^{D}\left\|\widehat{g}_t^d\right\|_{l^2}^2 \tag{A.4}$$

Then the optimization problem of the Eq. (A.1) can be defined as,

$$\begin{aligned}&\text{minimize} \quad p\left(\widehat{\mathbf{f}}_t\right)+q(\widehat{\mathbf{g}}_t)\\ &\text{subject to} \quad \widehat{\mathbf{f}}_t-\widehat{\mathbf{g}}_t=0\end{aligned} \tag{A.5}$$

Here, we denote the optimal value of the problem in Eq. (A.5) as,

$$u^*=\inf\left\{p\left(\widehat{\mathbf{f}}_t\right)+q(\widehat{\mathbf{g}}_t)\middle|\widehat{\mathbf{f}}_t-\widehat{\mathbf{g}}_t=0\right\} \tag{A.6}$$

Then the augmented Lagrangian function of the Eq. (A.5) is formed as,

$$\begin{aligned}L\left(\widehat{\mathbf{f}}_t,\widehat{\mathbf{g}}_t,\widehat{\mathbf{s}}\right) &= p\left(\widehat{\mathbf{f}}_t\right)+q(\widehat{\mathbf{g}}_t)+\widehat{\mathbf{s}}^T\left(\widehat{\mathbf{f}}_t-\widehat{\mathbf{g}}_t\right)+\frac{\eta}{2}\left\|\widehat{\mathbf{f}}_t-\widehat{\mathbf{g}}_t\right\|_{l^2}^2\\ &= \frac{1}{2}\left\|\sum_{d=1}^{D}\widehat{x}_t^d\cdot\widehat{f}_t^d-\widehat{y}_t\right\|_{l^2}^2+\frac{\lambda_1}{2}\sum_{d=1}^{D}\left\|\widehat{g}^d\right\|_{l^2}^2\\ &\quad+\frac{\lambda_2}{2}\sum_{d=1}^{D}\left\|\left(1+\widehat{\psi}(t)\right)*\widehat{w}*\left(\widehat{f}_t^d-\widehat{f}_{t-1}^d\right)\right\|_{l^2}^2\\ &\quad+\sum_{d=1}^{D}\widehat{s}^{d\,T}\left(\widehat{f}^d-\widehat{g}^d\right)+\frac{\eta}{2}\sum_{d=1}^{D}\left\|\widehat{f}_t^d-\widehat{g}^d\right\|_{l^2}^2\end{aligned} \tag{A.7}$$

where the $\widehat{\mathbf{s}}=\left(\widehat{s}^1,\ldots,\widehat{s}^D\right)$ is the Lagrange multiplier, and the η a penalty factor. Let $\widehat{\mathbf{h}}=\frac{1}{\eta}\widehat{\mathbf{s}}$, the Eq. (A.7) can be written into a concise form,

$$\begin{aligned}L\left(\widehat{\mathbf{f}}_t,\widehat{\mathbf{g}},\widehat{\mathbf{h}}\right) &= \frac{1}{2}\left\|\sum_{d=1}^{D}\widehat{x}_t^d\cdot\widehat{f}_t^d-\widehat{y}_t\right\|_{l^2}^2+\frac{\lambda_1}{2}\sum_{d=1}^{D}\left\|\widehat{g}^d\right\|_{l^2}^2\\ &\quad+\frac{\lambda_2}{2}\sum_{d=1}^{D}\left\|\left(1+\widehat{\psi}(t)\right)*\widehat{w}*\left(\widehat{f}_t^d-\widehat{f}_{t-1}^d\right)\right\|_{l^2}^2\\ &\quad+\frac{\eta}{2}\sum_{d=1}^{D}\left\|\widehat{f}_t^d-\widehat{g}^d+\widehat{h}^d\right\|_{l^2}^2\end{aligned} \tag{A.8}$$

The ADMM algorithm is adopted by alternately solving the following three subproblems,

$$\begin{cases}Q1:\widehat{\mathbf{f}}_t^{(i+1)}=\arg\min_{\widehat{\mathbf{f}}_t}\frac{1}{2}\left\|\sum_{d=1}^{D}\widehat{x}_t^d\cdot\widehat{f}_t^d-\widehat{y}_t\right\|_{l^2}^2\\ \quad+\frac{\lambda_2}{2}\sum_{d=1}^{D}\left\|\left(1+\widehat{\psi}(t)\right)*\widehat{w}*\left(\widehat{f}_t^d-\widehat{f}_{t-1}^d\right)\right\|_{l^2}^2+\frac{\eta}{2}\sum_{d=1}^{D}\left\|\widehat{f}_t^d-\widehat{g}^d+\widehat{h}^d\right\|_{l^2}^2\\ Q2:\widehat{\mathbf{g}}^{(i+1)}=\arg\min_{\widehat{\mathbf{g}}}\frac{\lambda_1}{2}\sum_{d=1}^{D}\left\|\widehat{g}^d\right\|_{l^2}^2+\frac{\eta}{2}\sum_{d=1}^{D}\left\|\widehat{f}_t^d-\widehat{g}^d+\widehat{h}^d\right\|_{l^2}^2\\ Q3:\widehat{\mathbf{h}}^{(i+1)}=\widehat{\mathbf{h}}^{(i)}+\widehat{\mathbf{f}}^{(i+1)}-\widehat{\mathbf{g}}^{(i+1)}\end{cases} \tag{A.9}$$

Here, we use the Eq. (A.7) to prove the convergence of our filter training method. It is obvious that the function $p\left(\widehat{\mathbf{f}}_t\right)$ and $q(\widehat{\mathbf{g}}_t)$ are closed, proper and convex. We define the residual $\mathbf{r}=\widehat{\mathbf{f}}_t-\widehat{\mathbf{g}}_t$, and assume that there exists a saddle point $\left(\widehat{\mathbf{f}}_t^*,\widehat{\mathbf{g}}_t^*,\widehat{\mathbf{s}}^*\right)$ of the $L\left(\widehat{\mathbf{f}}_t,\widehat{\mathbf{g}}_t,\widehat{\mathbf{s}}\right)$, which means,

$$L\left(\widehat{\mathbf{f}}_t^*,\widehat{\mathbf{g}}_t^*,\widehat{\mathbf{s}}\right)\leqslant L\left(\widehat{\mathbf{f}}_t^*,\widehat{\mathbf{g}}_t^*,\widehat{\mathbf{s}}^*\right)\leqslant L\left(\widehat{\mathbf{f}}_t,\widehat{\mathbf{g}}_t,\widehat{\mathbf{s}}^*\right) \tag{A.10}$$

It can be proved that the ADMM iterates satisfy the following:

(1) Residual convergence:
When the number of iterations $i\to\infty$, $\mathbf{r}\to\mathbf{0}$, i.e., the iterates approach feasibility.
(2) Objective convergence:
When the number of iterations $i\to\infty$, the value of the objective function $p\left(\widehat{\mathbf{f}}_t^i\right)+q\left(\widehat{\mathbf{g}}_t^i\right)\to u^*$.
(3) Dual variable convergence
When the number of iterations $i\to\infty$, $\widehat{\mathbf{s}}^i\to\widehat{\mathbf{s}}^*$, where $\widehat{\mathbf{s}}^*$ is a dual optimal point.

In order to prove the three properties mentioned above, we firstly construct the Lyapunov function as,

$$V^i=(1/\eta)\left\|\widehat{\mathbf{s}}^i-\widehat{\mathbf{s}}^*\right\|_2^2+\eta\left\|\widehat{\mathbf{g}}_t^i-\widehat{\mathbf{g}}_t^*\right\|_2^2 \tag{A.11}$$

and prove the following three inequalities:

$$V^{i+1}\leqslant V^i-\eta\left\|\mathbf{r}^{i+1}\right\|_2^2-\eta\left\|\widehat{\mathbf{g}}_t^{i+1}-\widehat{\mathbf{g}}_t^i\right\|_2^2 \tag{A.12}$$

$$u^{i+1}-u^*\leqslant-\left(\widehat{\mathbf{s}}^{i+1}\right)^T\mathbf{r}^{i+1}+\eta\left(\widehat{\mathbf{g}}_t^{i+1}-\widehat{\mathbf{g}}_t^i\right)^T\left(-\mathbf{r}^{i+1}-\left(\widehat{\mathbf{g}}_t^{i+1}-\widehat{\mathbf{g}}_t^*\right)\right) \tag{A.13}$$

$$u^* - u^{i+1} \leqslant (\widehat{\mathbf{s}}^*)^T \mathbf{r}^{i+1} \tag{A.14}$$

The proof of the above three inequalities is discussed as follows.

Proof of inequation (A.14)

Since the $(\widehat{\mathbf{f}}_t^*, \widehat{\mathbf{g}}_t^*, \widehat{\mathbf{s}}^*)$ is a saddle point for the augmented Lagrangian function shown in the Eq. (A.7), namely

$$L(\widehat{\mathbf{f}}_t^*, \widehat{\mathbf{g}}_t^*, \widehat{\mathbf{s}}^*) \leqslant L(\widehat{\mathbf{f}}_t^{i+1}, \widehat{\mathbf{g}}_t^{i+1}, \widehat{\mathbf{s}}^*) \tag{A.15}$$

Using $\widehat{\mathbf{f}}_t^* - \widehat{\mathbf{g}}_t^* = \mathbf{0}$, the left-hand side $L(\widehat{\mathbf{f}}_t^*, \widehat{\mathbf{g}}_t^*, \widehat{\mathbf{s}}^*) = u^*$. With the equation that $u^{i+1} = p(\widehat{\mathbf{f}}_t^{i+1}) + q(\widehat{\mathbf{g}}_t^{i+1})$, the Eq. (A.15) can be written as,

$$u^* \leqslant u^{i+1} + (\widehat{\mathbf{s}}^*)^T \mathbf{r}^{i+1} \tag{A.16}$$

which can prove the inequation (A.14).

Proof of inequation (A.13)

Assuming that $\widehat{\mathbf{f}}_t^{i+1}$ minimizes $L(\widehat{\mathbf{f}}_t, \widehat{\mathbf{g}}_t^i, \widehat{\mathbf{s}}^i)$, since the $p(\widehat{\mathbf{f}}_t)$ is closed, proper, convex, and it is subdifferentiable, and so is $L(\widehat{\mathbf{f}}_t, \widehat{\mathbf{g}}_t^i, \widehat{\mathbf{s}}^i)$. The optimality condition is,

$$0 \in \partial L(\widehat{\mathbf{f}}_t^{i+1}, \widehat{\mathbf{g}}_t^i, \widehat{\mathbf{s}}^i) = \partial p(\widehat{\mathbf{f}}_t^{i+1}) + \widehat{\mathbf{s}}^i + \eta(\widehat{\mathbf{f}}_t^{i+1} - \widehat{\mathbf{g}}_t^i) \tag{A.17}$$

Since $\widehat{\mathbf{s}}^{i+1} = \widehat{\mathbf{s}}^i + \eta \mathbf{r}^{i+1}$, namely $\widehat{\mathbf{s}}^i = \widehat{\mathbf{s}}^{i+1} - \eta \mathbf{r}^{i+1}$, the Eq. (A.17) can be rearranged as,

$$0 \in \partial p(\widehat{\mathbf{f}}_t^{i+1}) + (\widehat{\mathbf{s}}^{i+1} + \eta(\widehat{\mathbf{g}}_t^{i+1} - \widehat{\mathbf{g}}_t^i)) \tag{A.18}$$

This implies that $\widehat{\mathbf{f}}_t^{i+1}$ minimizes,

$$p(\widehat{\mathbf{f}}_t) + (\widehat{\mathbf{s}}^{i+1} + \eta(\widehat{\mathbf{g}}_t^{i+1} - \widehat{\mathbf{g}}_t^i))^T \widehat{\mathbf{f}}_t \tag{A.19}$$

A similar argument shows that $\widehat{\mathbf{g}}_t^{i+1}$ minimizes $q(\widehat{\mathbf{g}}_t) - (\widehat{\mathbf{s}}^{i+1})^T \widehat{\mathbf{g}}_t$. It follows that,

$$p(\widehat{\mathbf{f}}_t^{i+1}) + (\widehat{\mathbf{s}}^{i+1} + \eta(\widehat{\mathbf{g}}_t^{i+1} - \widehat{\mathbf{g}}_t^i))^T \widehat{\mathbf{f}}_t^{i+1} \leqslant p(\widehat{\mathbf{f}}_t^*) + (\widehat{\mathbf{s}}^{i+1} + \eta(\widehat{\mathbf{g}}_t^{i+1} - \widehat{\mathbf{g}}_t^i))^T \widehat{\mathbf{f}}_t^* \tag{A.20}$$

and that

$$q(\widehat{\mathbf{g}}_t^{i+1}) - (\widehat{\mathbf{s}}^{i+1})^T \widehat{\mathbf{g}}_t^{i+1} \leqslant q(\widehat{\mathbf{g}}_t^*) - (\widehat{\mathbf{s}}^{i+1})^T \widehat{\mathbf{g}}_t^* \tag{A.21}$$

Adding the two inequalities above, using $\widehat{\mathbf{f}}_t^* - \widehat{\mathbf{g}}_t^* = \mathbf{0}$, we can obtain that,

$$u^{i+1} - u^* \leqslant -(\widehat{\mathbf{s}}^{i+1})^T \mathbf{r}^{i+1} + (\eta(\widehat{\mathbf{g}}_t^{i+1} - \widehat{\mathbf{g}}_t^i))^T (\widehat{\mathbf{f}}_t^* - \widehat{\mathbf{f}}_t^{i+1}) \tag{A.22}$$

and the inequation (A.13) can be proved by substituting the equation into the Eq. (A.22)

Proof of inequation (A.12)

Adding the Eqs. (A.13) and (A.14), and multiplying through by 2 gives,

$$2(\widehat{\mathbf{s}}^{i+1} - \widehat{\mathbf{s}}^*)^T \mathbf{r}^{i+1} + 2\eta(\widehat{\mathbf{g}}_t^{i+1} - \widehat{\mathbf{g}}_t^i)^T \mathbf{r}^{i+1} + 2\eta(\widehat{\mathbf{g}}_t^{i+1} - \widehat{\mathbf{g}}_t^i)^T (\widehat{\mathbf{g}}_t^{i+1} - \widehat{\mathbf{g}}_t^*) \leqslant 0 \tag{A.23}$$

Substituting the $\widehat{\mathbf{s}}^{i+1} = \widehat{\mathbf{s}}^i + \eta \mathbf{r}^{i+1}$, the $2(\widehat{\mathbf{s}}^{i+1} - \widehat{\mathbf{s}}^*)^T \mathbf{r}^{i+1}$ in Eq. (A.23) can be written as,

$$2(\widehat{\mathbf{s}}^i - \widehat{\mathbf{s}}^*)^T \mathbf{r}^{i+1} + \eta \|\mathbf{r}^{i+1}\|_2^2 + \eta \|\mathbf{r}^{i+1}\|_2^2 \tag{A.24}$$

and substituting $\mathbf{r}^{i+1} = \frac{1}{\eta}(\widehat{\mathbf{s}}^{i+1} - \widehat{\mathbf{s}}^i)$ into the first two items of Eq. (A.24), we can obtain that,

$$\frac{2}{\eta}(\widehat{\mathbf{s}}^i - \widehat{\mathbf{s}}^*)^T (\widehat{\mathbf{s}}^{i+1} - \widehat{\mathbf{s}}^i) + \frac{1}{\eta}\|\widehat{\mathbf{s}}^{i+1} - \widehat{\mathbf{s}}^i\|_2^2 + \eta \|\mathbf{r}^{i+1}\|_2^2 \tag{A.25}$$

Since $\widehat{\mathbf{s}}^{i+1} - \widehat{\mathbf{s}}^i = (\widehat{\mathbf{s}}^{i+1} - \widehat{\mathbf{s}}^*) - (\widehat{\mathbf{s}}^i - \widehat{\mathbf{s}}^*)$, the Eq. (A.25) can be written as,

$$\frac{1}{\eta}\left(\|\widehat{\mathbf{s}}^{i+1} - \widehat{\mathbf{s}}^*\|_2^2 - \|\widehat{\mathbf{s}}^i - \widehat{\mathbf{s}}^*\|_2^2\right) + \eta \|\mathbf{r}^{i+1}\|_2^2 \tag{A.26}$$

Using the Eq. (A.26) to replace the in Eq. (A.23), we can obtain that,

$$\frac{1}{\eta}\left(\|\widehat{\mathbf{s}}^{i+1} - \widehat{\mathbf{s}}^*\|_2^2 - \|\widehat{\mathbf{s}}^i - \widehat{\mathbf{s}}^*\|_2^2\right) + \eta \|\mathbf{r}^{i+1}\|_2^2 + 2\eta(\widehat{\mathbf{g}}_t^{i+1} - \widehat{\mathbf{g}}_t^i)^T \mathbf{r}^{i+1} + 2\eta(\widehat{\mathbf{g}}_t^{i+1} - \widehat{\mathbf{g}}_t^i)^T (\widehat{\mathbf{g}}_t^{i+1} - \widehat{\mathbf{g}}_t^*) \leqslant 0 \tag{A.27}$$

Substituting the $\widehat{\mathbf{g}}_t^{i+1} - \widehat{\mathbf{g}}_t^* = (\widehat{\mathbf{g}}_t^{i+1} - \widehat{\mathbf{g}}_t^i) + (\widehat{\mathbf{g}}_t^i - \widehat{\mathbf{g}}_t^*)$ into the Eq. (A.28),

$$\eta \|\mathbf{r}^{i+1}\|_2^2 + 2\eta(\widehat{\mathbf{g}}_t^{i+1} - \widehat{\mathbf{g}}_t^i)^T \mathbf{r}^{i+1} + 2\eta(\widehat{\mathbf{g}}_t^{i+1} - \widehat{\mathbf{g}}_t^i)^T (\widehat{\mathbf{g}}_t^{i+1} - \widehat{\mathbf{g}}_t^*) \tag{A.28}$$

and the Eq. (A.28) can be rewritten as,

$$\eta \|\mathbf{r}^{i+1} + (\widehat{\mathbf{g}}_t^{i+1} - \widehat{\mathbf{g}}_t^i)\|_2^2 + \eta \|\widehat{\mathbf{g}}_t^{i+1} - \widehat{\mathbf{g}}_t^i\|_2^2 + 2\eta(\widehat{\mathbf{g}}_t^{i+1} - \widehat{\mathbf{g}}_t^i)^T (\widehat{\mathbf{g}}_t^i - \widehat{\mathbf{g}}_t^*) \tag{A.29}$$

Substituting the $\widehat{\mathbf{g}}_t^{i+1} - \widehat{\mathbf{g}}_t^i = (\widehat{\mathbf{g}}_t^{i+1} - \widehat{\mathbf{g}}_t^*) - (\widehat{\mathbf{g}}_t^i - \widehat{\mathbf{g}}_t^*)$ into the last two items of the Eq. (A.29) gives,

$$\eta \|\mathbf{r}^{i+1} + (\widehat{\mathbf{g}}_t^{i+1} - \widehat{\mathbf{g}}_t^i)\|_2^2 + \eta\left(\|\widehat{\mathbf{g}}_t^{i+1} - \widehat{\mathbf{g}}_t^*\|_2^2 - \|\widehat{\mathbf{g}}_t^i - \widehat{\mathbf{g}}_t^*\|_2^2\right) \tag{A.30}$$

Replace the $\eta \|\mathbf{r}^{i+1}\|_2^2 + 2\eta(\widehat{\mathbf{g}}_t^{i+1} - \widehat{\mathbf{g}}_t^i)^T \mathbf{r}^{i+1} + 2\eta(\widehat{\mathbf{g}}_t^{i+1} - \widehat{\mathbf{g}}_t^i)^T (\widehat{\mathbf{g}}_t^{i+1} - \widehat{\mathbf{g}}_t^*)$ in Eq. (A.27) using the Eq. (A.30), we can obtain that,

$$\frac{1}{\eta}\left(\|\widehat{\mathbf{s}}^{i+1} - \widehat{\mathbf{s}}^*\|_2^2 - \|\widehat{\mathbf{s}}^i - \widehat{\mathbf{s}}^*\|_2^2\right) + \eta \|\mathbf{r}^{i+1} + (\widehat{\mathbf{g}}_t^{i+1} - \widehat{\mathbf{g}}_t^i)\|_2^2 + \eta\left(\|\widehat{\mathbf{g}}_t^{i+1} - \widehat{\mathbf{g}}_t^*\|_2^2 - \|\widehat{\mathbf{g}}_t^i - \widehat{\mathbf{g}}_t^*\|_2^2\right) \leqslant 0 \tag{A.31}$$

According to the definition of the Eq. (A.11), the Eq. (A.31) can be rearranged as,

$$V^{i+1} \leqslant V^i - \eta \|\mathbf{r}^{i+1} + (\widehat{\mathbf{g}}_t^{i+1} - \widehat{\mathbf{g}}_t^i)\|_2^2 \tag{A.32}$$

Since,

$$\eta \|\mathbf{r}^{i+1} + (\widehat{\mathbf{g}}_t^{i+1} - \widehat{\mathbf{g}}_t^i)\|_2^2 = \eta \|\mathbf{r}^{i+1}\|_2^2 + \eta \|\widehat{\mathbf{g}}_t^{i+1} - \widehat{\mathbf{g}}_t^i\|_2^2 + 2\eta(\mathbf{r}^{i+1})^T (\widehat{\mathbf{g}}_t^{i+1} - \widehat{\mathbf{g}}_t^i) \tag{A.33}$$

to show the inequation (A.12), it now suffices to show that $2\eta(\mathbf{r}^{i+1})^T(\widehat{\mathbf{g}}_t^{i+1} - \widehat{\mathbf{g}}_t^i) > 0$, and it is known that,

$$q(\widehat{\mathbf{g}}_t^{i+1}) - (\widehat{\mathbf{s}}^{i+1})^T \widehat{\mathbf{g}}_t^{i+1} \leqslant q(\widehat{\mathbf{g}}_t^i) - (\widehat{\mathbf{s}}^{i+1})^T \widehat{\mathbf{g}}_t^i \tag{A.34}$$

and

$$q(\widehat{\mathbf{g}}_t^i) - (\widehat{\mathbf{s}}^i)^T \widehat{\mathbf{g}}_t^i \leqslant q(\widehat{\mathbf{g}}_t^{i+1}) - (\widehat{\mathbf{s}}^i)^T \widehat{\mathbf{g}}^{i+1} \tag{A.35}$$

Thus, $((\widehat{\mathbf{s}}^i) - (\widehat{\mathbf{s}}^{i+1}))^T (\widehat{\mathbf{g}}_t^{i+1} - \widehat{\mathbf{g}}_t^i) \leqslant 0$. Substituting $\widehat{\mathbf{s}}^{i+1} = \widehat{\mathbf{s}}^i + \eta \mathbf{r}^{i+1}$, the inequation (A.12) can be proved.

References

[1] B.D. Lucas, T. Kanade, Iterative image registration technique with an application to stereo vision, in: Proceedings of the 7th International Joint Conference on Artificial Intelligence, vol. 2, 1981, pp. 674–679, doi:10.1042/cs0730285.

[2] S. Avidan, Support vector tracking, IEEE Trans. Pattern Anal. Mach. Intell. 26 (8) (2004) 1064–1072, doi:10.1109/TPAMI.2004.53.

[3] M.S. Arulampalam, S. Maskell, N. Gordon, T. Clapp, A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking, IEEE Trans. Signal Process. 50 (2) (2002) 174–188, doi:10.1109/78.978374.

[4] C. Qian, S. Xu, S. Zhang, Robust visual tracking via weighted incremental subspace learning, in: 2010 2nd International Conference on Future Computer and Communication, vol. 2, 2010, pp. V2-26–V2-29, doi:10.1109/ICFCC.2010.5497295.

[5] D.S. Bolme, J.R. Beveridge, B.A. Draper, Y.M. Lui, Visual object tracking using adaptive correlation filters, in: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2010, pp. 2544–2550, doi:10.1109/CVPR.2010.5539960.
[6] Y. Qi, S. Zhang, L. Qin, Q. Huang, H. Yao, J. Lim, M.H. Yang, Hedging deep features for visual tracking, IEEE Trans. Pattern Anal. Mach. Intell. 41 (5) (2019) 1116–1130, doi:10.1109/TPAMI.2018.2828817.
[7] B. Li, J. Yan, W. Wu, Z. Zhu, X. Hu, High performance visual tracking with siamese region proposal network, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 8971–8980, doi:10.1109/CVPR.2018.00935.
[8] Q. Wang, L. Zhang, L. Bertinetto, W. Hu, P.H.S. Torr, Fast online object tracking and segmentation: aunifying approach, in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 1328–1338, doi:10.1109/CVPR.2019.00142.
[9] P. Voigtlaender, J. Luiten, P.H.S. Torr, B. Leibe, Siam R-CNN: visual tracking by re-detection, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 6577–6587, doi:10.1109/CVPR42600.2020.00661.
[10] Y. Wu, J. Lim, M. Yang, Object tracking benchmark, IEEE Trans. Pattern Anal. Mach. Intell. 37 (9) (2015) 1834–1848, doi:10.1109/TPAMI.2014.2388226.
[11] M. Mueller, N. Smith, B. Ghanem, A Benchmark and Simulator for UAV Tracking, vol. 9905, LNCS, 2016, pp. 445–461, doi:10.1007/978-3-319-46448-0_27.
[12] M.K. et. al, The visual object tracking VOT2017 challenge results, in: 2017 IEEE International Conference on Computer Vision Workshops (ICCVW), 2017, pp. 1949–1972, doi:10.1109/ICCVW.2017.230.
[13] H. Fan, L. Lin, F. Yang, P. Chu, G. Deng, S. Yu, H. Bai, Y. Xu, C. Liao, H. Ling, Lasot: a high-quality benchmark for large-scale single object tracking, in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 5369–5378, doi:10.1109/CVPR.2019.00552.
[14] J.F. Henriques, R. Caseiro, P. Martins, J. Batista, Exploiting the circulant structure of tracking-by-detection with kernels, in: Proceedings of the 12th European Conference on Computer Vision, 2012, pp. 702–715, doi:10.1007/978-3-642-33765-9_50.
[15] T. Surasak, I. Takahiro, C. Cheng, C. Wang, P. Sheng, Histogram of oriented gradients for human detection in video, in: 2018 5th International Conference on Business and Industrial Research (ICBIR), 2018, pp. 172–176, doi:10.1109/ICBIR.2018.8391187.
[16] J. van de Weijer, C. Schmid, J. Verbeek, D. Larlus, Learning color names for real-world applications, IEEE Trans. Image Process. 18 (7) (2009) 1512–1523, doi:10.1109/TIP.2009.2019809.
[17] J.F. Henriques, R. Caseiro, P. Martins, J. Batista, High-speed tracking with kernelized correlation filters, IEEE Trans. Pattern Anal. Mach. Intell. 37 (3) (2015) 583–596, doi:10.1109/TPAMI.2014.2345390.
[18] F. Li, P.L. Y. Yao, D. Zhangand W. Zuo, M. Yang, Integrating boundary and center correlation filters for visual tracking with aspect ratio variation, in: 2017 IEEE International Conference on Computer Vision Workshops (ICCVW), 2017, pp. 2001–2009, doi:10.1109/ICCVW.2017.234.
[19] A. Lukeic, T. Vojr, L.C. Zajc, J. Matas, M. Kristan, Discriminative correlation filter with channel and spatial reliability, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4847–4856, doi:10.1109/CVPR.2017.515.
[20] S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: towards real-time object detection with region proposal networks, IEEE Trans. Pattern Anal. Mach. Intell. 39 (6) (2017) 1137–1149, doi:10.1109/TPAMI.2016.2577031.
[21] Y. Seo, K. Shin, Image classification of fine-grained fashion image based on style using pre-trained convolutional neural network, in: 2018 IEEE 3rd International Conference on Big Data Analysis (ICBDA), 2018, pp. 387–390, doi:10.1109/ICBDA.2018.8367713.
[22] M. Danelljan, G. Bhat, F.S. Khan, M. Felsberg, Eco: efficient convolution operators for tracking, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 6931–6939, doi:10.1109/CVPR.2017.733.
[23] G. Bhat, J. Johnander, M. Danelljan, F.S. Khan, M. Felsberg, Unveiling the power of deep tracking, in: 2018 Proceedings of the European Conference on Computer Vision, in: LNCS, vol. 11206, 2018, pp. 483–498, doi:10.1007/978-3-030-01216-8_30.
[24] M. Danelljan, G. Bhat, F.S. Khan, M. Felsberg, Atom: accurate tracking by overlap maximization, in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 4655–4664, doi:10.1109/CVPR.2019.00479.
[25] G. Bhat, M. Danelljan, L. Van Gool, R. Timofte, Learning discriminative model prediction for tracking, in: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 6181–6190, doi:10.1109/ICCV.2019.00628.
[26] B. Li, W. Wu, Q. Wang, F. Zhang, J. Xing, J. Yan, Siamrpn++: evolution of siamese visual tracking with very deep networks, in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 4277–4286, doi:10.1109/CVPR.2019.00441.
[27] J. Borui, L. Ruixuan, M. Jiayuan, X. Tete, J. Yuning, Acquisition of localization confidence for accurate object detection, in: 2018 Proceedings of the European Conference on Computer Vision, 2018, pp. 816–832, doi:10.1007/978-3-030-01264-9_48.
[28] K.N. Dai, Y. Zhang, D. Wang, J.H. Li, H.C. Lu, X.Y. Yang, High-performance long-term tracking with meta-updater, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 6297–6306, doi:10.1109/CVPR42600.2020.00633.
[29] L. Huang, X. Zhao, K. Huang, Globaltrack: a simple and strong baseline for long-term tracking, 2020, Proceedings of the AAAI Conference on Artificial Intelligence (2020) 11037–11044, doi:10.1609/AAAI.v34i07.6758.
[30] B. Yan, H.J. Zhao, D. Wang, H.C. Lu, X.Y. Yang, 'Skimming-perusal' tracking: a framework for real-time and robust long-term tracking, in: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 2385–2393, doi:10.1109/ICCV.2019.00247.
[31] M. Danelljan, F.S. Khan, M. Felsberg, J. Van De Weijer, Adaptive color attributes for real-time visual tracking, in: 2014 IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1090–1097, doi:10.1109/CVPR.2014.143.
[32] L. Bertinetto, J. Valmadre, S. Golodetz, O. Miksik, P.H.S. Torr, Staple: complementary learners for real-time tracking, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1401–1409, doi:10.1109/CVPR.2016.156.
[33] Y. Li, J. Zhu, A scale adaptive kernel correlation filter tracker with feature integration, Computer Vision - ECCV 2014 Workshops, 2015,pp. 254–265, doi:10.1007/978-3-319-16181-5_18.
[34] M. Danelljan, G. Hger, F.S. Khan, M. Felsberg, Learning spatially regularized correlation filters for visual tracking, in: 2015 IEEE International Conference on Computer Vision (ICCV), 2015, pp. 4310–4318, doi:10.1109/ICCV.2015.490.
[35] H.K. Galoogahi, A. Fagg, S. Lucey, Learning background-aware correlation filters for visual tracking, in: 2017 IEEE International Conference on Computer Vision (ICCV), 2017, pp. 1144–1152, doi:10.1109/ICCV.2017.129.
[36] Q. Guo, R. Han, W. Feng, Z. Chen, L. Wan, Selective spatial regularization by reinforcement learned decision making for object tracking, IEEE Trans. Image Process. 29 (2020) 2999–3013, doi:10.1109/TIP.2019.2955292.
[37] K. Dai, D. Wang, H. Lu, C. Sun, J. Li, Visual tracking via adaptive spatially-regularized correlation filters, in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 4665–4674, doi:10.1109/CVPR.2019.00480.
[38] W. Feng, R. Han, Q. Guo, J. Zhu, S. Wang, Dynamic saliency-aware regularization for correlation filter-based object tracking, IEEE Trans. Image Process. 28 (7) (2019) 3232–3245, doi:10.1109/TIP.2019.2895411.
[39] M. Danelljan, G. Hger, F.S. Khan, M. Felsberg, Discriminative scale space tracking, IEEE Trans. Pattern Anal. Mach. Intell. 39 (8) (2017) 1561–1575, doi:10.1109/TPAMI.2016.2609928.
[40] F. Li, C. Tian, W. Zuo, L. Zhang, M. Yang, Learning spatial-temporal regularized correlation filters for visual tracking, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 4904–4913, doi:10.1109/CVPR.2018.00515.
[41] Z. Huang, C. Fu, Y. Li, F. Lin, P. Lu, Learning aberrance repressed correlation filters for real-time UAV tracking, in: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 2891–2900, doi:10.1109/ICCV.2019.00298.