



# Robust visual tracker combining temporal consistent constraint and adaptive spatial regularization

Yi Zhang<sup>1</sup> · Guixi Liu<sup>1</sup> · Haoyang Zhang<sup>1</sup> · Hanlin Huang<sup>1</sup>

Received: 7 June 2020 / Accepted: 11 December 2020 / Published online: 11 January 2021  
© The Author(s), under exclusive licence to Springer-Verlag London Ltd. part of Springer Nature 2021

## Abstract

Existing discriminative correlation filters suffer from the defects of potential spatial distractors and the degradation of appearance model caused by hard-temporal correlation. Aiming at this issue, a robust tracker which combines the adaptive spatial regularization and the temporal consistent constraint is proposed in this paper. First, we propose to take the extracted saliency map of the background as a reference weight to construct the spatial regularization term, with which the perceived performance of the filter against distractors is enhanced by learning the spatial sparse constraint adaptively. Second, we further implement the temporal consistent regularization formed by capturing dynamic appearance information from multiple historical frames with a high-confidence strategy to mitigate the model degradation. Third, we employ the alternating direction method of multipliers to solve the constrained optimization problem efficiently, thereby the computational complexity can be reduced. The concrete experimental results on OTB-2013, OTB-2015, Temple-Color-128 and VOT2016 benchmarks demonstrate that our tracker outperforms several state-of-the-art algorithms.

**Keywords** Correlation filter tracking · Temporal consistent constraint · Adaptive spatial regularization · High-confidence strategy

## 1 Introduction

Visual tracking is one of the most significant topics in computer vision. It has comprehensive application prospects in intelligent surveillance, human–machine interaction, vehicle navigation and so on [1]. The core mission of generic tracking is to estimate the trace of a target in a series of consecutive frames on the basis of the initial specified state. In recent decades, the developing computing capability and robustness of correlation filter-based trackers lead to the substantial progress of it. However, visual tracking remains to be a complex and challenging problem due to the complicated interference factors in practical scenarios, such as deformation, occlusion and background clutter.

Generally, tracking algorithms can be classified into either generative or discriminative tracking according to

modeling method. The generative model-based method [2–5] can be regarded as a template matching process realized by searching for the candidate patch which is the most relevant to the target model. Discriminative tracking approaches [6, 7] provide a tracking-by-detection framework. They distinguish target from various negative samples through a learned classifier. Recently, discriminative correlation filter-based (DCF) tracking methods [8–11] have attracted extensive attention in the visual tracking community and displayed superior characteristics in multiple challenging benchmarks. We attribute this competitive performance to the implementation of circulant structures for efficient training and detection in the Fourier domain.

However, all negative examples adopted in the filter training process are generated through the cyclic property of correlation, which cannot truly represent the negative patch in real-world scenarios. Due to the limitation of the pseudo-negative sample, the tracker tends to drift under the heavy influence of background distractors. Addressing this problem, Danelljan et al. [12] propose a spatially regularized correlation filter for learning a more discriminative

✉ Guixi Liu  
gxliu@xidian.edu.cn

<sup>1</sup> School of Mechano-Electronic Engineering, Xidian University, Xi'an 710071, Shaanxi, China

model by retaining more negative information on a larger image region. The spatial regularization component with fixed weights ignores the diversity of the target and leads to an inaccurate description of the feature in case of background clutters and deformations. Galoogahi et al. [13] increase the proportion of real patches by preprocessing the image content with a binary mask, hence the filter could be learned from real negative samples extracted densely from the background. This approach requires a larger cropped region to offer more visual information of the background, which results in the increase of the computational burden during the whole training process. Moreover, Mueller et al. [14] derive a tracking framework with explicit incorporation of background content, which collects context patches around the target as hard negative examples and makes the context information regress to zeros. It would have the potential risk of model drift that provides limited background information in cluttered scenarios when such context-aware trackers taking context samples at the fixed position.

In addition, the classifier updates with an aggressive learning rate for handling appearance variations over time due to the extremely limited reliable positive samples in online classification. Addressing this problem, Li et al. [15] suggest a passive updating strategy which mitigates the filter degradation by introducing a temporal regularization that keeps the learned filter close to previous ones. Based on online passive-aggressive learning, the adaptability of the target template to appearance differences is improved. However, the temporal regularization component in aforementioned work is trained on the basis of the previous learned filter, which maintains diversity of the target appearance through two consecutive frames. The hard matching may incorporate inaccurate historical information continually that eventually leads to failure in complex scenarios.

In this paper, we aim at learning a joint correlation filter consisting of temporal consistent constraint and adaptive spatial regularization, with which we could enrich the target appearance diversity and perceive the background variations simultaneously. Specifically, we take the extracted saliency map of the background as a reference weight to construct a spatial regularization term which could learn the target-aware spatial constraint adaptively. Moreover, we introduce the temporal consistent component according to multiple historical target appearances selected by the high-confidence strategy, thereby eliminating the limitation of hard matching between two consecutive frames. For the proposed constrained optimization formulation, we apply the ADMM algorithm [16] to obtain the closed-form solution. Furthermore, we perform comprehensive experiments on tracking benchmarks to verify the

effectiveness of the proposal. The main contributions are as follows:

- An adaptive spatial constraint is learned by integrating a spatial regularization component into the DCF framework, which takes the saliency map of distractors as the reference weight.
- A temporal consistent constraint is constructed with the captured dynamic target appearance information by multiple reliable historical frames selected according to high-confidence strategy, by which a robust visual tracker combining temporal consistent constraint and adaptive spatial regularization (TSCF) is formulated.
- An ADMM algorithm is implemented to solve TSCF efficiently, where the closed-form solution is obtained.
- A thorough discussion and comparison between the proposed algorithm and several state-of-the-art trackers in details are illustrated on the OTB-2013 [17], OTB-2015 [18], Temple-Color-128 [19] and VOT2016 [20] datasets.

The rest of this paper is summarized as follows: Section 2 gives a brief review of some relevant literature. Section 3 presents a detailed description of the proposed tracking method. Section 4 provides experimental results and relevant discussions. Section 5 gives the conclusion.

## 2 Related works

### 2.1 Discriminative correlation filters

Discriminative correlation filter-based approaches with high efficiency have been widely applied in visual tracking community. Their efficiency comes from the element-wise multiplication implemented in the frequency domain. The minimum output sum of squared error tracker proposed by Bolme et al. [8] is the earliest DCF-based method that adopts only grayscale features to learn a filter. In practice, the single-channel gray information is too limited to express the characteristics of the target in complex scenarios. Henriques et al. [9] present a dense sampling technique to construct numerous training examples, with which exploit kernel trick to handle the nonlinear problem effectively. On this basis, multi-channel histogram features are utilized for performance improving in [10]. However, such algorithms cannot solve the variation of the target scale. Considering the impact of scale changing, an efficient method of training one-dimensional correlation filter based on a scale pyramid for obtaining the optimal target scale is demonstrated in [21, 22]. These approaches improve the tracking accuracy at the cost of speed reduction. Moreover, they also have high-precision requirement on target location in practical applications. Li et al. [23]

suggest a framework integrating 1D boundary and 2D center filters to deal with the scale variation. However, the average running time of this method is only 1.25 frames per second. In view of the feature representation, Danelljan et al. [11] extend trackers with multi-channel color attributes, which exhibits favorable discriminative characteristics with appearance description. But this tracker drifts away easily when the surroundings share similar visual cues with the object. Bertinetto et al. [24] learn a complementary tracker that is inherently robust to both deformations and illumination changes by merging histogram of oriented gradients (HOG) features and color histograms at the response map level. The fusion coefficient is fixed throughout the tracking process, which may not be optimal, hence shows the potential for improvement in accuracy. In order to integrate the multi-resolution feature map effectively, a continuous-domain learning formulation is introduced in [25] to obtain accurate localization thereafter. The trained continuous convolution filter is sparse, and it is prone to over-fitting with high-dimensional deep features. Therefore, Danelljan et al. [26] reduce the number of model parameters and increase tracking performance by introducing a factorized convolution operator for learning the filter with significant energy. From the perspective of the development process, various tracking algorithms are derived based on the correlation filter framework [27–30].

Among all the discriminative correlation algorithms mentioned above, background region and appearance diversity are not considered in the tracking process that inevitably brings unreliable distractors information. On this basis, we propose a robust tracker consisting of adaptive spatial regularization and temporal consistent constraint in this paper. We construct a spatial regularization term to learn the adaptive spatial weight, which can effectively suppress the background distractors in different tracking scenarios. Furthermore, we propose to learn reliable appearance information under the temporal consistent constraint, which improves the adaptability of the appearance model to target deformation.

## 2.2 Spatial constraint correlation filters

The traditional DCF-based trackers exploit the circulant structure to solve the ridge regression problem. The potential periodic assumption caused by the circulant matrices derives the redundant background information which hampers the learned model severely. To alleviate this issue, Danelljan et al. [12] introduce a spatial weight function into the DCF formulation to decrease the significance of background. The predefined spatial constraint keeps a negative Gaussian distribution in the tracking process that cannot guarantee the coefficient to be zero outside of the bounding box. As a result, the filter may

learn unexpected background information when the target undergoes appearance deformations. Galoogahi et al. [13] take background patches into account and then exploit a binary mask to crop the central image to increase the proportion of the real training sample. However, the image patches extracted by the mask do not explicitly exploit the shape information of the target. Mueller et al. [14] incorporate contextual information into the learned filter to suppress the background clutter. The context patches around the target are regressed to zeros that increases the risk of spurious detection on complex patterns. Moreover, the regression method of cropping image patches at fixed positions ignores the importance of pixel at different locations. [31] alleviates the impact of the surrounding background with a spatial reliability mask constructed by the target likelihood and the prior probability. But the feature adopted by this approach is sensitive to illumination variation.

As described above, [12–14] cannot suppress the background information completely in the positive sample because the unreliable region in the tracking box is not considered in the tracking process, and the shape information of the target is not effectively utilized by these methods. The spatial reliability adopted in [31] is sensitive to target deformation and has the risk of degradation. Compared with trackers mentioned above, we introduce a spatial regularization component into the discriminative correlation model while taking the extracted background reliability map as the reference weight. The spatial constraint learned by regularization takes the importance of pixel and shape variation of the target into account. Thereby, the unreliable region is penalized effectively by the learned spatial constraint.

## 2.3 Temporal regularized correlation filters

Generally, the target appearance changes over time during the whole tracking process, so it is critical for trackers to construct a robust appearance model. To maintain memory of the appearance description, Li et al. [15] formulate a robust tracker integrated temporal regularization which learns the appearance model from two consecutive frames. The filter adapts to appearance difference effectively by updating passively in sequence with occlusion and deformation. However, most temporal constraint-based discriminative correlation trackers ignore the consistency of the target appearance in temporal dimension. They learn the object diversity by constructing hard matching based on adjacent historical filter. This leads to poor generalization and model degradation when the visual relevance between the current and previous frames decreased in complex scenarios. Different from the above tracking methods, we filter out the frame with high relevance from the continuous

historical frames and put forward the temporal consistent constraint according to reliable filters. We apply a high-confidence mechanism to learn rich information of target appearance, which breaks through the limitation of hard matching between two consecutive frames. Moreover, reliable correlation in temporal dimension can further improve the representation ability of the appearance model.

### 3 Proposed tracking method

Motivated by the discussion above, we propose a robust tracker which combines the adaptive spatial regularization and the temporal consistent constraint in this paper. To be specific, first, we propose to integrate the constructed background saliency map into the correlation filter-based tracker, which can suppress the background information successfully. Second, we formulate the temporal consistent regularization according to the high-confidence strategy to improve the adaptability of appearance model. Finally, we use the ADMM algorithm to solve the constrained optimization model and obtain the closed-form solutions of all subproblems. Compared with the traditional spatial constraint method with fixed coefficients, we take the saliency map as a reference weight to formulate a spatial regularization term, hence the spatial constraint could be learned adaptively through quadratic regularization. By iterative solution, the penalty coefficients residing outside the target region have higher weights that can effectively suppress the distractors. Moreover, in order to overcome the limitation of the hard matching between two consecutive frames, we construct a temporal consistent constraint on the basis of reliable historical frames which are selected according to the high-confidence mechanism to mitigate the filter degradation. To evaluate the reliability of historical targets, we propose a dynamic threshold measurement to distinguish spurious detection. In addition, we exploit HOG features to train a CF model separately for fast scale estimation. The overall tracking flowchart of our proposed method is shown in Fig. 1.

Compared with the existing CF-based tracking algorithm, the proposed method achieves a balance between accuracy and robustness during online tracking in the given sequences as shown in Fig. 2. We attribute this to the fact that our TSCF tracker emphasizes the incorporation of the adaptive spatial constraint and the temporal consistent regularization separately. The learned spatial weight suppresses surrounding distractors, hence the tracking accuracy is effectively boosted. The temporal constraint component integrates the diversity of the target appearance, which makes the model more robust against challenging patterns.

#### 3.1 Adaptive spatial constraint model

The periodic assumption of the cyclic property produces distractors if the surroundings share similar visual cues with the target. In this paper, we extract the background saliency map from the search region and take it as a reference weight to formulate the spatial regularization component, which assigns higher penalty coefficients to the background by learning an adaptive spatial weight. The background saliency map that displays the spatial distribution information of the image is composed of the negative Gaussian shaped matrix and the background log-likelihood. We obtain the likelihood map of the image through the Bayesian formula, which calculates the probability of each color component belongs to the background region. This probabilistic model is prone to confusion when the target is surrounded by distractors similar to its color. In order to improve the performance in this situation, we combine the likelihood with a negative Gaussian matrix to form the background saliency map. This extension provides an effective suppression of the background.

The probability that a pixel  $p_n$  which is labeled as  $l_n \in L$  (foreground = 0, background = 1) at location  $n$  belongs to the background color model  $r^t$  via Bayesian rule in the current frame  $t$  can be depicted as [32]:

$$P(r^t|n) = \frac{P(n|l_n = 1, H_b^{t-1})}{P(n|l_n = 0, H_f^{t-1}) + P(n|l_n = 1, H_b^{t-1})} \quad (1)$$

where  $H_f^{t-1}$  and  $H_b^{t-1}$  denote the color histograms corresponding to the foreground object and background in the previous frame  $t - 1$ . The spatial saliency map is described with the proper combination of the probability model and the negative Gaussian matrix, which assigns higher penalty weights to region residing outside the target. The Gaussian function is constructed as follows:

$$G(p_r, p_c) = g_1 \cdot \left( \left( \frac{p_r}{w_{\text{tar}}} \right)^2 + \left( \frac{p_c}{h_{\text{tar}}} \right)^2 \right) + g_2 \quad (2)$$

where  $(p_r, p_c)$  represents the pixel coordinate of the input image,  $(w_{\text{tar}}, h_{\text{tar}})$  is the target scale,  $g_2$  is the minimum value of the predefined Gaussian window, and  $g_1$  determines the decay speed of weights from edge to center. The reference weight is obtained by simply multiplying the probability model by the two-dimensional Gaussian function. Therefore, the construction formula of the reference weight is as follows:

$$w^{\text{ref}} = G(p_r, p_c) \cdot P(r^t|n) \quad (3)$$

where  $w^{\text{ref}}$  is the reference weight matrix. By applying the Gaussian model centered at the target position on the likelihood map, the distractors with similar color to the



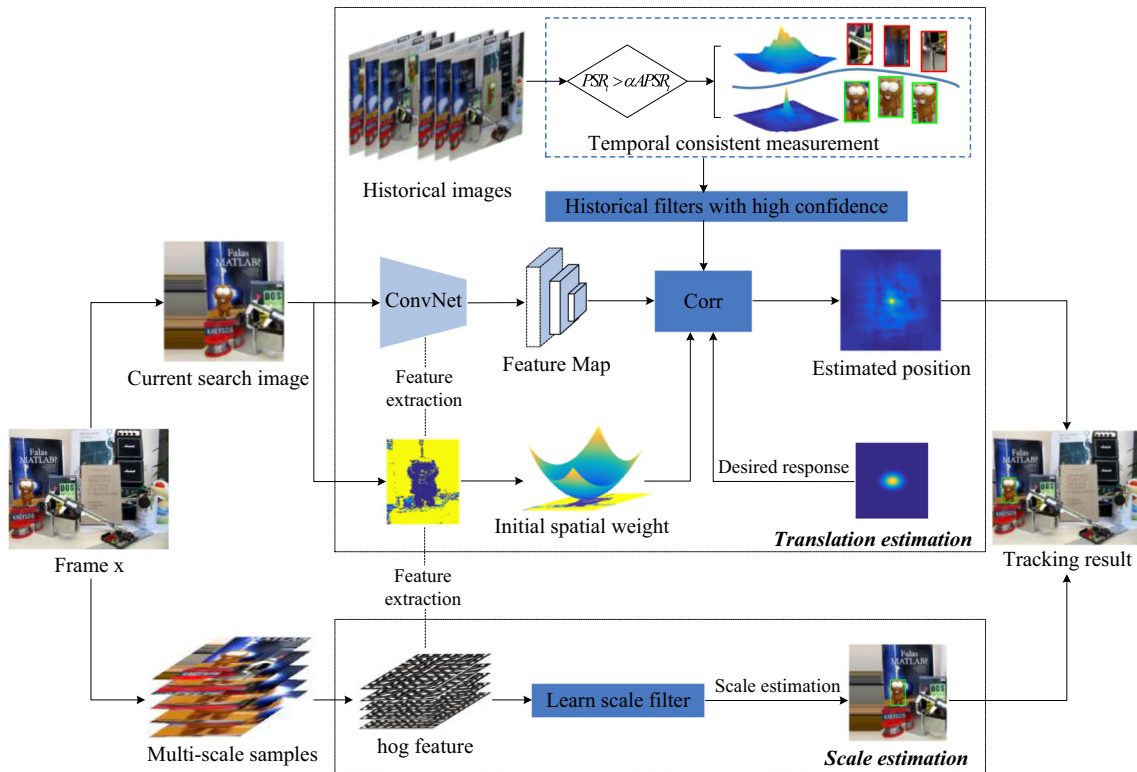


Fig. 1 The overall tracking flowchart of the TSCF tracker

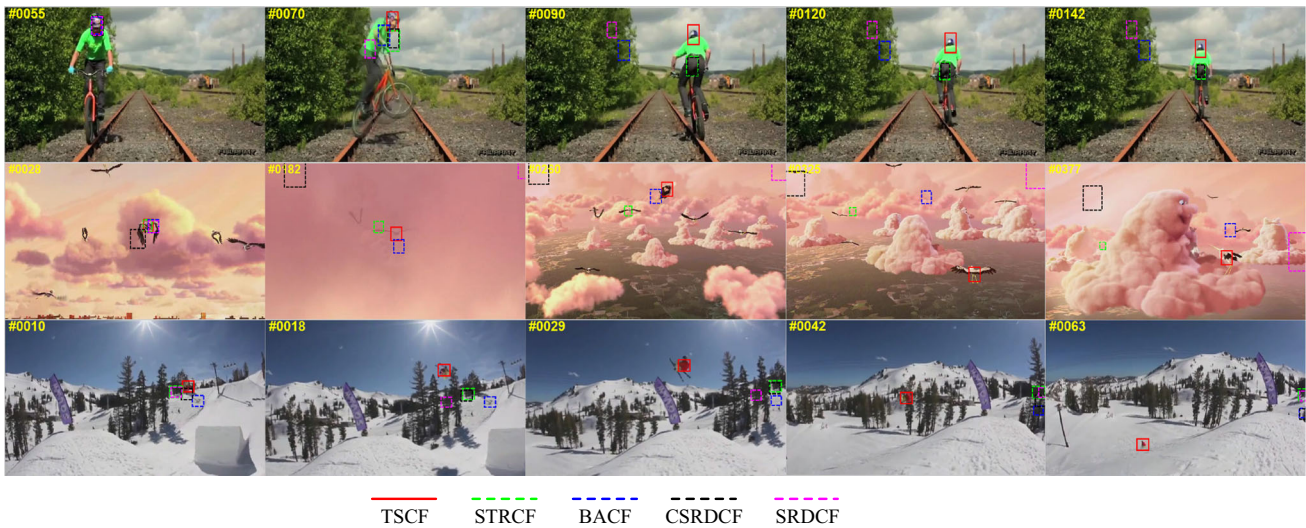


Fig. 2 Visualized tracking results of our TSCF tracker, STRCF, BACF, CSRDCF and SRDCF on 3 challenging video sequences of Biker, bird and skiing. Our tracker performs well throughout the tracking process

target can be further suppressed. Furthermore, we introduce the saliency map as a reference weight into the correlation filter-based model to learn the spatial weight adaptively. The adaptive spatial constraint model can be constructed as follows:

$$\operatorname{argmin}_f \frac{1}{2} \left\| \sum_{d=1}^{N_d} x^d * f^d - y \right\|^2 + \frac{\lambda_1}{2} \sum_{d=1}^{N_d} \|w \cdot f^d\|^2 + \frac{\lambda_2}{2} \|w - w^{\text{ref}}\|^2 \quad (4)$$

where  $x^d$  represents the  $d$ -th channel of the vectorized image feature.  $f^d$  represents the  $d$ -th channel of the filter.  $N_d$  is the number of channels in total.  $y$  is the desired output

response.  $w^{\text{ref}}$  is the reference weight.  $\lambda_1$  and  $\lambda_2$  are the regularization parameters. The symbols  $*$  and  $\cdot$  stand for circular convolution and element-wise multiplication, respectively.

By learning the spatial constraint with quadratic regularization, an adaptive penalty coefficient matrix is optimized in the training process. Thus, the tracker is more robust against cluttered background. Figure 3 shows the visualization of the learned spatial weights. It can be seen from the figure that the adaptive spatial constraint assigns a higher weight to the background region in the positive sample to suppress the background information, and it can adjust adaptively according to the target in different scenarios. Compared with the algorithm using the fixed coefficient constraint, the adaptive spatial constraint we proposed takes the reliability of the tracking region into consideration, which can suppress the distractors within the tracking box effectively.

### 3.2 Temporal consistent mechanism

The traditional DCF methods adopt a fixed learning rate to update a model without considering the reliability of the changing appearance. It makes the tracker bearing a high risk of model fluctuation and corruption. Existing temporal regularized trackers rely on the previous filter excessively, thereby the model may degrade by the introduction of inaccurate detection. In this case, we propose a temporal consistent mechanism to extract historical frames with high confidence, thus the temporal consistent regularization is constructed based on reliable appearance information. By capturing the reliable information of the target appearance from continuous historical frames, the adaptability of the model to the appearance changes is improved significantly. In this paper, we consider the peak-to-sidelobe ratio (PSR) [8] as the evaluation index of target reliability. It can be defined as:

$$\text{PSR}_t = \frac{R_{\text{max}}(x) - \mu_s(R_t(x))}{\sigma_s(R_t(x))} \quad (5)$$

where  $R_{\text{max}}(x)$  denotes the maximum value of the tracking response map corresponding to the image patch  $x$  in frame

$t$ .  $\mu_{s\lfloor}$  and  $\sigma_{s\lfloor}$  are the mean value and standard deviation of the sidelobe, respectively. Generally, the reliability of the tracking results is determined by the predefined experiential threshold. However, this approach is not universally suitable because the response amplitude of different video sequences is not uniform in the same range. Aiming at this problem, this paper applies a dynamic evaluation threshold calculated by the average PSR (APSR) value of historical frames:

$$\text{APSR}_t = \frac{\sum_{i=1}^T \text{PSR}^i}{T}, T = t - n + 1 \quad (6)$$

where  $\sum_{i=1}^T \text{PSR}^i$  represents the sum of PSR values of all tracking results,  $\text{PSR}^i$  is the PSR value of the  $i$ -th tracking response. The parameter  $T$  is the total number of frames from the initial image to the current image, and  $n$  is the initial frame number. In the case when the initial frame number is 1, the value of  $T$  is the same as that of the current frame  $t$ .

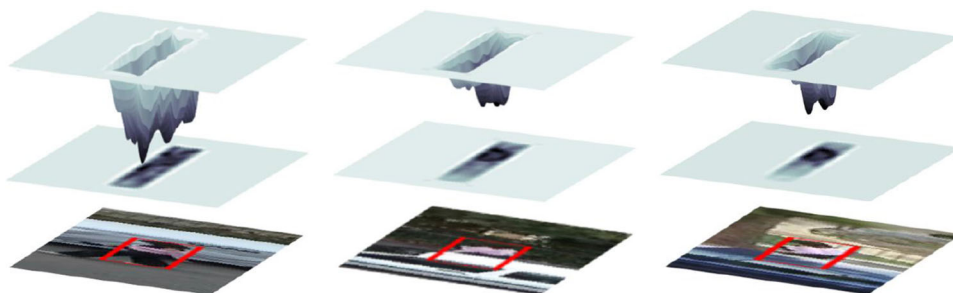
We employ APSR to describe the tracking quality in advance. The historical filter is reliable only when the certain condition is met: the PSR is higher than the average value APSR with a certain ratio  $\alpha$ . The evaluation criterion is shown in Eq. (7):

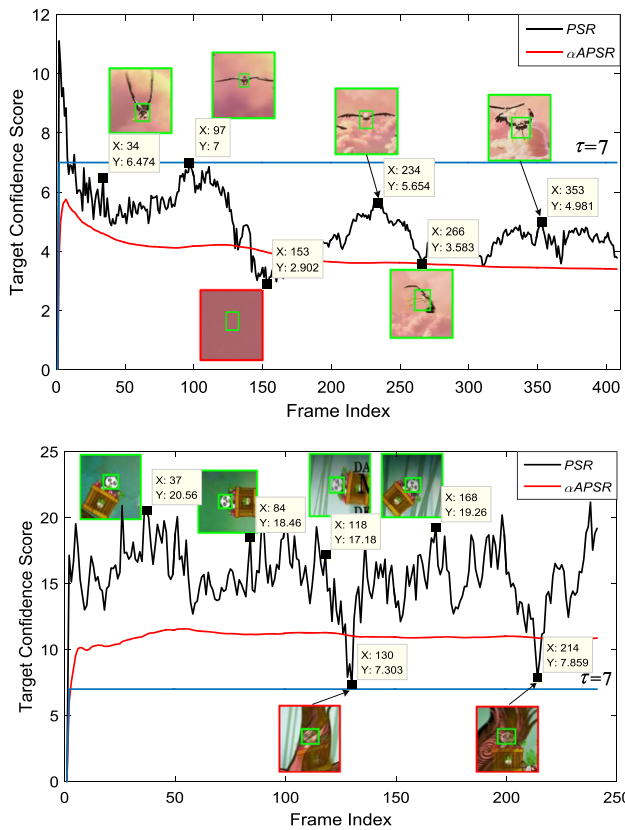
$$\text{PSR}_t > \alpha \text{APSR}_t \quad (7)$$

This evaluation strategy can effectively prevent the filter from learning inaccurate target representation information. To verify the validity of the dynamic threshold more intuitively, we compare the variation curves of different evaluation indicators of tracking results in typical challenge video sequences.

Figure 4 shows the variation trends of PSR and  $\alpha$  APSR in different video sequences. By comparing the curve of the two sequences, it can be seen that the fixed experiential threshold is not universally applicable. In panda, the target is subject to the cluttered background at frame 130 resulting in a significant decrease in PSR value to 7.303, but still higher than the predefined threshold  $\tau$  because the overall confidence value is at a relatively high level. The tracking results at this state introduce background information into the training process, which leads to poor

**Fig. 3** Visualization of the learned spatial weights. The target and the distractors are given different weights according to the different reliability in the tracking region





**Fig. 4** PSR and  $\alpha$  APSR variation curves corresponding to the video sequences bird1 (top) and panda (bottom).  $\tau$  is the predefined exponential threshold

discrimination of the filter. Meanwhile, it is obvious that the overall confidence score in bird1 is lower than the value of  $\tau$ , which means the predefined exponential threshold may misjudge the reliable tracking result. Therefore, exploiting the exponential threshold is not sufficient to reflect the tracking quality. In this work, the proposed historical average threshold can change adaptively for the sequences with different tracking response values, which breaks the limitation of fixed mode and evaluates the target consistency efficiently. In addition, the comparison of the same sequence demonstrates that the dynamic threshold is more accurate than the fixed threshold in distinguishing whether the target appearance is continuous. Through the target appearance consistency evaluation, a number of historical frames with high matching degree are screened for enriching the information of target appearance diversity.

### 3.3 Objective function of our TSCF tracker

Based on the above discussion, we combine the adaptive spatial constraint and the temporal consistent regularization into the CF-based tracker. This minimization problem takes the form:

$$\operatorname{argmin}_f \frac{1}{2} \left\| \sum_{d=1}^{N_d} x_t^d * f^d - y \right\|^2 + \frac{\lambda_1}{2} \sum_{d=1}^{N_d} \|w \cdot f^d\|^2 + \frac{\lambda_2}{2} \|w - w^{\text{ref}}\|^2 + \frac{\gamma}{2} \sum_{f_i} \|f - f_i\|^2 \quad (8)$$

where the second term denotes the adaptive spatial regularization.  $w$  identifies pixels that should be ignored in  $f^d$ , and the third term introduces a reference weight to prevent degradation of the constraint matrix. By learning an adaptive spatial constraint to suppress the background distractors, a more discriminative classifier can be learned on a huge set of negative patches. The last term is the temporal regularization component, in which  $f_i$  represents the screened historical frames with consistent appearance.  $\gamma$  is the regularization parameter. The variation of the target appearance can be effectively simulated according to the L2-norm minimization.

The model in Eq. (8) is a convex function, with which the global optimal solution can be obtained via the ADMM method. By introducing a dual variable  $h = f$  and the stepsize parameter  $\mu$ , the augmented Lagrangian form of Eq. (8) can be formulated as:

$$\mathcal{L}(f, h, w, s) = \frac{1}{2} \left\| \sum_{d=1}^{N_d} x_t^d * f^d - y \right\|^2 + \frac{\lambda_1}{2} \sum_{d=1}^{N_d} \|w \cdot h^d\|^2 + \frac{\lambda_2}{2} \|w - w^{\text{ref}}\|^2 \quad (9)$$

$$+ \sum_{d=1}^{N_d} (s^d)^T (f^d - h^d) + \frac{\mu}{2} \sum_{d=1}^{N_d} \|f^d - h^d\|^2 + \frac{\gamma}{2} \sum_{f_i} \|f - f_i\|^2$$

where  $s$  denotes the Lagrange multiplier,  $\mu$  is the penalty parameter that controls the convergence rate. Equation (10) is obtained when we set  $z = s/\mu$ :

$$\mathcal{L}(f, h, w, s) = \frac{1}{2} \left\| \sum_{d=1}^{N_d} x_t^d * f^d - y \right\|^2 + \frac{\lambda_1}{2} \sum_{d=1}^{N_d} \|w \cdot h^d\|^2 + \frac{\lambda_2}{2} \|w - w^{\text{ref}}\|^2 + \frac{\mu}{2} \sum_{d=1}^{N_d} \|f^d - h^d + s^d\|^2 + \frac{\gamma}{2} \sum_{f_i} \|f - f_i\|^2 \quad (10)$$

According to the ADMM method, the objective function can be decomposed into multiple subproblems. The iterative solution steps are:

$$\begin{cases} f^{(i+1)} = \underset{f}{\operatorname{argmin}} \frac{1}{2} \left\| \sum_{d=1}^{N_d} x_t^d * f^d - y \right\|^2 + \frac{\mu}{2} \sum_{d=1}^{N_d} \|f^d - h^d + s^d\|^2 + \frac{\gamma}{2} \sum_{f_i} \|f - f_i\|^2 \\ h^{(i+1)} = \underset{h}{\operatorname{argmin}} \frac{\lambda_1}{2} \sum_{d=1}^{N_d} \|w \cdot h^d\|^2 + \frac{\mu}{2} \sum_{d=1}^{N_d} \|f^d - h^d + s^d\|^2 \\ w^{(i+1)} = \underset{w}{\operatorname{argmin}} \frac{\lambda_1}{2} \sum_{d=1}^{N_d} \|w \cdot h^d\|^2 + \frac{\lambda_2}{2} \|w - w^{\text{ref}}\|^2 \\ s^{(i+1)} = s^{(i)} + f^{(i+1)} - h^{(i+1)} \end{cases} \quad (11)$$

Each subproblem can be, respectively, solved in details as follows:

Subproblem *f*: For the computational efficiency, we transform the objective function of subproblem of *f* into Fourier domain according to the Parseval’s theorem:

$$\underset{f}{\operatorname{argmin}} \frac{1}{2} \left\| \sum_{d=1}^{N_d} \hat{x}_t^d \cdot \hat{f}^d - \hat{y} \right\|^2 + \frac{\mu}{2} \sum_{d=1}^{N_d} \|\hat{f}^d - \hat{h}^d + \hat{s}^d\|^2 + \frac{\gamma}{2} \sum_{f_i} \|\hat{f} - \hat{f}_i\|^2 \quad (12)$$

Considering the processing of all channels of each pixel, we let  $\mathcal{V}_i(f) \in \mathcal{R}^{N_d}$  represent the vector composed of the *i*-th element of *f* along all  $N_d$  channels. The closed-form solution for  $\mathcal{V}_i(\hat{f})$  is obtained by the following derivation:

$$\mathcal{V}_i(\hat{f}) = \left( \mathcal{V}_i(\hat{x})\mathcal{V}_i(\hat{x})^T + \mu I + \sum_{f_i} \gamma \right)^{-1} \left( \mathcal{V}_i(\hat{x})\hat{y} + \mu\mathcal{V}_i(\hat{h}) - \mu\mathcal{V}_i(\hat{s}) + \gamma \sum_{f_i} \mathcal{V}_i(\hat{f}_i) \right) \quad (13)$$

Since  $\mathcal{V}_i(\hat{x})$  is a column vector and  $\mathcal{V}_i(\hat{x})\mathcal{V}_i(\hat{x})^T$  is a rank-1 matrix, Eq. (13) can be re-expressed by the Sherman–Morrison formula:

$$\mathcal{V}_i(\hat{f}) = \frac{1}{\mu + \sum_{f_i} \gamma} \left( I - \frac{\mathcal{V}_i(\hat{x})\mathcal{V}_i(\hat{x})^T}{\mu + \sum_{f_i} \gamma + \mathcal{V}_i(\hat{x})^T \mathcal{V}_i(\hat{x})} \right) \left( \mathcal{V}_i(\hat{x})\hat{y} + \mu\mathcal{V}_i(\hat{h}) - \mu\mathcal{V}_i(\hat{s}) + \gamma \sum_{f_i} \mathcal{V}_i(\hat{f}_i) \right) \quad (14)$$

Then, the optimal *f* can be obtained by the inverse discrete Fourier transform (IDFT) of  $\mathcal{V}_i(\hat{f})f$ .

Subproblem *h*: By taking the derivative of the second equation in Eq. (11), the closed-form solution of *h* is

$$h = (\lambda_1 W^T W + \mu I)^{-1} (\mu f + \mu s) = \frac{\mu(f + s)}{\lambda_1 w \cdot w + \mu I} \quad (15)$$

where  $W = \operatorname{diag}(w)$  is the  $N_d MN \times N_d MN$  block diagonal matrix.

Subproblem *w*: The adaptive spatial constraint weight is denoted as *w*, we apply an additional ADMM solver to

optimize it. By introducing the penalty parameter  $\eta$  and the auxiliary variable  $v = w$ , the augmented Lagrange formula of subproblem *w* can be expressed as:

$$\mathcal{L}(w, v, s) = \frac{\lambda_1}{2} \sum_{d=1}^{N_d} \|w \cdot h^d\|^2 + \frac{\lambda_2}{2} \|v - w^{\text{ref}}\|^2 + s^T (w - v) + \frac{\eta}{2} \|w - v\|^2 \quad (16)$$

Let  $u = s/\mu$ , the above formulation is equivalent to the following one:

$$\mathcal{L}(w, v, u) = \frac{\lambda_1}{2} \sum_{d=1}^{N_d} \|w \cdot h^d\|^2 + \frac{\lambda_2}{2} \|v - w^{\text{ref}}\|^2 + \frac{\eta}{2} \|w - v + u\|^2 \quad (17)$$

The closed-form solution of the subproblem *w* obtained by the derivation of Eq. (17) is:

$$w = \frac{\eta(v - u)}{\lambda_1 \sum_{d=1}^{N_d} h^d \cdot h^d + \eta I} \quad (18)$$

Similarly, the solution of subproblem *v* is:

$$v = \frac{\lambda_2 w^{\text{ref}} + \eta(w + u)}{\lambda_2 + \eta} \quad (19)$$

In the process of solving subproblem *w*, the updating scheme of Lagrange multiplier *u* and the selection of the stepsize parameter  $\eta$  are as follows:

$$u^{(i+1)} = u^{(i)} + w^{(i+1)} - v^{(i+1)} \quad (20)$$

$$\eta^{(i+1)} = \min(\eta_{\max}, \varepsilon \eta^{(i)}) \quad (21)$$

where  $\varepsilon$  denotes the scale factor. Different  $\eta$  in each iteration improves the convergence with the less dependent on the initial choice of parameter. The background can be suppressed effectively by learning the adaptive spatial weight *w* through an additional ADMM solver.

Lagrange Multiplier *s*: The updating scheme of Lagrange multiplier vector *s* is as follows:

$$s^{(i+1)} = s^{(i)} + f^{(i+1)} - h^{(i+1)} \quad (22)$$

$f^{(i+1)}$  and  $h^{(i+1)}$  are the solutions of the subproblem in the  $(i + 1)$ -th iteration. And the stepsize parameter is update with  $\mu^{(i+1)} = \min(\mu_{\max}, \beta \mu^{(i)})$ .

### 3.4 Online detection and scale estimation

In the detection stage, the target’s position is detected by correlating the learned filter with image feature. Let *x* represents the *d*-th channel of the vectorized image feature extracted from the current frame, the response map at all locations can be calculated by the following function:



$$r = \mathcal{F}^{-1} \left( \sum_{d=1}^{N_d} x^d \cdot \hat{f}^d \right) \quad (23)$$

where  $\mathcal{F}^{-1}$  represents the IDFT operator. The maximum value in the response map corresponds to the target location.

the historical frame in a simple way that makes the model maintain stabilized to the target appearance changes. The outline of our TSCF tracker is summarized in Algorithm 1.

---

**Algorithm 1** The TSCF tracking algorithm.

---

**Input:** Image patch  $I_t$ , previous target position  $p_{t-1}$ , scale  $s_{t-1}$ , location filter  $f_{t-1}$ , adaptive spatial constraint  $w_{t-1}$ , scale filter  $f_{t-1}^s$ .

**Output:** Estimated target position  $p_t$  and scale  $s_t$  in frame  $t$ .

**1: repeat**

2: Crop an image patch  $x_t$  and scale samples set centered at  $(x_{t-1}, y_{t-1})$  from  $I_t$ .

3: **if**  $t = 1$  **then**

4: Extract the background log-likelihood map to construct the initial reference weight  $w^{ref}$  of the spatial constraint.

5: **end if**

6: Extract hog and deep features centered at  $(x_{t-1}, y_{t-1})$  from  $I_t$ .

7: Optimize the spatial constraint  $w$ , correlation filter  $f$  and scale filter  $s$ .

8: Apply Eq. (7) to generate a reliable filter set from the historical filters.

9: Compute correlation confidence score for each candidate by multiply the filter  $f$  with the image patch features  $x$  according to Eq. (23).

10: Estimate the new target position  $p_t$  and compute current scale factor.

11: Update appearance model  $x_t$  by linear interpolation.

**12: until the end of video sequences.**

---

For scale estimation, the method in [21] is adopted to learn a one-dimensional filter based on the scale pyramid with HOG features in this paper. Specifically, we extract multiple scale candidates centered at the same predicted location and thereby obtain the optimal scale by calculating the maximum response value. Compared with multi-scale deep features, exploiting shallow features for scale estimating can significantly improve the efficiency of the tracking algorithm.

### 3.5 Model update

In this work, we update the appearance model adaptively to accommodate the appearance variations of the target. The online adaptation at frame  $t$  is formulated as:

$$x_t^{\text{model}} = (1 - \theta)x_{t-1}^{\text{model}} + \theta x_t^{\text{model}} \quad (24)$$

where  $\theta$  denotes the online learning rate parameter and  $x_{t-1}^{\text{model}}$  denotes the template model of frame  $t - 1$ . This updating approach combines the current information with

## 4 Experiments

Comprehensive experimental evaluation and discussion of the proposed TSCF tracker are demonstrated in this section. Section 4.1 is the experiment setup that describes the benchmark datasets and evaluation indicators in experiments. Experimental details and parameter settings are provided in Sec. 4.2. Section 4.3 evaluates the performance of the proposal and the related trackers on different benchmarks. Section 4.4 is the attribute-based analysis of all trackers in comparison. Section 4.5 gives a qualitative comparison with related trackers on several representative sequences. Furthermore, an evaluation on VOT2016 is displayed in Sec. 4.6. Section 4.7 demonstrates the effectiveness of each component of the proposed tracker with an ablation study. Section 4.8 studies the impact of the key parameters on tracking performance. Section 4.9 discusses the failure cases and analyzes the shortcomings of the proposed method in detail. Finally, Sec. 4.10 summarizes

all the observations and lessons learned from this research for future work propositions.

#### 4.1 Experiment step and methodology

*Datasets and evaluation metrics:* This paper performs the evaluation on the OTB-2013, OTB-2015, Temple-Color-128 and VOT2016 datasets. OTB-2013 dataset contains 50 fully annotated sequences, and OTB-2015 is an extension of the former with 100 sequences. These two datasets are manually labeled with 11 different attributes, including illumination variation (IV), scale variation (SV), occlusion (OCC), deformation (DEF), motion blur (MB), fast motion (FM), in-plane rotation (IPR), out-of-plane rotation (OPR), out-of-view (OV), background clutters (BC), low resolution (LR). The tracking performance is evaluated by the distance precision plots and overlap success plots in one-pass evaluation (OPE).

Temple-Color-128 dataset contains 128 color sequences in different tracking scenarios. The evaluation metric is the same as the OTB benchmark.

VOT2016 dataset consists of 60 challenging video sequences, with which it takes the expected average overlap (EAO), accuracy value (A) and robustness value (R) as the evaluation metric.

*Experimental platform:* We implement our TSCF tracker on Matlab with the MatConvNet toolbox. The experiments are performed on a 2.90 GHz Intel i5-9400F Core CPU with 16 GB RAM and a NVIDIA GeForce RTX 2070 GPU. The proposed tracker runs at a speed of 17.2 frames per second (FPS) on the OTB-2015 dataset.

#### 4.2 Implementation parameters

For all experiments, the regularization parameters of spatial and temporal components are set as  $\lambda_1 = 1.2$ ,  $\lambda_2 = 0.001$  and  $\gamma = 1.1$ , respectively. Within the ADMM optimization process of the filter, the initial value, maximum value and scale factor in the update step of the stepsize parameter are set as  $\mu^{(0)} = 0.89$ ,  $\mu_{\max} = 100$  and  $\beta = 11$ . The maximum number of iterations is 2. Similarly, optimization parameters of the spatial constraint are set as  $\eta^{(0)} = 1$ ,  $\eta_{\max} = 10^3$  and  $\varepsilon = 10$  while the iteration is 3 to make sure the accuracy and efficiency are balanced. For the scale estimation, five scales are considered and the scale step is chosen as 1.01. The learning rate of target appearance model is 0.0186. In addition, we employ the public available codes of other trackers for fair comparison.

#### 4.3 Overall performance evaluation of the TSCF tracker

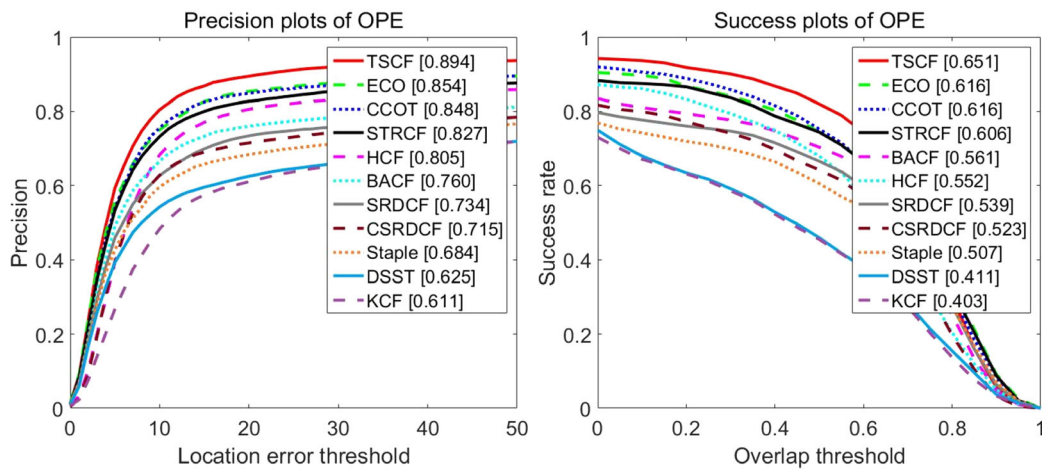
To verify the effectiveness of the tracking performance, we evaluate the proposed TSCF tracker on OTB-2013, OTB-2015 and Temple-Color-128 datasets with comparison to 10 representative trackers including KCF [10], DSST [21], Staple [24], SRDCF [12], STRCF [15], CSRDCF [31], BACF [13], HCF [33], CCOT [25] and ECO [26]. Among them, ECO, CCOT and HCF are convolutional neural networks (CNNs)-based tracking methods, while others are correlation filter-based trackers. The overall performance evaluation results are shown in Fig. 5.

Figure 5 illustrates the distance precision and overlap success plots of all compared trackers on OTB-2013, OTB-2015 and Temple-Color-128 benchmarks. It can be seen that our method achieves superior performance on all three datasets. Specifically, our TSCF tracker ranks first with the area under the curve (AUC) value of 68.7% and the distance precision (DP) score of 91.3% on OTB-2015, which are significantly improved 21.3% and 22.4% compared with the baseline tracker KCF. By learning the spatial constraint adaptively, our tracker is increased by 12.5%, 10.3% and 13.1% in distance precision over congeneric trackers CSRDCF, BACF and SRDCF, respectively. Compared to the STRCF that adopts single previous filter for the temporal regularization, our TSCF tracker meets 5.6% tracking precision improvement through acquiring the appearance diversity information from multi-frame constraints. Our proposal still presents satisfactory performance with an accuracy of 3.0%, 4.6% and 7.6% higher than ECO, CCOT and HCF trackers which behave excellent properties based on deep features. Additionally, a favorable result also realized on the Temple-Color-128 dataset: our tracker improves the precision and overlap rate by 1.3% and 0.9% respectively than the second-best tracker. Overall, the proposed TSCF tracker achieves superiority and effectiveness against state-of-art trackers.

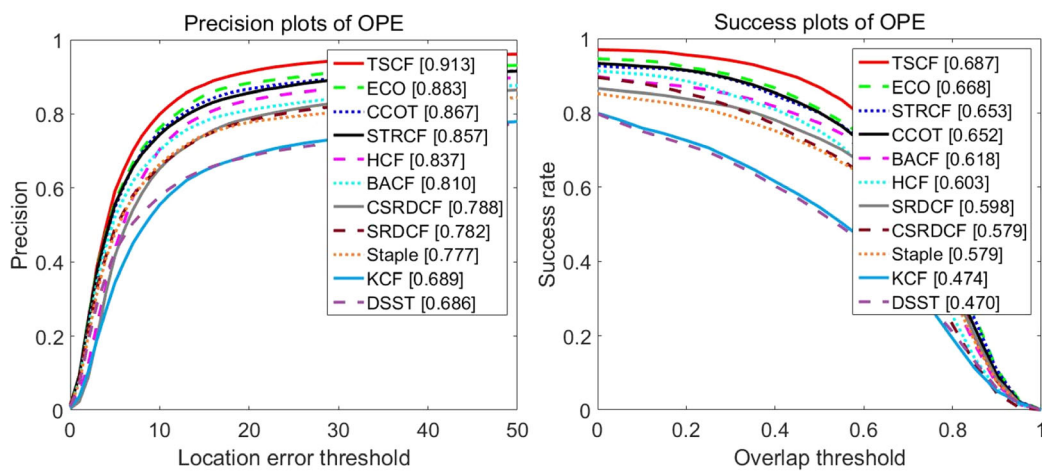
#### 4.4 Attribute-based comparison

In this section, an attribute-based analysis of all compared trackers is carried out on the OTB-2015 dataset. Table 1 shows the DP score comparison results of all compared trackers based on 11 attributes.

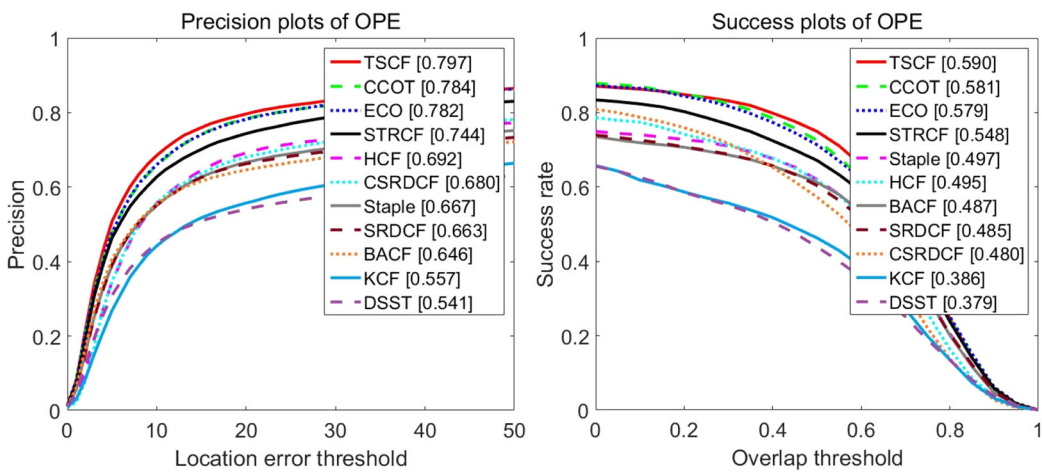
From Table 1, the proposed tracker achieves the best performance in 7 of the challenge attributes including IV, SV, DEF, FM, IPR, OPR and BC, especially it outperforms in terms of BC, DEF, OPR and IV with 4.8%, 4.7%, 3.4% and 2.5% improvement respectively than the second-best tracker. For sequences with attributes OCC and MB, our approach enjoys the second rank with scores of 88.8% and



(a) Results on the OTB-2013 dataset



(b) Results on the OTB-2015 dataset



(c) Results on the Temple-Color-128 dataset

Fig. 5 Precision and success plots on the OTB-2013, OTB-2015 and Temple-Color-128 datasets using an OPE

84.6%. Conclusively, the above experimental results reveal that our tracker achieves favorable performance in multiple challenging tracking scenarios. We attribute this to the combination of the adaptive spatial constraint and the temporal consistent regularization, which not only enable the filter to suppress background distractors but also maintain the target diversity effectively. The success plot of different trackers with 11 attributes on OTB-2015 is shown in Fig. 6. From the results in the figure, we can see that our proposed tracker has the highest AUC score on 8 attributes of BC, DEF, IPR, IV, OCC, OPR, OV and SV. Similarly, the proposed tracker also obtains excellent results on the Temple-Color-128 benchmark that effectively exhibit the competitiveness of the algorithm. We report the success plots for 6 attributes on the Temple-Color-128 benchmark in Fig. 7.

#### 4.5 Qualitative evaluation

We perform a qualitative analysis on 16 representative challenge sequences for evaluating the tracking performance intuitively. For clear visualization, the tracking results of the proposed method and competitive trackers are displayed in Fig. 8. The videos (from top to bottom) are Ball\_ce2, Bird1, Busstation\_ce1, Eagle\_ce, Face\_ce, Fish\_ce1, Fish\_ce2, Messi\_ce, Panda, Railwaystation\_ce, Singer\_ce2, Skating\_ce2, Surf\_ce1, Surf\_ce4, Yo-yos\_ce2, Yo-yos\_ce3. Our tracker performs well in the video sequences with multiple challenging attributes.

**Deformation** In the Bird1 sequence, the target has obvious deformation due to the change of the posture. Our tracker is the only one works successfully during the whole tracking process, and the remaining 10 trackers lost the target. The temporal consistent constraint proposed in this paper can effectively maintain the diversity of the target's

appearance, which makes our tracker identify and locate the target accurately even if the appearance changes greatly. Panda is a long-term sequence with the challenge of deformation. At frame 539, SRDCF, KCF and DSST drift away to the background region as the continuous deformation of the target, while Staple, ECO, CCOT, STRCF, CSRDCF and HCF obtain inaccurate location and size of the target. At frame 709, Staple also lost the target. Due to the deformation of the target and the low video resolution, ECO drifts at frame 980, while only our tracker and CSRDCF track the target successfully. In general, our tracking algorithm performs well on video sequences with the deformation attribute.

**Background clutter** The target in the sequences of Eagle\_ce, Face\_ce and Singer\_ce2 suffers from the influence of background clutter. At frame 60 of Eagle\_ce sequence, ECO, STRCF, BACF, SRDCF, KCF and DSST cannot identify the specified target, while CCOT, CSRDCF and Staple accumulate plenty of background information in the tracking box. At frame 71, only our tracker and HCF can locate the target accurately. Since the adaptive spatial constraint proposed in this work can effectively suppress the unreliable information within the tracking box, our tracker can cope well with the problem of model drift caused by background clutter. Face\_ce sequence contains a lot of distractors which are similar to the target. At frame 34, only our tracker performs well, the remaining 10 trackers identify the background of the search region as the target. Till the 619th frame, our tracker is the only one that can locate the target accurately. In Singer\_ce2 sequence, 7 trackers including ECO, CCOT, HCF, CSRDCF, SRDCF, KCF and DSST drift away from the target because of the background clutter. Due to the application of the adaptive spatial constraint, our tracker can handle the challenge of background clutter effectively.

**Table 1** Precision comparison results of 11 attributes on the OTB-2015 benchmark. The top three results are highlighted in red, blue and green, respectively

	TSCF	ECO	CCOT	STRCF	HCF	BACF	CSRDCF	SRDCF	Staple	KCF	DSST
ALL	0.913	0.883	0.867	0.857	0.837	0.810	0.788	0.782	0.777	0.689	0.686
IV	0.912	0.873	0.887	0.837	0.839	0.803	0.755	0.781	0.778	0.724	0.720
SV	0.876	0.858	0.850	0.827	0.789	0.755	0.733	0.733	0.713	0.624	0.642
OCC	0.888	0.890	0.853	0.812	0.774	0.731	0.721	0.727	0.724	0.632	0.602
DEF	0.902	0.855	0.803	0.842	0.789	0.769	0.760	0.730	0.747	0.619	0.560
MB	0.846	0.867	0.838	0.797	0.772	0.715	0.731	0.745	0.676	0.578	0.544
FM	0.865	0.846	0.842	0.780	0.786	0.772	0.717	0.753	0.693	0.604	0.556
IPR	0.899	0.858	0.880	0.812	0.864	0.792	0.766	0.743	0.768	0.701	0.711
OPR	0.906	0.872	0.865	0.837	0.802	0.767	0.724	0.731	0.727	0.666	0.653
OV	0.839	0.841	0.843	0.710	0.602	0.705	0.640	0.561	0.619	0.453	0.428
BC	0.893	0.837	0.821	0.845	0.825	0.776	0.743	0.755	0.727	0.692	0.680
LR	0.811	0.841	0.851	0.743	0.869	0.741	0.746	0.666	0.610	0.560	0.602



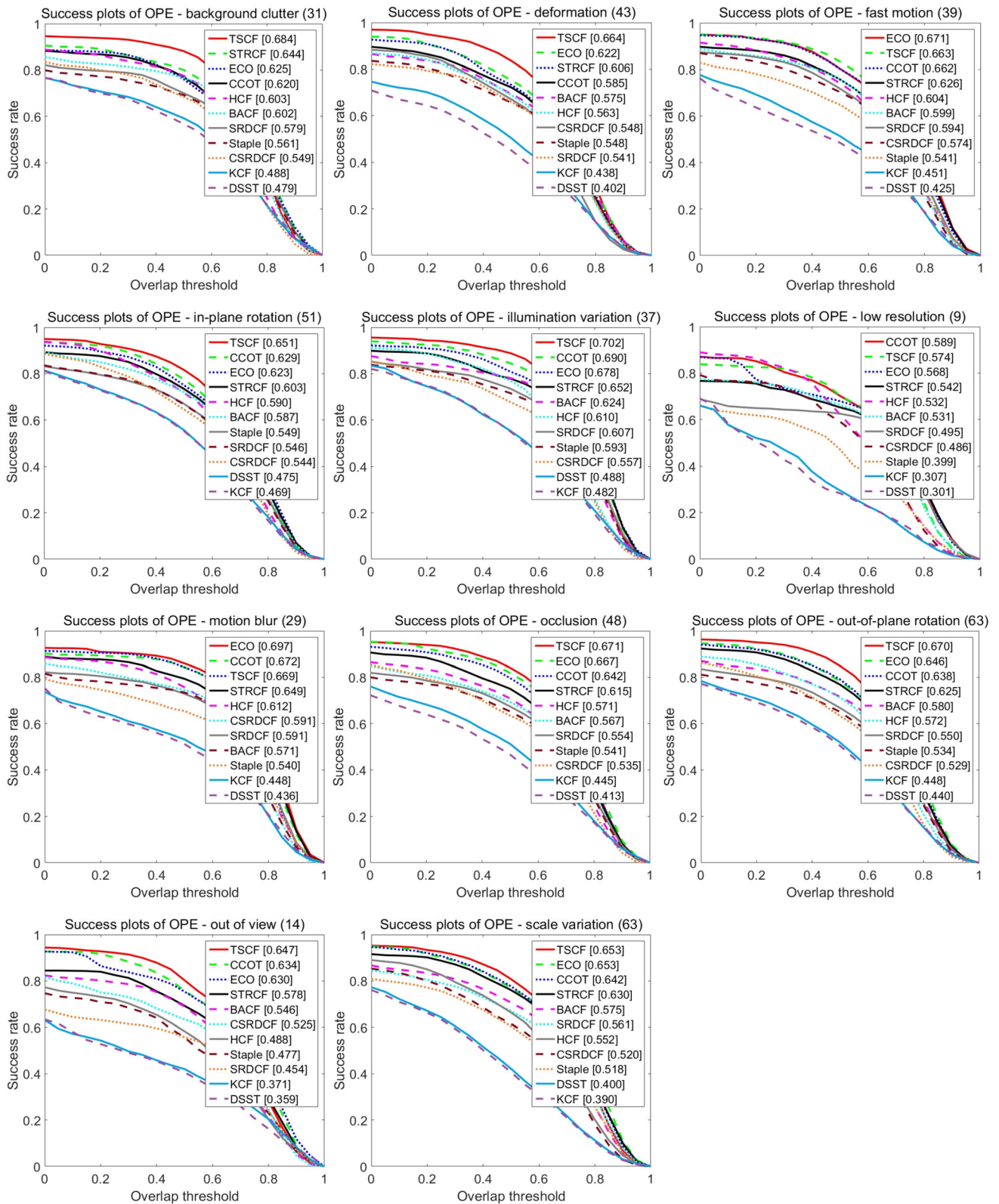
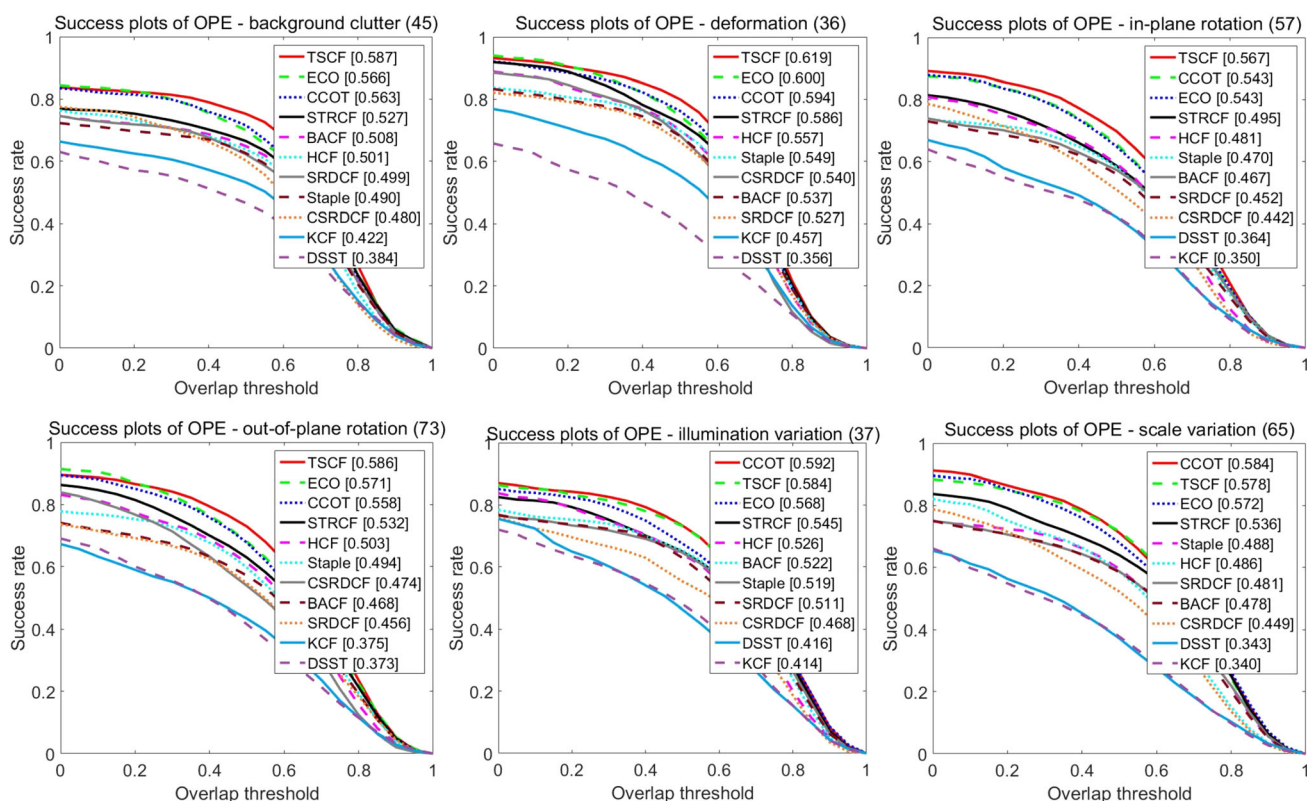


Fig. 6 Success plots of 11 attributes on the OTB-2015 dataset



**Fig. 7** Success plots of 6 attributes on the Temple-Color-128 dataset

**Occlusion** *Ball\_ce2*, *Busstation\_ce1*, *Fish\_ce1*, *Fish\_ce2*, *Messi\_ce*, and *Railwaystation\_ce* are typical video sequences with the challenge of target occlusion. In *Ball\_ce2*, only our tracker, ECO, CCOT, and HCF achieve accurate location in the whole tracking process. At frame 51 of *Busstation\_ce1*, all trackers except our proposal have the problem of model drift due to the occluded target. At frame 116, only our tracker and CCOT can identify the target in the scenario where the target is occluded. In sequences of *Fish\_ce1* and *Fish\_ce2*, our tracker can also cope well with the problem of target occlusion. The occlusion occurs at frame 190 of *Messi\_ce* sequence, only our tracker, CCOT and SRDCF can track the target successfully in subsequent frames. In *Railwaystation\_ce* sequence, the target is occluded heavily for several times, our tracker, ECO and SRDCF achieve favorable results in all frames. In conclusion, our tracker performs well under the challenge of occlusion.

**In-plane rotation** The target faces the challenge of rotation due to the changing of posture in *Skating\_ce2*, *Surf\_ce1* and *Surf\_ce4* sequences. At the beginning of those videos, most trackers can track the target accurately. In *Skating\_ce2*, all the trackers except for ours give inaccurate estimation results when the target rotates. At frame 17 of *Surf\_ce4* sequence, only our tracker gives the exact location of the target. BACF and SRDCF lost the target at

frame 76, while the tracking boxes of HCF, CCOT, ECO, STRCF and DSST contain only a part of the target and introduce a lot of background information. Till the 99th frame, our tracker is still able to locate the target accurately.

**Fast motion** In the whole tracking process of *Yo-yos\_ce2* and *Yo-yos\_ce3* sequences, only our tracker, STRCF and CCOT can track the target steadily, the other 8 trackers fail to track the target when it moves rapidly. Therefore, the tracker proposed in this paper can handle the challenge of fast motion well.

#### 4.6 Evaluation on VOT2016

Different from the OTB benchmark, the VOT dataset is dominated by short-term sequences with relatively high resolution. The tracker resets automatically whenever the overlap does not exist between the prediction and ground-truth. We further perform an evaluation of our algorithm in terms of accuracy, robustness and expected average overlap (EAO) on VOT2016. The accuracy metric is the average overlap rate of the bounding box. The robustness is measured by the number of tracking failures. And the expected average overlap is a quite important metric for evaluation which averages the no-reset overlap of trackers. The compared trackers in this experiment include STRCF,





**Fig. 8** Qualitative evaluation of our proposed method with comparison to 10 trackers (denoted in different colors and lines) on 16 challenging video sequences

BACF, SRDCF, HCF and Staple, which are related to our TSCF tracker. From the evaluation results listed in Table 2, it can be seen that our tracker achieves the best EAO score (0.2925) with the best accuracy (0.55) and a favorable robustness score (0.65). The visualization of expected overlap curves and scores on the VOT2016 dataset can be found in Fig. 9. Generally, our implementation performs better than most of other trackers.

#### 4.7 Ablation analysis

To verify the contribution of the components in the proposed tracker, we perform an ablation analysis on the OTB-2015 dataset. We propose three variants of our approach for further analysis. TSCF-T is the proposed TSCF tracker without considering the spatial consistent, TSCF-S is the TSCF tracker that does not apply the temporal consistent regularization, and TSCF-TS is the TSCF tracker that contains neither. For comparison, we report precision and success plots in Fig. 10, and results for 11 attributes in Table 3. From the comparison results in Fig. 10, the TSCF-S with adaptive spatial constraints introduced into the baseline tracker achieves an AUC improvement of 0.9% and a DP improvement of 0.9% relative to the TSCF-TS tracker, while the decreases of 1.0% and 2.4% are obtained compare with our TSCF tracker. Similarly, when we incorporate the proposed temporal consistent regularization in the TSCF-TS, the precision is increased by 2.3%, in which case the precision score is reduced by 1% compared with our TSCF tracker. The data results for 11 attributes in Table 3 reflect the contributions of each component sufficiently. In general, these data speak volumes about the effectiveness of our proposal.

#### 4.8 Key parameters analysis

The parameter values in the experiment are obtained from extensive experiments under the condition of balanced precision and speed. In this section, we analyze the effect of the key parameters of our method on tracking performance in detail, including the spatial regularization coefficients  $\lambda_1$ ,  $\lambda_2$ , and the temporal consistent coefficient  $\gamma$ . Among them,  $\lambda_1$  adjusts the fitting degree of the filter with

penalty weights,  $\lambda_2$  constrains the optimization of the spatial weight, and  $\gamma$  is used to smooth the learning of the filter. We verify the validity of the parameters by comparing the precision and success rate with different parameter values on the OTB-2015 benchmark. Specifically, we adjust the parameters in the experiment until we find a parameter that can fully demonstrate the tracking performance. For fair comparison, we keep the other parameters unchanged when adjusting one parameter. The comparison results and rankings are given in Fig. 11.

In Fig. 11, the influence of the spatial regularization parameters and the temporal consistent parameter on tracking performance is displayed in detail. By adjusting the parameters in the experiment, we can intuitively observe and select the value that can make the tracking algorithm achieve the optimal performance. From (a) and (b) in Fig. 11, we can see that the algorithm can obtain better tracking accuracy when the spatial regularization parameters are set as  $\lambda_1 = 1.2$  and  $\lambda_2 = 0.001$ , respectively. In Fig. 11 (c), the overall tracking performance reaches the peak value when the temporal constraint parameter is  $\gamma = 1.1$ . According to the analysis, we set the above optimal parameter values in all experiments to ensure satisfactory tracking performance of the proposed algorithm.

#### 4.9 Failure case analysis

Despite that the proposed method achieves satisfactory results on the above tracking benchmarks, it still has some limitations. Specifically, our proposal underperforms on attribute challenges such as low resolution, drastic scale variation, fast motion and blur. This section provides a discussion about the failure cases observed during the experiment. Figure 12 shows several failure cases of our TSCF tracker. The videos (from top to bottom) are Pool\_ce3, TennisBall\_ce, Ball\_ce1, Cup\_ce, Trans, Diving.

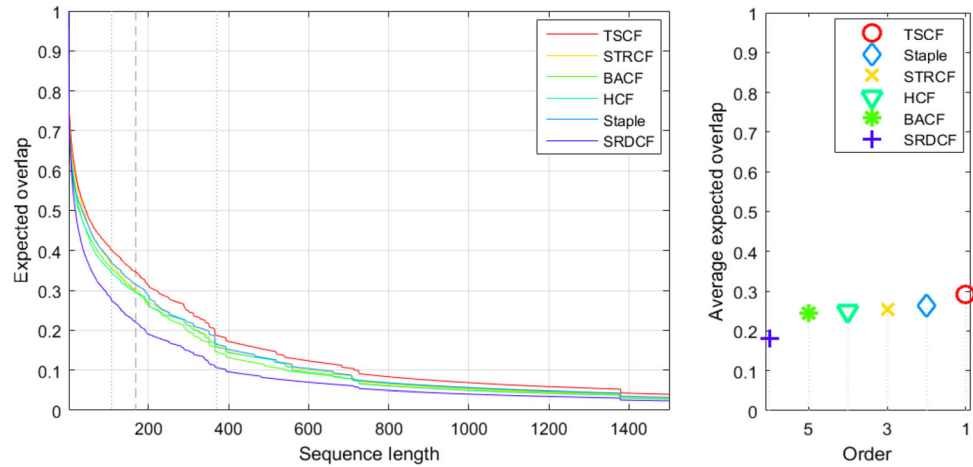
Pool\_ce3 and TennisBall\_ce are typical video sequences with low resolution. In Fig. 12(a), it can be seen that since the target is too small, the tracker can hardly extract the main features of the target. The lack of appearance information directly leads to tracking failure. In addition, there is also the problem of motion blur in these two videos, which makes the feature of the target more difficult to identify. Figure 12(b) shows two failed cases caused by fast motion and blur. From the video Ball\_ce1, we can see that our tracker can locate the target well before it moves rapidly. At frame 12, motion blur is generated due to the fast movement of the target. At this time, the external contour of the object is blurred and the details is lost, which causes the tracker to drift to the background area. Similarly, the object in the video Cup\_ce is also difficult to

**Table 2** Performance evaluation on the VOT2016 benchmark. The top three results are highlighted in red, blue and green, respectively

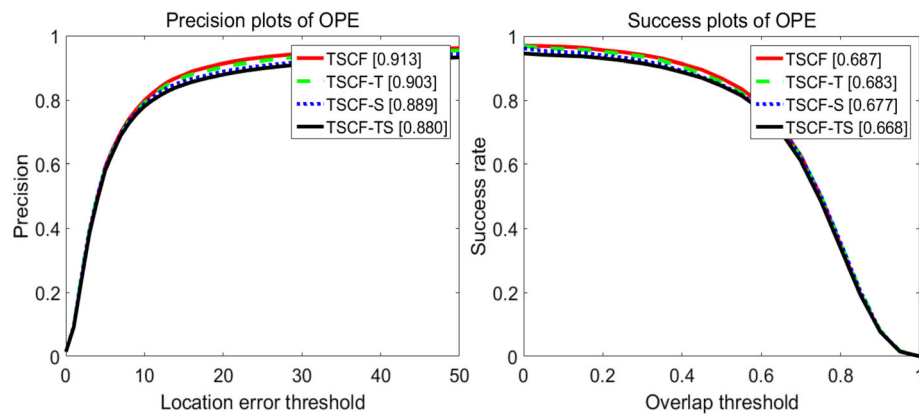
	TSCF	Staple	STRCF	BACF	HCF	SRDCF
EAO	0.2925	0.2628	0.2545	0.2435	0.2483	0.1810
Accuracy	0.5476	0.5398	0.5258	0.5383	0.5235	0.5216
Robustness	0.65	0.58	0.59	0.55	0.63	0.42



**Fig. 9** Expected overlap curves and scores on the VOT2016 benchmark



**Fig. 10** Experiment results of different cases on the OTB-2015 dataset using an OPE



**Table 3** Ablation analysis with 11 attributes on the OTB-2015 benchmark

	ALL	IV	SV	OCC	DEF	MB	FM	IPR	OPR	OV	BC	LR
TSCF	0.913	0.912	0.876	0.888	0.902	0.846	0.865	0.899	0.906	0.839	0.893	0.811
TSCF-T	0.903	0.894	0.861	0.867	0.898	0.817	0.839	0.879	0.889	0.823	0.871	0.805
TSCF-S	0.889	0.899	0.854	0.844	0.883	0.821	0.828	0.866	0.878	0.823	0.854	0.811
TSCF-TS	0.880	0.870	0.843	0.837	0.862	0.820	0.823	0.864	0.872	0.810	0.829	0.800

identify because of its fast motion and blur attributes. In this case, the representation of general features is too limited for constructing a robust target model, which inevitably leads to tracking failure. Moreover, the scale estimation of our TSCF tracker is proportional, so it is difficult to fit the exact size when the aspect ratio of bounding box changes. For the sequences of Trans and Diving in Fig. 12(c), the aspect ratio of the target keeps changing during the tracking process, while the fixed aspect ratio is limited in describing the shape change of the target. Based on the above analysis, the problems in tracking failure cases may be mitigated to some extent by extracting more efficient features and setting bounding boxes with different aspect ratios.

### 4.10 Discussion

According to the above analysis and statement of experimental results, the adaptive spatial regularization and the temporal consistent constraint contained in our proposed method improve the tracking performance of the baseline algorithm to a certain extent respectively. Firstly, different from previous tracking methods [12, 15], our spatial constraint weight is not a matrix with fixed coefficients and the shape of the target is taken into account. Therefore, our proposal can adaptively suppress background information in complex tracking scenarios. The experimental data and graphs show that the constraint scheme is effective. Furthermore, the temporal consistent constraint is constructed

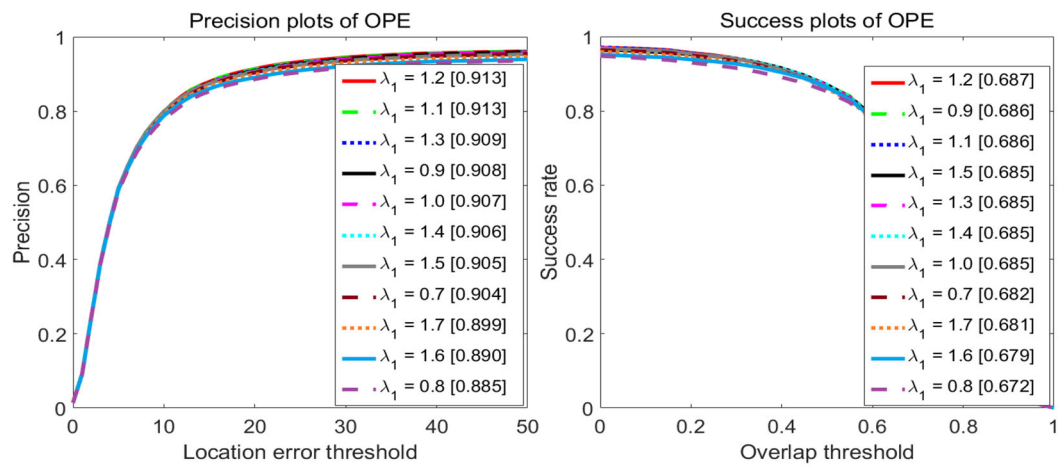
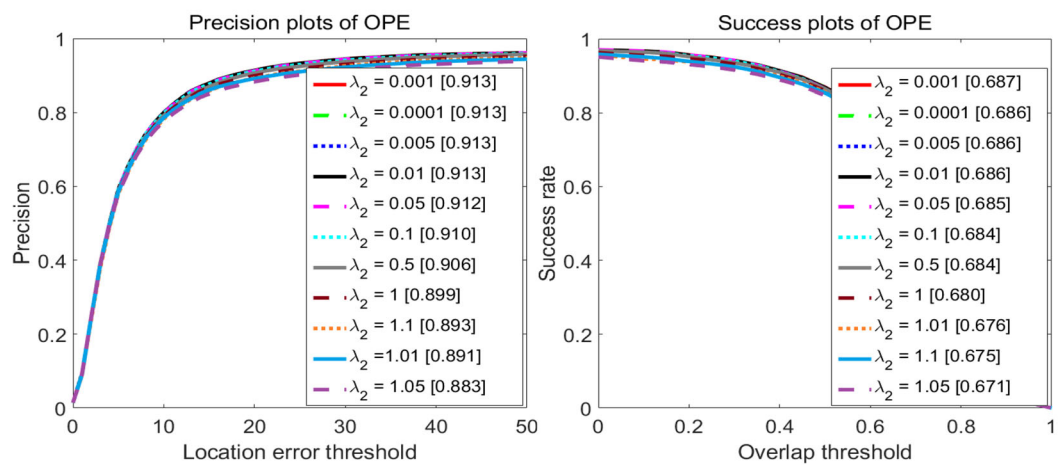
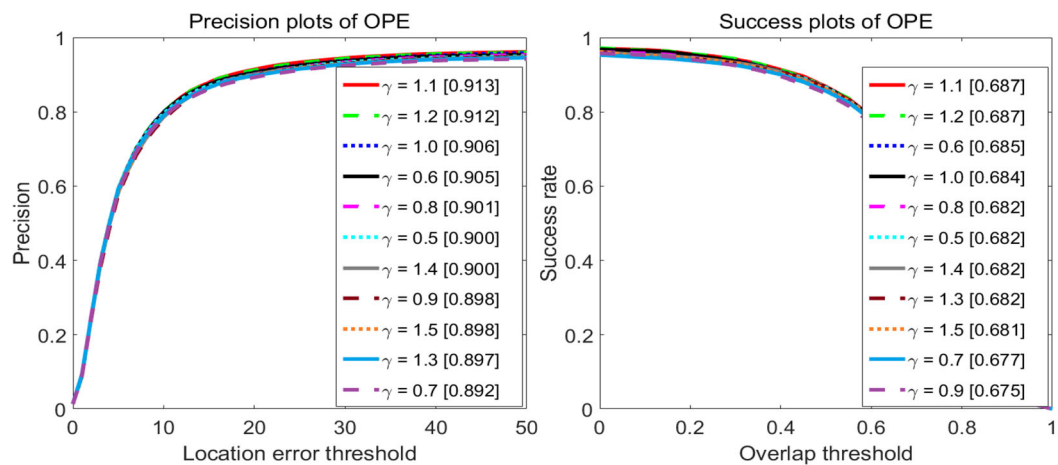
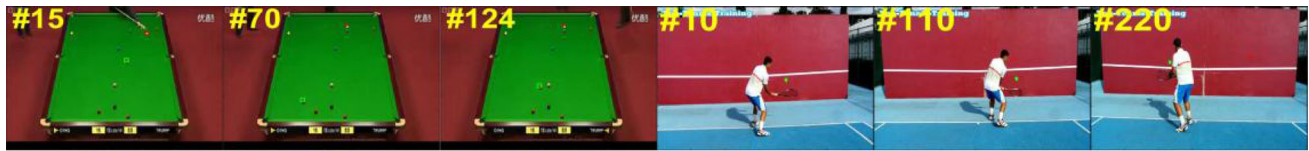
(a) Comparison of the fine-tuning spatial regularization parameter  $\lambda_1$ (b) Comparison of the fine-tuning spatial regularization parameter  $\lambda_2$ (c) Comparison of the fine-tuning temporal consistency parameter  $\gamma$ 

Fig. 11 Precision and success plots of different parameters



(a) Low resolution



(b) Fast motion and blur



(c) Scale variation

—
—  
 TSCF                                      Ground-truth

Fig. 12 Failure cases of our proposed method

on the basis of a high-confidence strategy. Considering of the appearance consistency of the target in temporal dimension, we extract reliable appearance information of the target from historical frames, which breaks the limitation of hard matching between two consecutive frames in [15]. The consistent constraint method can effectively prevent the introduction of inaccurate historical information. From the experimental results, both the accuracy and the robustness of the tracker achieve favorable improvement.

In the experiment, we observe that the algorithm proposed in this paper has some limitations. From the analysis of the failure cases, we can find that our method needs further research and improvement when dealing with the challenges of low resolution, fast motion and aspect ratio variation. Among them, the failure of the former two is mainly due to the difficulty of building a robust target model with general features. For the problem of scale variation, an inaccurate tracking box is inevitably obtained when the premise of scale estimation is a fixed aspect ratio coefficient. In the future, we plan to explore how to extract more effective features to represent the target, and further design a scale search mechanism with different aspect ratios while ensuring efficient tracking performance.

## 5 Conclusion

A joint correlation tracker consisting of temporal consistent constraint and adaptive spatial regularization is proposed in this paper. We learn a spatial weight adaptively by introducing the background map as a reference constraint into the correlation filter, which leads to favorable discrimination against cluttered background. Moreover, we extract reliable historical filters to construct the temporal consistent regularization according to the high-confidence mechanism, which enriches the diversity of the target appearance. With the unified tracking-detection framework, our approach meets robust and efficient visual tracking in extremely challenging situations. Additionally, the experimental results of qualitative and quantitative evaluations on various benchmarks verified that the proposed tracker achieves favorable tracking performance in terms of both robustness and accuracy, thereby outperforms several state-of-the-art trackers.

**Acknowledgements** This work is supported in part by National Natural Science Foundation of China (Grant No. 61972307), the Foundation of Preliminary Research Field of China (Grant No. 61405170206) and the 13th Five-Year Equipment Development Project of China (Grant No. 41412010202).

## Compliance with ethical standards

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

## References

1. Yilmaz A, Javed O, Shah M (2006) Object tracking: a survey. *ACM Comput Surv* 38(4):1–45
2. Ross DA, Lim J, Lin RS, Yang MH (2008) Incremental learning for robust visual tracking. *Int J Comput Vis* 77(1–3):125–141
3. Laura SL, Erik LM (2012) Distribution fields for tracking. In: *IEEE conference on computer vision and pattern recognition*
4. Kwon J, Lee KM (2010) Visual tracking decomposition. In: *IEEE conference on computer vision and pattern recognition*, pp 1269–1276
5. Vojir T, Noskova J, Matas J (2013) Robust scale-adaptive mean-shift for tracking. *Pattern Recogn Lett* 47:652–663
6. Hare S, Golodetz S, Saffari A, Vineet V, Cheng MM, Hicks SL, Torr PHS (2016) Struck: structured output tracking with kernels. *IEEE Trans Pattern Anal Mach Intell* 38(10):2096–2109
7. Kalal Z, Mikolajczyk K, Matas J (2012) Tracking-learning-detection. *IEEE Trans Pattern Anal Mach Intell* 34(7):1409–1422
8. Bolme DS, Beveridge JR, Draper BA, Lui YM (2010) Visual object tracking using adaptive correlation filters. In: *IEEE conference on computer vision and pattern recognition*
9. Henriques JF, Caseiro R, Martins P, Batista J (2012) Exploiting the circulant structure of tracking-by-detection with kernels. In: *European conference on computer vision*, pp 702–715
10. 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
11. Danelljan M, Khan FS, Felsberg M, Weijer J (2014) Adaptive color attributes for real-time visual tracking. In: *IEEE conference on computer vision and pattern recognition*, pp 1090–1097
12. Danelljan M, Häger G, Khan FS, Felsberg M (2016) Learning spatially regularized correlation filters for visual tracking. In: *IEEE international conference on computer vision*, pp 4310–4318
13. Galoogahi HK, Fagg A, Lucey S (2017) Learning background-aware correlation filters for visual tracking. In: *IEEE International conference on computer vision*, pp 1144–1152
14. Mueller M, Smith N, Ghanem B (2017) Context-aware correlation filter tracking. In: *IEEE conference on computer vision and pattern recognition*, pp 1387–1395
15. Li F, Tian C, Zuo WM, Zhang L, Yang MH (2018) Learning spatial-temporal regularized correlation filters for visual tracking. In: *IEEE conference on computer vision and pattern recognition*, pp 4904–4913
16. Boyd S, Parikh N, Chu E, Peleato B, Eckstein J (2011) Distributed optimization and statistical learning via the alternating direction method of multipliers. *Found Trends Mach Learn* 3(1):1–122
17. Wu Y, Lim J, Yang MH (2013) Online object tracking: a benchmark. In: *IEEE conference on computer vision and pattern recognition*, pp 2411–2418
18. Wu Y, Lim J, Yang MH (2015) Object tracking benchmark. *IEEE Trans Pattern Anal Mach Intell* 37(9):1834–1848
19. Liang PP, Blasch E, Ling HB (2015) Encoding color information for visual tracking: algorithms and benchmark. *IEEE Trans Image Process* 24(12):5630–5644
20. Kristan M, Leonardis A, Matas J, Felsberg M, Pflugfelder RP, Cehovin L, Vojir T, Häger G, Lukežič A, Fern G et al (2016) The visual object tracking VOT2016 challenge results. In: *European conference on computer vision*
21. Danelljan M, Häger G, Khan FS, Felsberg M (2014) Accurate scale estimation for robust visual tracking. In: *British machine vision conference*, pp 1–11
22. Li Y, Zhu JK (2014) A scale adaptive kernel correlation filter tracker with feature integration. In: *European conference on computer vision*, pp 254–265
23. Li F, Yao YJ, Li PH, Zhang D, Zuo WM, Yang MH (2017) Integrating boundary and center correlation filters for visual tracking with aspect ratio variation. In: *IEEE international conference on computer vision*
24. Bertinetto L, Valmadre J, Golodetz S, Miksik O, Torr PHS (2016) Staple: complementary learners for real-time tracking. In: *IEEE conference on computer vision and pattern recognition*, pp 1401–1409
25. Danelljan M, Robinson A, Khan FS, Felsberg M (2016) Beyond correlation filters: learning continuous convolution operators for visual tracking. In: *European conference on computer vision*, pp 472–488
26. Danelljan M, Bhat G, Khan FS, Felsberg M (2017) ECO: efficient convolution operators for tracking. In: *European conference on computer vision*
27. Ma C, Yang XK, Zhang CY, Yang MH (2015) Long-term Correlation Tracking. In: *IEEE conference on computer vision and pattern recognition*, pp 5388–5396
28. Wang MM, Liu Y, Huang ZY (2017) Large margin object tracking with circulant feature maps. In: *IEEE conference on computer vision and pattern recognition*, pp 4800–4808
29. Valmadre J, Bertinetto L, Henriques J, Vedaldi A, Torr PHS (2017) End-to-end representation learning for correlation filter based tracking. In: *IEEE conference on computer vision and pattern recognition*, pp 5000–5008
30. Wang Q, Gao J, Xing JL, Zhang MD, Hu WM (2017) DCFNet: discriminant correlation filters network for visual tracking. In: *IEEE conference on computer vision and pattern recognition*
31. Lukežič A, Vojir T, Zajc LČ, Matas J, Kristan M (2017) Discriminative correlation filter with channel and spatial reliability. In: *IEEE conference on computer vision and pattern recognition*, pp 4847–4856
32. Bibby C, Reid I (2008) Robust real-time visual tracking using pixel-wise posteriors. In: *European conference on computer vision*, pp 831–844
33. Ma C, Huang JB, Yang XK, Yang MH (2015) Hierarchical convolutional features for visual tracking. In: *IEEE international conference on computer vision*, pp 3074–3082

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