Applied Intelligence (2022) 52:7566–7581
https://doi.org/10.1007/s10489-021-02825-1

Visual tracking for UAV using adaptive spatio-temporal regularized correlation filters

Libin Xu[1] · Mingliang Gao[1] · Qilei Li[2] · Guofeng Zou[1] · Jinfeng Pan[1] · Jun Jiang[3]

Accepted: 8 September 2021 / Published online: 1 October 2021

Abstract
The advance of visual tracking has provided unmanned aerial vehicle (UAV) with the intriguing capability for various practical applications. With promising performance and efficiency, discriminative correlation filter (DCF)-based trackers have drawn significant attention and undergone remarkable progress. However, the boundary effect and filter degradation remain two intractable problems. In this work, we propose a novel Adaptive Spatio-Temporal Regularized Correlation Filters (ASTR-CF) model to address the two problems. The ASTR-CF model simultaneously optimizes the spatial and temporal regularization weights adaptively, and it is optimized by the alternating direction method of multipliers (ADMM) effectively. Extensive experiments on 4 UAV tracking benchmarks have proven the superiority of the proposed ASTR-CF compared with more than 30 state-of-the-art trackers in terms of accuracy and speed.

Keywords Visual tracking · Correlation filters · Spatio-temporal regularization · UAV

1 Introduction

Visual tracking is an established yet rapidly developed research area in computer vision. It aims to estimate the spatial trajectory of a target in image sequences given its initial state. Specially, visual tracking of unmanned aerial vehicles (UAVs) draws much attention benefiting from their inherent advantages, *e.g.*, easy deployment, high mobility, large field of vision and uniform scale [14]. Meanwhile, it has enabled many new applications in computer vision, such as visual surveillance [13, 43], aerial navigation [50, 61], and obstacle avoidance [41, 46]. Unlike the generic object tracking, UAV-based tracking is to locate a certain target from a low-altitude aerial perspective, which poses new challenges, *e.g.*, rapid changes in scale and perspective, limited pixels in the target region, and multiple similar disruptors [65].

Generally, visual tracking models can be classified into two categories, namely generative trackers and discriminative trackers. The former trackers manage to build models to represent the appearance of the target and search the most similar candidate region with minimal reconstruction error. While the later trackers treat tracking as an online classification task and train a classifier to distinguish the target from the candidate area.

Recently, discriminative correlation filter (DCF)-based trackers boost the tracking performance to a higher level [26, 32, 35]. One prominent merit of the DCF-based tracker is the efficient in the training and detection, as they can be transferred into the Fourier domain and operated by element-wise multiplication. The overall framework of DCF-based trackers is depicted in Fig. 1.

However, it is still challenging for DCF-based trackers to achieve high performance tracking for an arbitrary object in unconstrained scenarios. The main obstacles include spatial boundary effect, temporal filter degeneration, and the limited feature representation capacity [33]. Learning DCF in the frequency domain comes with a high cost, especially from circularly shifted examples around the foreground target. Consequently, it results in the unwanted boundary effect, which has an adverse impact on tracking accuracy [8]. This dilemma can be alleviated by applying additional

✉ Mingliang Gao
mlgao@sdut.edu.cn

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

[2] School of Electronic Engineering and Computer Science, Queen Mary University of London, London, E1 4NS, UK

[3] School of Computer Science, Southwest Petroleum University, Chengdu 610500, China

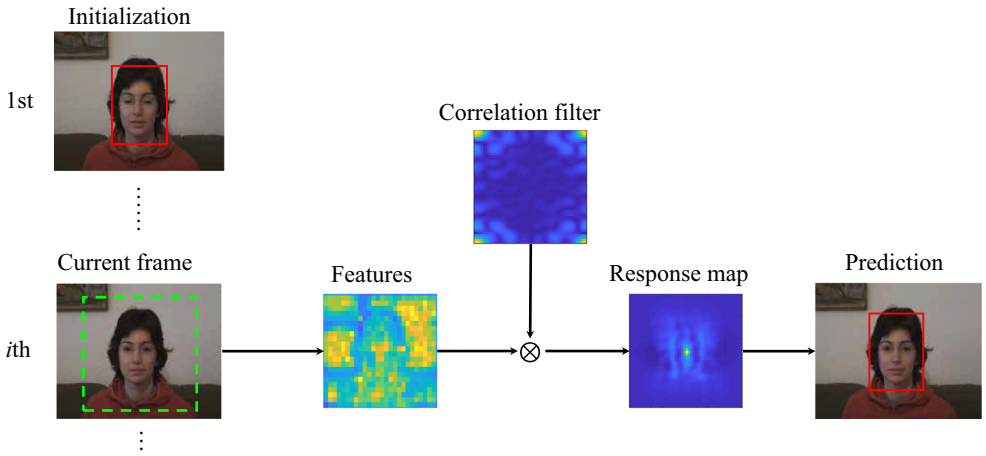

Fig. 1 Overall framework of the DCF-based trackers. Initially, the CF is trained using image patches cropped from the target in the first frame. Then, the image patch at the predicted position is cropped and features are extracted. Subsequently, a response map is calculated by the cross-correlation operation between the features and the filter. Last, the location with the maximum response is predicted as the new location of the target

predefined spatial constraints on filter coefficients. For example, Danelljan et al. [8] introduced the Spatially Regularized Discriminative Correlation Filters to mitigate the boundary effect. Galoogahi et al. [19] multiplied the filters directly with a binary matrix to generate real positive and negative samples for model training. Those aforementioned two spatial constraints are widely adopted in the subsequent works [6, 27]. On the other hand, the appearance model in DCF-based trackers is updated via a linear interpolation approach, thus it cannot adapt to ubiquitous appearance change. This results in the filter degradation inevitably. To address the problem of filter degradation, some solutions, *e.g.*, training set management [9, 11, 36], temporal restriction, tracking confidence verification [52] and over-fitting alleviation [45], are proposed. Among them, temporal regularization has been proven to be an effective way [30, 34]. Moreover, to build a robust appearance model, deep learning-based trackers have drawn much attention, *e.g.*, deep feature-based trackers [21, 22, 37] and Siamese neural network-based trackers[2, 20, 49, 59]. Although the deep learning-based trackers promote the tracking accuracy for generic visual tracking, the tracking speed of these methods is limited due to the complex calculation. Also, it is very hardware-dependent (especially on GPU), which is not in conformity with the requirements for UAV (*e.g.*, lightweight and low-power dissipation).

In this work, we proposed an adaptive spatio-temporal regularized correlation filters (ASTR-CF) model to address the issues of boundary effect and filter degradation. The merits of the ASTR-CF are summarized as follows.

1. The ASTR-CF model can effectively estimate an object-aware spatial regularization and a context-aware temporal regularization adaptively and simultaneously.
2. The ASTR-CF model can be effectively optimized via the alternating direction method of multipliers (ADMM), where each sub-problem has the analytic solution.

We perform comprehensive experiments on 4 UAV tracking benchmarks with more than 30 state-of-the-art trackers. Experimental results indicate the superiority of the proposed ASTR-CF tracker in terms of both accuracy and speed.

2 Related work

This section briefly reviews the DCF-based tracking approaches given their outstanding performance in recent tracking competitions [16, 28, 38, 57]. One of the seminal works is MOSSE [4], which formulates the tracking task as discriminative filter learning. To generate more background samples in the learning stage, the circulant matrix concept is introduced to DCF by CSK [25] with a padded search window. Additionally, spatio-temporal context [55] and kernel tricks [26] are used to improve the learning formulation.

Despite the success of DCF, it remains a challenge to achieve high performance tracking for an arbitrary object in unconstrained scenarios due to the inherent spatial boundary effect and temporal filter degradation [30]. To solve these problems, spatial regularization and temporal regularization are introduced to the DCF framework as constrains for model optimization.

2.1 Spatial regularization

Learning DCF in the frequency domain inevitably incurs the boundary effect due to the periodic assumption. To alleviate the boundary effect, Danelljan et al. [8] proposed the SRDCF tracker by introducing a spatially penalized coefficient to focus on the information near the target center. Fu et al. [17] proposed a Part-Based Background-Aware Tracking (PBBAT) method. The part-based strategy endows PBBAT the ability against boundary effect and object occlusion compared with the holistic appearance model. Huang et al. [27] proposed the Aberrance Repressed Correlation

Filters (ARCF) tracker to suppress the distortion effectively. Moreover, Fu et al. [18] proposed an object saliency-aware Dual Regularized Correlation Filters (DRCF) model by introducing a dual regularization strategy to suppress the boundary effect. Dai et al. [6] proposed an Adaptive Spatially Regularized Correlation Filters (ASRCF) model, which could estimate an object-aware spatial regularization and obtain more reliable filter coefficients during the tracking process.

2.2 Spatio-temporal regularization

In DCF-based trackers, the appearance model is updated by linear interpolation. Thus, it cannot adapt to ubiquitous appearance changes. To address this problem, Li et al. [30] proposed the Spatial and Temporal Regularization Correlation Filter (STRCF) by introducing the temporal regularization module to SRDCF and incorporating both spatial and temporal regularization into DCF. The STRCF is a rational approximation of the full SRDCF formulation on multiple training images, and it can be exploited for simultaneous DCF learning and model updating. In addition, Li et al. [34] introduced the intermittent context learning strategy to restrain filter degradation.

In this work, we propose a novel Adaptive Spatio-Temporal Regularized Correlation Filters (ASTR-CF) model based on STRCF. The proposed ASTR-CF can adaptively estimate an object-aware spatial regularization and context-aware temporal regularization. Meanwhile, it can be efficiently optimized by the ADMM algorithm [5]. The tracking framework of the proposed ASTR-CF is presented in Fig. 2.

3 Proposed method

In this section, we first revisit CF [26], SRDCF [8], and STRCF [30]. Then, the ASTR-CF model is presented. Finally, an ADMM [5] algorithm is developed to optimize the proposed model.

3.1 Objective function of ASTR-CF

CF: The standard multi-channel CF model in the spatial domain aims to minimize the following objective function [26],

$$E(\mathbf{H}) = \frac{1}{2}\left\|\mathbf{y} - \sum_{k=1}^{K}\mathbf{x}^k * \mathbf{h}^k\right\|_2^2 + \frac{1}{2}\sum_{k=1}^{K}\left\|\mathbf{h}^k\right\|_2^2, \tag{1}$$

where, $\mathbf{x}^k \in \mathbb{R}^{T\times 1}(k = 1, 2, \ldots, K)$ and $\mathbf{h}^k \in \mathbb{R}^{T\times 1}(k = 1, 2, \ldots, K)$ denote the extracted feature and the filter trained feature in the t-th frame, respectively. T denotes the length of feature. The vector $\mathbf{y} \in \mathbb{R}^{T\times 1}$ is the desired response (*i.e.*, the Gaussian-shaped ground truth) and $*$ denotes the convolution operator. $\mathbf{H} = \left[\mathbf{h}^1, \mathbf{h}^2, \ldots, \mathbf{h}^K\right]$ is the matrix representing the filter from all the K channels.

The standard CF model suffers from periodic repetitions on boundary positions caused by circulant shifted samples, which inevitably degrades the tracking performance. To solve this problem, several spatial constraints have been introduced to alleviate unexpected boundary effects. The representative methods are SRDCF [8] and STRCF [30].

SRDCF The SRDCF method [8] introduces a spatial regularization to penalize the filter coefficients with respect to their spatial locations and the objective function is formulated as,

$$E(\mathbf{H}) = \frac{1}{2}\left\|\mathbf{y} - \sum_{k=1}^{K}\mathbf{x}^k * \mathbf{h}^k\right\|_2^2 + \frac{1}{2}\sum_{k=1}^{K}\left\|\widetilde{\mathbf{w}} \odot \mathbf{h}^k\right\|_2^2, \tag{2}$$

where, $\widetilde{\mathbf{w}}$ is a negative Gaussian-shaped spatial weight to make the learned filter have a high response around the center of the tracked object. Although SRDCF suppresses the adverse boundary effects effectively, it increases the computational burden for two reasons. **(i)** The failure of exploiting circulant matrix structure. **(ii)** The large linear equations and Gauss-Seidel solver.

STRCF The STRCF method [30] adopts a spatial-temporal regularized module to CF and the objective function is formulated as,

$$\begin{aligned} E(\mathbf{H}) = {} & \frac{1}{2}\left\|\mathbf{y} - \sum_{k=1}^{K}\mathbf{x}^k * \mathbf{h}^k\right\|_2^2 + \frac{1}{2}\sum_{k=1}^{K}\left\|\widetilde{\mathbf{w}} \odot \mathbf{h}^k\right\|_2^2 \\ & + \frac{\mu}{2}\sum_{k=1}^{K}\left\|\mathbf{h}^k - \mathbf{h}_{t-1}^k\right\|_2^2, \end{aligned} \tag{3}$$

where, $\mathbf{x}^k \in \mathbb{R}^{T\times 1}(k = 1, 2, \ldots, K)$ is the extracted feature with length T in frame t. $\mathbf{h}^k, \mathbf{h}_{t-1}^k \in \mathbb{R}^{T\times 1}(k = 1, 2, 3, \ldots, K)$ denotes the filter of the k-th channel trained in the t-th and (t-1)-th frame, respectively. The spatial regularization weight $\widetilde{\mathbf{w}}$ is imitated from SRDCF [8] to reduce the boundary effect, and temporal regularization is firstly proposed to restrict filter variation by penalizing the difference between the current and previous filters.

However, as aforementioned, the spatial regularization and temporal penalty strength of STRCF [30] are fixed. Therefore, it cannot adapt well to the appearance variation in the unforeseeable aerial tracking scenarios.

The proposed ASTR-CF model To tackle these issues, we propose a novel adaptive spatio-temporal regularized method

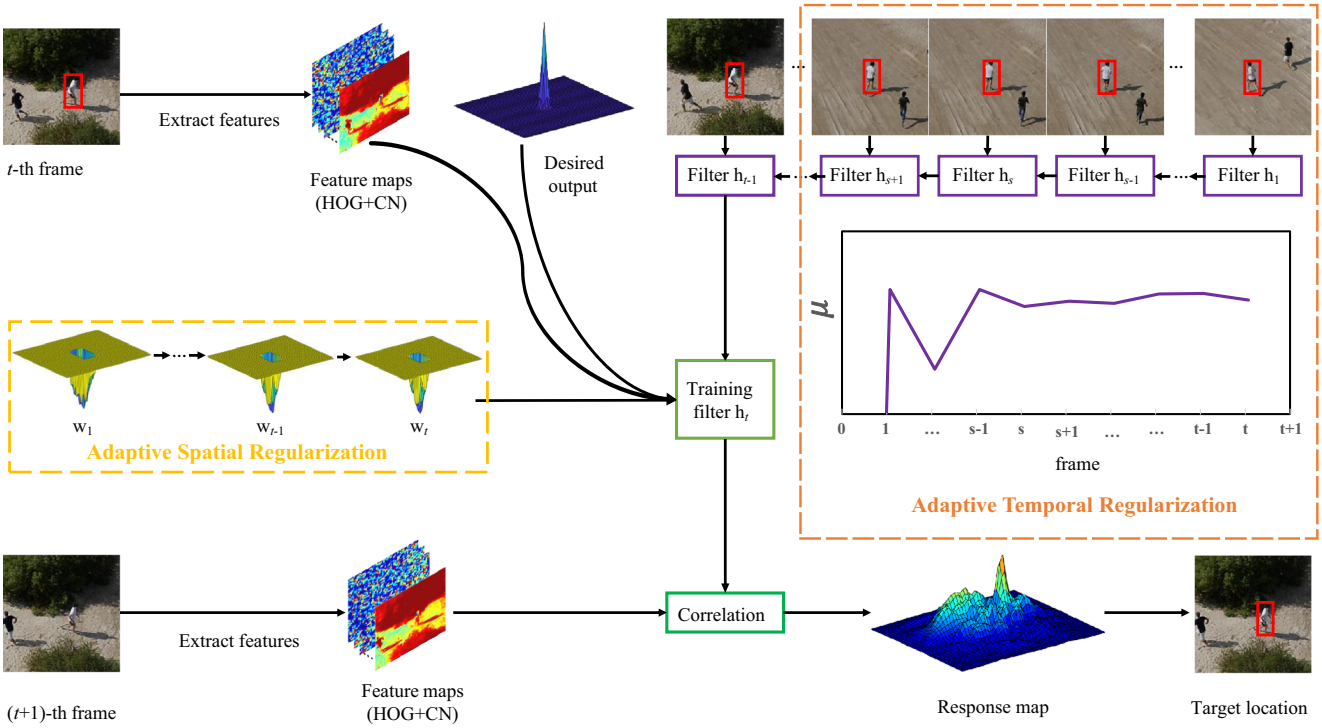

Fig. 2 Tracking framework of the proposed ASTR-CF model. In the training stage, a training patch is cropped at the estimated location of the target at the t-th frame. Here, $\mathbf{w}$ and μ are object-aware spatial regularization weight and context-aware temporal regularization parameter, respectively. $\mathbf{w}$ is flexible in different frames, and it introduces prior information to avoid model degradation. μ can be adaptively adjusted according to the response map variations. We extract the feature maps (HOG [7] and Color Names [48]) combined with prior filter $\mathbf{h}_{t-1}$, spatial regularization weight $\mathbf{w}$ and desired output $\mathbf{y}$ to train the current filter $\mathbf{h}_t$. At the $(t+1)$-th frame, the trained filter $\mathbf{h}_t$ and the feature map $\mathbf{x}_t$ of the current frame generate a response map through cross-correlation operations. Finally, the target is located based on the maximum value of the response map. More details are presented in Section 3.2

to learn multi-channel CFs effectively. The objective function of the proposed ASTR-CF model is defined as follows,

$$\begin{aligned} E(\mathbf{H},\mathbf{w},\mu) = &\frac{1}{2}\left\|\mathbf{y}-\sum_{k=1}^{K}\mathbf{x}^k * \mathbf{h}^k\right\|_2^2 \\ &+\left(\frac{\lambda_1}{2}\sum_{k=1}^{K}\left\|\mathbf{w}\odot\mathbf{h}^k\right\|_2^2+\frac{\lambda_2}{2}\|\mathbf{w}-\widetilde{\mathbf{w}}\|_2^2\right) \\ &+\left(\frac{\mu}{2}\sum_{k=1}^{K}\left\|\mathbf{h}^k-\mathbf{h}_{t-1}^k\right\|_2^2+\frac{1}{2}\|\mu-\tilde{\mu}\|_2^2\right), \end{aligned} \tag{4}$$

where, the first term is the ridge regression term that convolves the training data $\mathbf{X}=\left[\mathbf{x}^1,\mathbf{x}^2,\ldots,\mathbf{x}^K\right]$ with the filter $\mathbf{H}=\left[\mathbf{h}^1,\mathbf{h}^2,\ldots,\mathbf{h}^K\right]$ to fit the Gaussian-distributed ground truth $\mathbf{y}$. The second term introduces an adaptive spatial regularization on the filter $\mathbf{H}$. The spatial weight $\mathbf{w}$ requires to be optimized to approximate a reference weight $\widetilde{\mathbf{w}}$. This constraint introduces prior information on $\mathbf{w}$ and avoids model degradation. λ_1 and λ_2 are the regularization parameters of the second term. The third term introduces an adaptive temporal regularization, where $\tilde{\mu}$ and μ denote the reference and optimized context-aware temporal regularization parameter, respectively [33]. $\tilde{\mu}$ is denoted as,

$$\tilde{\mu}=\frac{\zeta}{1+\log\left(\nu||\Pi||_2+1\right)},\quad \|\Pi\|_2\le\phi, \tag{5}$$

where, $\Pi=\left[\left|\Pi^1\right|,\left|\Pi^2\right|,\ldots,\left|\Pi^T\right|\right]$ denotes the response variations, ϕ denotes threshold. ζ and ν denote hyperparameters.

3.2 Optimization of ASTR-CF

We formulate the objective function *i.e.*, Eq. (4), in the frequency domain using Parseval's theorem, and convert it into the equality constrained optimization form,

$$\begin{aligned} E(\mathbf{H},\widehat{\mathbf{G}},\mathbf{w},\mu) = &\frac{1}{2}\left\|\widehat{\mathbf{y}}-\sum_{k=1}^{K}\widehat{\mathbf{x}}^k\odot\widehat{\mathbf{g}}^k\right\|_2^2+\frac{\lambda_1}{2}\sum_{k=1}^{K}\left\|\mathbf{w}\odot\mathbf{h}^k\right\|_2^2 \\ &+\frac{\lambda_2}{2}\|\mathbf{w}-\widetilde{\mathbf{w}}\|_2^2+\frac{\mu}{2}\sum_{k=1}^{K}\left\|\widehat{\mathbf{g}}^k-\widehat{\mathbf{g}}_{t-1}^k\right\|_2^2 \\ &+\frac{1}{2}\|\mu-\tilde{\mu}\|_2^2, \\ s.t.,\ \widehat{\mathbf{g}}^k=&\sqrt{T}\mathbf{F}\mathbf{h}^k,\ k=1,2,\ldots,K. \end{aligned} \tag{6}$$

Where, $\widehat{\mathbf{G}} = \left[\widehat{\mathbf{g}}^1, \widehat{\mathbf{g}}^2, \ldots, \widehat{\mathbf{g}}^K\right] (\widehat{\mathbf{g}}^k = \sqrt{T}\mathbf{F}\mathbf{h}^k, k = 1, 2, \ldots, K)$ is an auxiliary variable matrix. The symbol $\widehat{\cdot}$ denotes the discrete Fourier transform form of a given signal. $\mathbf{F}$ is the orthonormal $T \times T$ matrix of complex basis vectors to map any T dimensional vectorized signal into the Fourier domain. The model in Eq. (6) is bi-convex, and it can be minimized to obtain a local optimal solution using ADMM [5]. The augmented Lagrangian form of Eq. (6) can be formulated as,

$$\begin{aligned} L(\mathbf{H}, \widehat{\mathbf{G}}, \mathbf{w}, \mu, \widehat{\mathbf{V}}) &= E(\mathbf{H}, \widehat{\mathbf{G}}, \mathbf{w}, \mu) \\ &+ \frac{\gamma}{2} \sum_{k=1}^{K} \left\| \widehat{\mathbf{g}}^k - \sqrt{T}\mathbf{F}\mathbf{h}^k \right\|_2^2 \\ &+ \sum_{k=1}^{K} \left(\widehat{\mathbf{v}}^k\right)^{\mathrm{T}} \left(\widehat{\mathbf{g}}^k - \sqrt{T}\mathbf{F}\mathbf{h}^k\right), \end{aligned} \tag{7}$$

where, γ denotes the step size regularization parameter, $\mathbf{V}$ is the Lagrange multiplier, and $\widehat{\mathbf{V}}$ is the corresponding Fourier transform. By introducing $\mathbf{s}^k = \frac{1}{\gamma}\mathbf{v}^k$, the optimization of Eq. (7) is equivalent to solving,

$$\begin{aligned} L(\mathbf{H}, \widehat{\mathbf{G}}, \mathbf{w}, \mu, \widehat{\mathbf{S}}) &= E(\mathbf{H}, \widehat{\mathbf{G}}, \mathbf{w}, \mu) \\ &+ \frac{\gamma}{2} \sum_{k=1}^{K} \left\| \widehat{\mathbf{g}}^k - \sqrt{T}\mathbf{F}\mathbf{h}^k + \widehat{\mathbf{s}}^k \right\|_2^2. \end{aligned} \tag{8}$$

Then, the ADMM [5] algorithm is applied by alternately solving the following 5 subproblems.

Subproblem $\widehat{\mathbf{G}}$ If $\mathbf{H}$, $\mathbf{w}$, μ, and $\widehat{\mathbf{S}}$ are given, the optimal $\widehat{\mathbf{G}}^*$ can be estimated by solving the optimization problem as,

$$\begin{aligned} \widehat{\mathbf{G}}^* = \underset{\widehat{\mathbf{G}}}{\operatorname{argmin}} \Bigg\{ &\frac{1}{2} \left\| \mathbf{y} - \sum_{k=1}^{K} \widehat{\mathbf{x}}^k \odot \widehat{\mathbf{g}}^k \right\|_2^2 + \frac{\mu}{2} \sum_{k=1}^{K} \left\| \widehat{\mathbf{g}}^k - \widehat{\mathbf{g}}_{t-1}^k \right\|_2^2 \\ &+ \frac{\gamma}{2} \sum_{k=1}^{K} \left\| \widehat{\mathbf{g}}^k - \sqrt{T}\mathbf{F}\mathbf{h}^k + \widehat{\mathbf{s}}^k \right\|_2^2 \Bigg\}. \end{aligned} \tag{9}$$

However, it is difficult to optimize Eq. (9) due to its high computation complexity. Thus, it can be simplified via processing each pixel of all channels by,

$$\begin{aligned} \mathcal{V}_j^*(\widehat{\mathbf{G}}) = \underset{\mathcal{V}_j(\widehat{\mathbf{G}})}{\operatorname{argmin}} \Bigg\{ &\frac{1}{2} \left\| \widehat{\mathbf{y}}_j - \mathcal{V}_j(\widehat{\mathbf{X}})^{\mathrm{T}} \mathcal{V}_j(\widehat{\mathbf{G}}) \right\|_2^2 \\ &+ \frac{\mu}{2} \left\| \mathcal{V}_j(\widehat{\mathbf{G}}) - \mathcal{V}_j\left(\widehat{\mathbf{G}}_{t-1}\right) \right\|_2^2 \\ &+ \frac{\gamma}{2} \left\| \mathcal{V}_j(\widehat{\mathbf{G}}) + \mathcal{V}_j(\widehat{\mathbf{S}}) - \mathcal{V}_j(\sqrt{T}\mathbf{F}\mathbf{H}) \right\|_2^2 \Bigg\}, \end{aligned} \tag{10}$$

where, $\mathcal{V}_j(\widehat{\mathbf{X}}) \in \mathbb{C}^{K \times 1}$ denotes the values on the pixel j $(j = 1, 2, \ldots, T)$ in all K channels of $\widehat{\mathbf{X}}$. Then, the analytical solution of Eq. (10) can be obtained as,

$$\mathcal{V}^*(\widehat{\mathbf{G}}) = \frac{1}{\mu + \gamma} \left[\mathbf{I} - \frac{\mathcal{V}_j(\widehat{\mathbf{X}}) \mathcal{V}_j(\widehat{\mathbf{X}})^{\mathrm{T}}}{\mu + \gamma + \mathcal{V}_j(\widehat{\mathbf{X}})^{\mathrm{T}} \mathcal{V}_j(\widehat{\mathbf{X}})} \right] \rho, \tag{11}$$

here,

$$\rho = \mathcal{V}_j(\widehat{\mathbf{X}}) \widehat{\mathbf{y}}_j + \mu \left[\mathcal{V}_j\left(\widehat{\mathbf{G}}_{t-1}\right)\right] + \gamma \left[\mathcal{V}_j(\sqrt{T}\mathbf{F}\mathbf{H}) - \mathcal{V}_j(\widehat{\mathbf{S}})\right]. \tag{12}$$

The derivation of Eq. (12) uses the Sherman Morrsion formula,

$$\left(\mathbf{A} + \mathbf{u}\mathbf{v}^{\mathrm{T}}\right)^{-1} = \mathbf{A}^{-1} - \frac{\mathbf{A}^{-1}\mathbf{u}\mathbf{v}^{\mathrm{T}}\mathbf{A}^{-1}}{1 + \mathbf{v}^{\mathrm{T}}\mathbf{A}^{-1}\mathbf{u}}, \tag{13}$$

where, $\mathbf{u}$ and $\mathbf{v}$ are two column vectors and $\mathbf{u}\mathbf{v}^{\mathrm{T}}$ is a rank-one matrix.

Subproblem H If $\widehat{\mathbf{G}}$, $\mathbf{w}$, μ, and $\widehat{\mathbf{S}}$ are given, the optimal $\mathbf{H}^*$ can be obtained as,

$$\begin{aligned} \mathbf{h}^{k^*} &= \underset{\mathbf{h}^k}{\operatorname{argmin}} \left\{ \frac{\lambda_1}{2} \left\| \mathbf{w} \odot \mathbf{h}^k \right\|_2^2 + \frac{\gamma}{2} \left\| \widehat{\mathbf{g}}^k - \sqrt{T}\mathbf{F}\mathbf{h}^k + \widehat{\mathbf{s}}^k \right\|_2^2 \right\} \\ &= \left[\lambda_1 \mathbf{W}^{\mathrm{T}}\mathbf{W} + \gamma T \mathbf{I}\right]^{-1} \gamma T \left(\mathbf{g}^k + \mathbf{s}^k\right) \\ &= \frac{\gamma T \left(\mathbf{g}^k + \mathbf{s}^k\right)}{\lambda_1 (\mathbf{w} \odot \mathbf{w}) + \gamma T}, \end{aligned} \tag{14}$$

where, $\mathbf{W} = \operatorname{diag}(\mathbf{w}) \in \mathbb{R}^{T \times T}$. Eq. (14) shows that the solution of $\mathbf{h}^k$ merely requires the element-wise multiplication and the inverse fast Fourier transform (*i.e.*,$\mathbf{g}^k = \frac{1}{\sqrt{T}}\mathbf{F}^{\mathrm{T}}\widehat{\mathbf{g}}^k$ and $\mathbf{s}^k = \frac{1}{\sqrt{T}}\mathbf{F}^{\mathrm{T}}\widehat{\mathbf{s}}^k$).

Solving w: If $\mathbf{H}$, $\widehat{\mathbf{G}}$, μ and $\widehat{\mathbf{S}}$ are fixed, the closed-form solution regrading $\mathbf{w}$ can be determined as,

$$\begin{aligned} \mathbf{w}^* &= \underset{\mathbf{w}}{\operatorname{argmin}} \left\{ \frac{\lambda_1}{2} \sum_{k=1}^{K} \left\| \mathbf{w} \odot \mathbf{h}^k \right\|_2^2 + \frac{\lambda_2}{2} \|\mathbf{w} - \widetilde{\mathbf{w}}\|_2^2 \right\} \\ &= \left[\lambda_1 \sum_{k=1}^{K} \left(\mathbf{N}^k\right)^{\mathrm{T}} \mathbf{N}^k + \lambda_2 \mathbf{I} \right]^{-1} \lambda_2 \widetilde{\mathbf{w}} \\ &= \frac{\lambda_2 \widetilde{\mathbf{w}}}{\lambda_1 \sum_{k=1}^{K} \mathbf{h}^k \odot \mathbf{h}^k + \lambda_2 \mathbf{I}}, \end{aligned} \tag{15}$$

where, $\mathbf{N}^k = \operatorname{diag}\left(\mathbf{h}^k\right) \in \mathbb{R}^{T \times T}$.

The visualization of the learned weights $\mathbf{w}$ along with the change of target's appearance is shown in Fig. 3. It shows that the adaptive spatial regularization works well in introducing large penalties on the unreliable regions and encourages the learned filter to focus more on the reliable regions in the next iteration. Thus, the ASTR-CF can obtain more reliable filter coefficients during the tracking process.

Solving μ Given other variables $\mathbf{H}$, $\widehat{\mathbf{G}}$, $\mathbf{w}$, and $\widehat{\mathbf{S}}$, the optimal solution of μ can be determined as,

$$\begin{aligned} \mu^* &= \underset{\mu}{\operatorname{argmin}} \left\{ \frac{\mu}{2} \sum_{k=1}^{K} \left\| \widehat{\mathbf{g}}^k - \widehat{\mathbf{g}}_{t-1}^k \right\|_2^2 + \frac{1}{2} \|\mu - \tilde{\mu}\|_2^2 \right\} \\ &= \tilde{\mu} - \frac{1}{2} \sum_{k=1}^{K} \left\| \widehat{\mathbf{g}}^k - \widehat{\mathbf{g}}_{t-1}^k \right\|_2^2. \end{aligned} \tag{16}$$

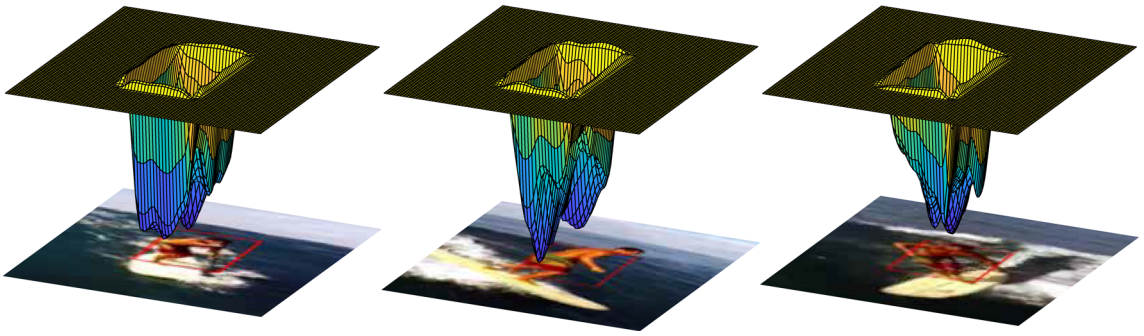

Fig. 3 Visualization of the proposed adaptive spatial regularization weight $\mathbf{w}$

Specifically, adaptive temporal regularization is introduced to restrict filter variation by penalizing the difference between the current and previous filter adaptively. Thus, ASTR-CF can gain a more robust appearance model than SRDCF and STRCF when the target is suffered from significant appearance variations. The visualization of the adaptive temporal regularization parameter μ along with the variations of response is shown in Fig. 4.

Lagrangian Multiplier Update We update Lagrangian multipliers as,

$$\widehat{\mathbf{S}}^{i+1} = \widehat{\mathbf{S}}^{i} + \widehat{\mathbf{G}}^{i+1} - \widehat{\mathbf{H}}^{i+1}, \tag{17}$$

where, i and $i+1$ denote the iteration index.

By solving the aforementioned five subproblems iteratively, we can optimize the objective function effectively and obtain the optimal filter $\widehat{\mathbf{G}}$, object-aware spatial regularization weight $\mathbf{w}$ and context-aware temporal regularization parameter μ in frame t. Then, $\widehat{\mathbf{G}}$ is used for detection in frame $t+1$. The pseudocode of the filter training is summarized in Algorithm 1.

Algorithm 1 Training algorithm for ASTR-CF.

Input: The feature map $\mathbf{X}_t$ in the t-th frame, the Gaussian-shaped response $\mathbf{y}$, the previous filter $\mathbf{H}_{t-1}$, and the ADMM iteration N.

Output: Optimized filter.

```
for t = 1 : video frames do
    Update temporal regularized reference parameter μ̃_t by (5).
    Update spatial regularized reference weight w̃_t = w_{t−1}.
    for i=1 : N do
        Calculate Ĝ* by (11) and (12).
        Calculate H* by (14).
        Update w* by (15).
        Update μ* by (16).
        Update the Lagrangian multipliers Ŝ by (17).
    end
    Obtain the optimized filter Ĝ = Ĝ*^N for detection of the next frame.
end
```

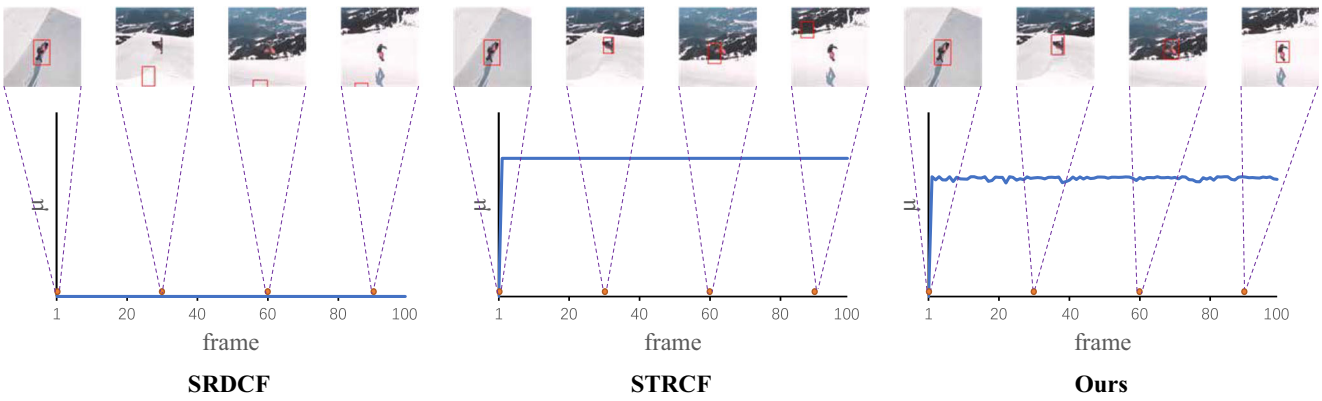

Fig. 4 Visualization of the proposed adaptive temporal regularization parameter μ

3.3 Target localization

The location of the target can be determined in the Fourier domain as,

$$\widehat{\mathcal{R}}_t = \sum_{k=1}^{K} \widehat{\mathbf{x}}^k \odot \widehat{\mathbf{g}}_{t-1}^k, \tag{18}$$

where, $\mathcal{R}_t$ and $\widehat{\mathcal{R}}_t$ denote the response map and its Fourier transform. After obtaining the response map, the optimal location can be obtained based on the maximum response.

4 Experiments and discussions

4.1 Evaluation metrics

We conduct quantitative and qualitative experiments on 4 popular UAV tracking benchmarks, *i.e.*, UAV123@10fps [38], DTB70[31], VisDrone2018-SOT-test-dev [56], and UAVDT [14]. Two evaluation metrics, namely success rate and precision [57], are adopted. These two evaluation metrics have been widely adopted to evaluate the tracking accuracy in the visual tracking domain. The success rate denotes the percentage of frames in which the Intersection Over Union (IOU) exceeds a threshold. The precision denotes the percentage of frames whose estimated location falls in the given threshold distance from the ground truth. In this work, performance evaluation on these benchmarks is based on the One Pass Evaluation (OPE) rule [57].

Given the tracked bounding box r_t and the ground truth bounding box r_g, the overlap score is defined as,

$$S = \frac{|r_t \bigcap r_g|}{|r_t \bigcup r_g|}. \tag{19}$$

Where, $\bigcap$ and $\bigcup$ represent the intersection and union of two regions, respectively. $|\cdot|$ denotes the number of pixels in the region. The tracking accuracy is evaluated by success rate and precision, and these two evaluation indices are measured by overlap score S and center location error (CLE), respectively. Area Under Curve (AUC) and Distance Precision (denoted by percentage of frames whose CLE $\leq$ 20 pixels) are adopted to rank the success rate and precision of different trackers. Moreover, the tracking speed is measured by Frames Per Second (FPS).

4.2 Experimental setup

The performance evaluation are implemented using MATLAB R2017a on a PC with an i7-8700K processor (3.7GHz), 32GB RAM and NVIDIA GTX 1080Ti GPU. For the parameters of ASTR-CF tracker, we set $\lambda_1 = 1$, $\lambda_2 = 0.001$, $\nu = 2 \times 10^{-5}$, and $\zeta = 13$. The threshold of ϕ is 3000, and the ADMM iteration is set to 4. The scheme for selecting γ (initially set to 1) is $\gamma^{i+1} = \min\left(\gamma_{\max},\ \beta\gamma^i\right)$ ($\beta = 10$, $\gamma_{\max} = 10000$). To make a fair comparison, the compared results are based on the codes or results which are publicly available.

4.3 Quantitative evaluation

4.3.1 Evaluation on UAV123@10fps benchmark

The UAV123@10fps benchmark [38] is down sampled from the UAV123 benchmark, which contains 123 UAV sequences, among which 115 sequences are captured by a UAV camera and 8 sequences are rendered by a UAV simulator. The UAV123@10fps benchmark provides a comprehensive sampling of tracking nuisances that are ubiquitous in low-altitude UAV videos. To the best of our knowledge, it is the first benchmark to address and analyze the performance of the state-of-the-art trackers on a comprehensive set of annotated aerial sequences that exhibit specific tracking nuisances.

Fifteen state-of-the-art trackers are employed for comparison, including *i.e.*, AutoTrack [33], STRCF [30], BACF [19], ECO-HC [11], RaF [62], DCF-CA [39], MOSSE-CA [39], Staple-CA [39], SAMF-CA [39], SAMF-AT [3], Staple [1], SRDCF [8], ARCF-H [27], LADCF-HC [58] and fDSST [12]. The comparative results are depicted in Fig. 5. It shows that the ASTR-CF tracker ranks first in terms of both success rate and precision. It surpasses the baseline STRCF [30] by 2.2% and 5.1% in terms of success rate and precision, respectively.

To further analyze the effectiveness of the proposed ASTR-CF model, we evaluate it on UAV123@10fps benchmark [38] with different attributes. UAV123@10fps [38] has defined 12 challenging attributes, including Aspect Ratio Change (ARC), Background Clutter (BC), Camera Motion (CM), Fast Motion (FM), Full Occlusion (FOC), Illumination Variation (IV), Low Resolution (LR), Out-of-View (OV), Partial Occlusion (POC), Similar Object (SOB), Scale Variation (SV) and Viewpoint Change (VC). Evaluation of different trackers with 12 challenging attributes on UAV123@10fps benchmark in terms of precision is listed in Table 1. One can see that the proposed tracker performs well in all the challenging situations. Particularly, it has a significant improvement on 6 attributes, *e.g.,* FM, IV, LR, POC, SV and VC. The success rate and precision of these attributes are shown in Fig. 6. Compared with the baseline STRCF [30], the success rate and precision for ASTR-CF notably improved by 3.9% and 9.1% in the attribute of IV. This phenomenon can be attributed to the fact that the learned filters can alleviate unexpected noises within the object region by introducing the adaptive spatial regularization.

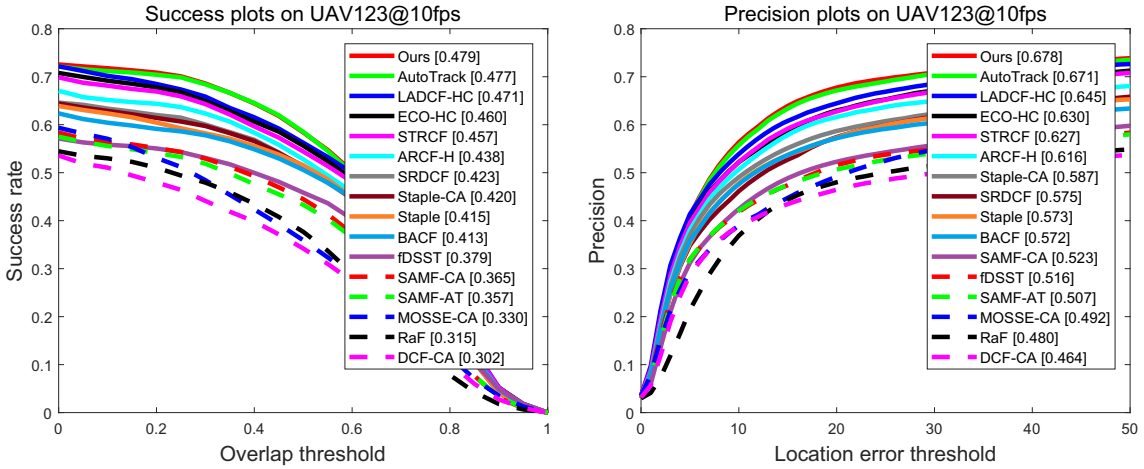

Fig. 5 Performance evaluation on the UAV123@10fps in terms of success rate and precision

4.3.2 Evaluation on DTB70 benchmark

The DTB70 benchmark [31] contains 70 challenging UAV image sequences in which the targets are surfed from large-scale changes and aspect ratios in various cluttered scenes. We compare the ASTR-CF tracker with 15 state-of-the-art trackers, *i.e.*, AutoTrack [33], STRCF [30], BACF [19], ECO-HC [11], RaF [62], DCF-CA [39], MOSSE-CA [39], Staple-CA [39], SAMF-CA [39], SAMF-AT [3], Staple [1], SRDCF [8], ARCF-H [27], LADCF-HC [58] and fDSST [12]. Fig. 7 depicts the performance evaluation of the trackers in terms of success rate and precision. Overall, the proposed ASTR-CF outperforms most of the competing trackers in terms of precision, with slightly lower (0.2%) than AutoTrack [33]. It is noteworthy that the proposed tracker surpasses its counterpart SRDCF [8] and STRCF [30] by 22% and 6.6%, respectively. For the success rate, the proposed tracker achieves the best result among all the other trackers. Compared with AutoTrack [33], ASTR-CF model brings in a reference weight $\widetilde{\mathbf{w}}$ which introduces prior information on $\mathbf{w}$ (*i.e.*, weight of spatial regularization) to avoid model degradation.

Table 1 Evaluation of different trackers with 12 challenging attributes on UAV123@10fps in terms of precision. The top-3 results are shown in red, green, and blue fonts, respectively. The proposed tracker outperforms the conventional CF based methods and other sophisticate models in 9 situations

	ARC	BC	CM	FM	FOC	IV	LR	OV	POC	SOB	SV	VC	Average
DCF-CA	0.335	0.318	0.366	0.231	0.307	0.312	0.360	0.313	0.367	0.470	0.411	0.365	0.464
RaF	0.398	0.280	0.463	0.386	0.315	0.373	0.271	0.442	0.377	0.476	0.437	0.379	0.480
MOSSE-CA	0.376	0.351	0.411	0.255	0.324	0.349	0.400	0.375	0.396	0.507	0.446	0.375	0.492
SAMF-AT	0.412	0.328	0.425	0.322	0.400	0.387	0.395	0.421	0.438	0.537	0.473	0.390	0.507
fDSST	0.418	0.319	0.432	0.350	0.379	0.395	0.399	0.447	0.451	0.533	0.471	0.400	0.587
SAMF-CA	0.408	0.328	0.466	0.391	0.359	0.369	0.365	0.429	0.419	0.536	0.474	0.404	0.523
BACF	0.478	0.425	0.532	0.407	0.336	0.430	0.431	0.421	0.467	0.605	0.525	0.491	0.572
Staple	0.459	0.409	0.499	0.356	0.388	0.438	0.408	0.441	0.507	0.612	0.519	0.485	0.573
SRDCF	0.472	0.389	0.527	0.427	0.418	0.436	0.431	0.492	0.504	0.585	0.531	0.474	0.575
Staple-CA	0.480	0.446	0.511	0.319	0.408	0.469	0.448	0.466	0.518	0.637	0.534	0.496	0.587
ARCF-H	0.530	0.428	0.558	0.402	0.392	0.483	0.486	0.449	0.533	0.667	0.575	0.516	0.587
STRCF	0.524	0.477	0.602	0.488	0.426	0.493	0.509	0.523	0.559	0.630	0.580	0.537	0.627
ECO-HC	0.549	0.540	0.606	0.447	0.458	0.512	0.531	0.510	0.558	0.654	0.582	0.548	0.630
LADCF-HC	0.561	0.443	0.637	0.503	0.432	0.481	0.550	0.537	0.564	0.653	0.608	0.549	0.645
AutoTrack	0.598	0.502	0.647	0.525	0.444	0.550	0.532	0.554	0.584	0.664	0.629	0.588	0.671
Ours	0.603	0.507	0.653	0.554	0.446	0.584	0.536	0.533	0.598	0.679	0.636	0.607	0.678

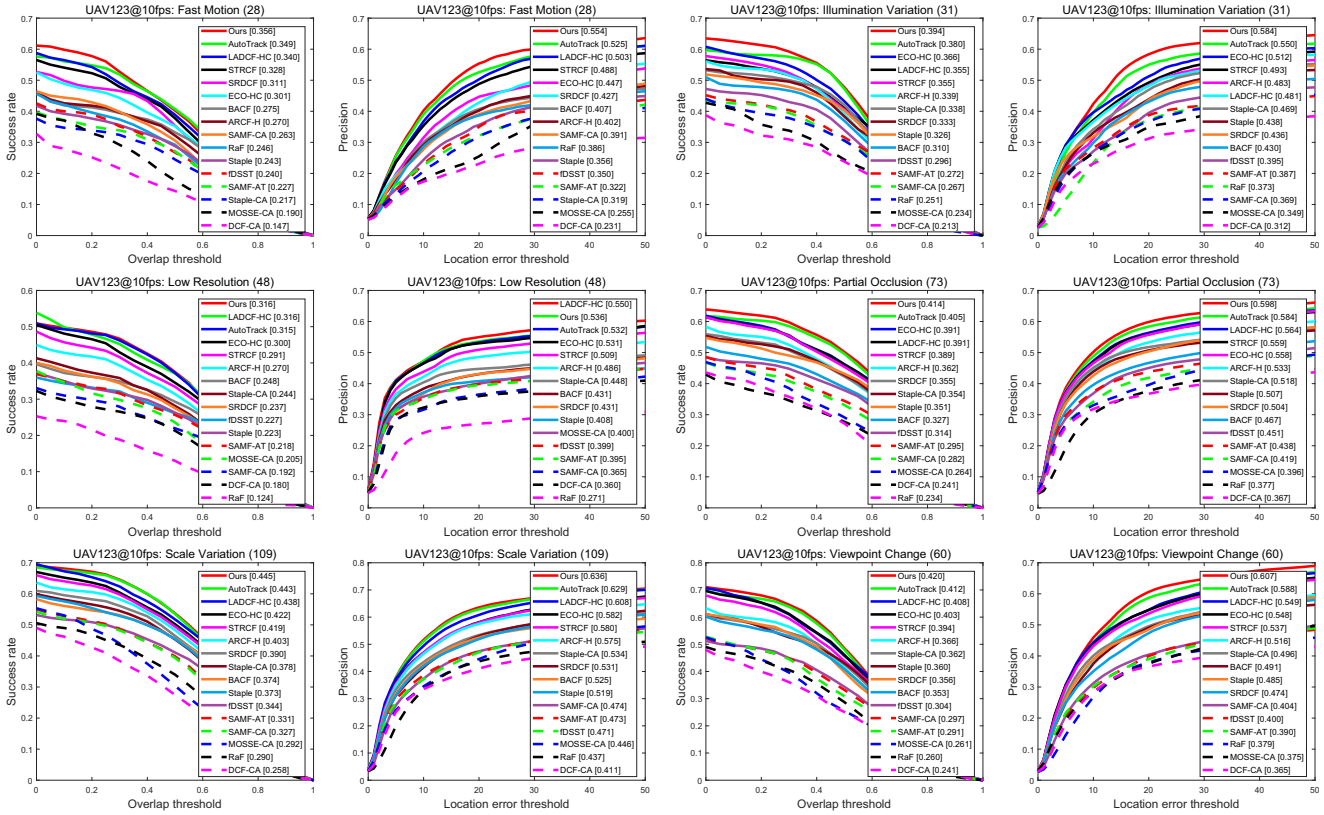

Fig. 6 Evaluation of different trackers with 6 challenging attributes on UAV123@10fps in terms of the success rate and precision. The title of each plot indicates the number of videos labelled with the respective attribute. The proposed ASTR-CF performs the best in these challenging situations

4.3.3 Evaluation on VisDrone2018-SOT-test-dev benchmark

The VisDrone2018-SOT-test-dev [56] benchmark includes 35 sequences with 29, 367 frames, and provides fully annotated bounding boxes of the targets as well as several useful attributes, *e.g.*, occlusion, background clutter, and camera motion. The targets in these sequences include pedestrians, cars, buses, and animals. The ASTR-CF tracker is compared with 15 state-of-the-art trackers, *i.e.*, AutoTrack [33], STRCF [30], BACF [19], ECO-HC [11], RaF [62], DCF-CA [39], MOSSE-CA [39], Staple-CA [39], SAMF-CA [39], SAMF-AT [3], Staple [1], SRDCF [8], ARCF-H [27], LADCF-HC [58] and fDSST [12]. As shown in Fig. 8, the ASTR-CF tracker outperforms most of the state-of-the-art trackers and achieves comparable performance with ECO-HC [11] both in terms of success

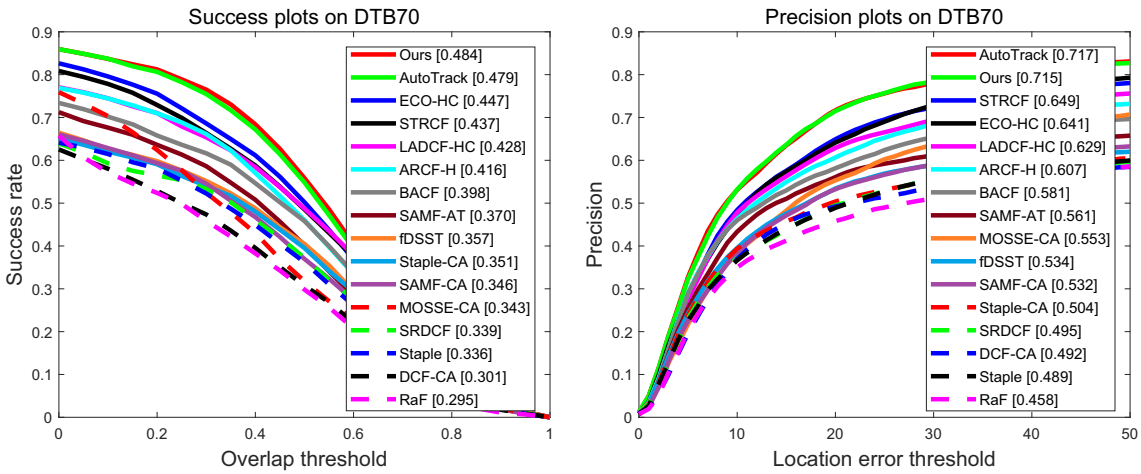

Fig. 7 Performance evaluation on DTB70 benchmark in terms of success rate and precision

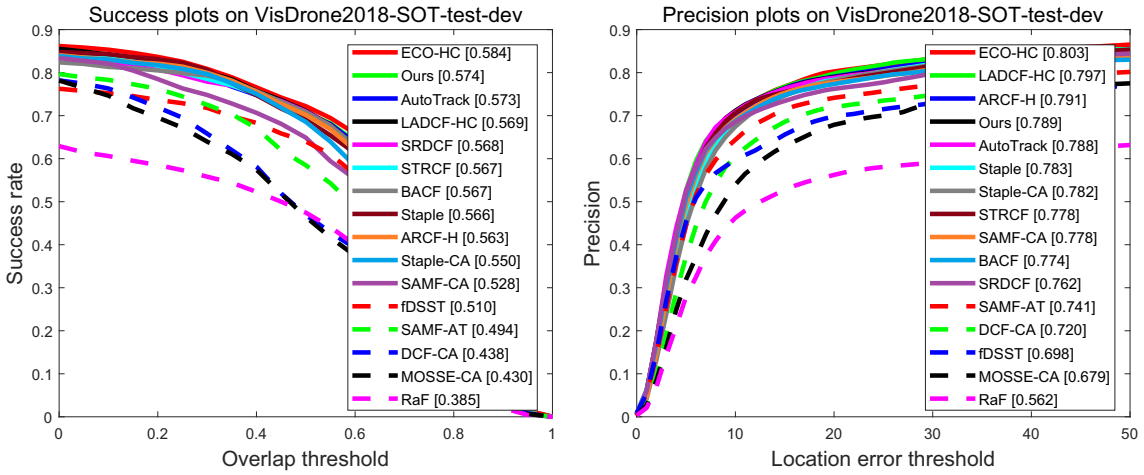

Fig. 8 Evaluation of different trackers on VisDrone2018-SOT-test-dev benchmark in terms of success rate and precision

rate and precision. This can be contributed to the stabilizing learning of the filter, especially in scenarios where a new sample is affected by sudden changes, such as out-of-view, and ratio change. The proposed tracker has advantages over ECO-HC in the situations of illumination variation and fast motion. Evaluation details of different trackers with 4 challenging attributes on VisDrone2018-SOT-test-dev benchmark [56] are shown in Fig. 9.

4.3.4 Evaluation on UAVDT benchmark

The UAVDT [14] benchmark focuses on complex scenarios with new level challenges (*i.e.*, about 80, 000 representative frames from 10 hours raw videos) and 14 kinds of attributes (*e.g.*, weather condition, flying altitude, camera view, vehicle category, and occlusion). The average, min, and max length of a sequence are 778.69, 83, and 2,970 respectively. We compare the ASTR-CF with 19 state-of-the-art deep learning-based trackers, *i.e.*, ASRCF [6], DeepSTRCF [30], UDT+ [54], ADNet [60], CFNet [47], CREST [44], ECO [11], IBCCF [29], MCPF [63], PTAV [15], C-COT [10], GOTURN [24], HDT [42], MDNet [40], SiameseFC [2], STCT [51], RSST [64] MCCT [53] and CFWCR [23]. The performance evaluation in terms of success rate and precision are shown in Fig. 10 and the corresponding tracking speed is shown in Table 2, respectively. Figure 10 depicts that the proposed ASTR-CF outperforms most of the deep learning-based trackers and only MDNet [40] (the winner of VOT2015) ranks higher than the ASTR-CF tracker both in terms of success rate and precision. It is worth mentioning that, with only hand-crafted features, ASTR-CF tracker outperforms DeepSTRCF [30], which

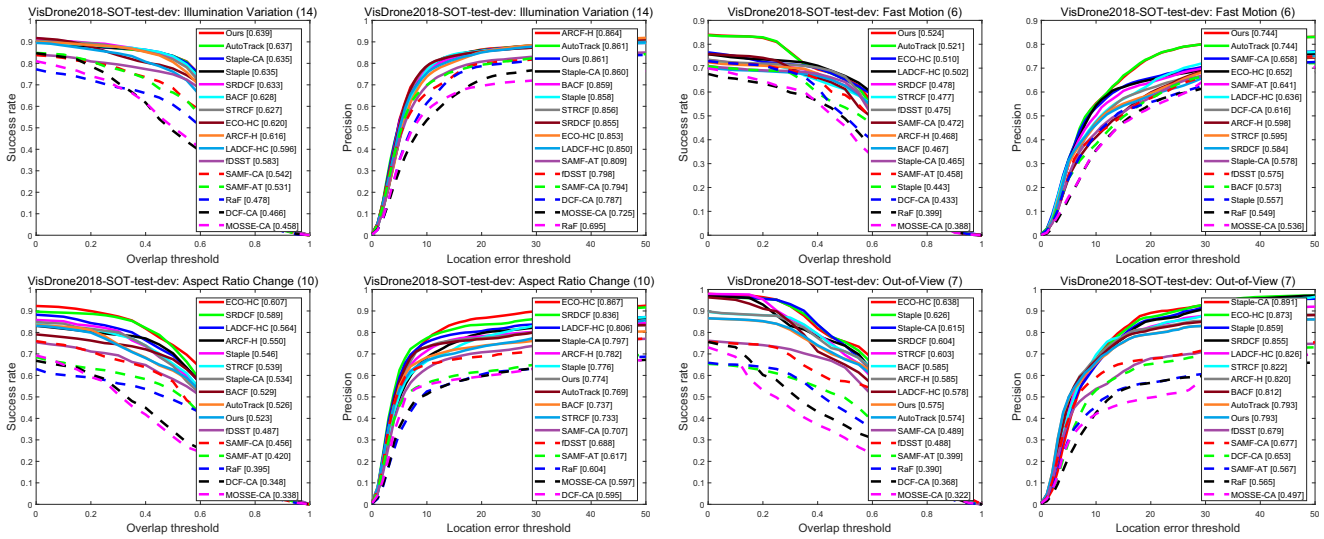

Fig. 9 Evaluation of different trackers with 4 challenging attributes (*i.e.,* IV, FM, ARC and OV), on VisDrone2018-SOT-test-dev benchmark in terms of success rate and precision

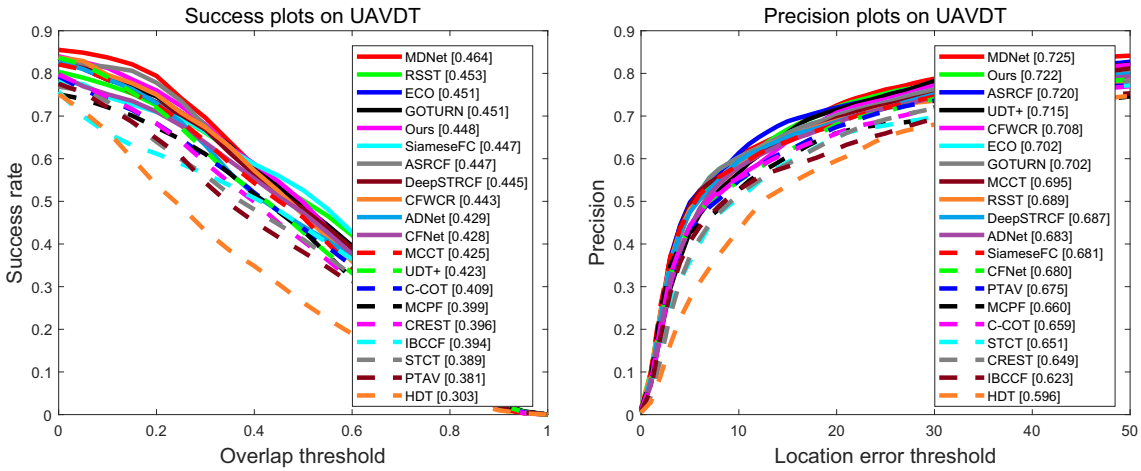

Fig. 10 Performance evaluation between ASTR-CF and 19 deep trackers on UAVDT in terms of success rate and precision

is the deep version of our baseline STRCF [30] in terms of both success rate and precision. For the success rate, the proposed tracker (44.8%) is still higher than that of DeepSTRCF [30] (44.5%), and our precision surpasses DeepSTRCF [30] by 3.5%. Table 2 shows that the proposed ASTR-CF performs the best in tracking speed. Especially, the speed of the ASTR-CF tracker (62.11fps) is more than 60 times faster than MDNet [40] (0.96fps). It is worth mentioning that some deep learning-based trackers above (*e.g.* UDT+, CFNet) can run in real-time on GPU. However, on a UAV mobile device solely with CPU, they can hardly satisfy the real-time needs.

4.3.5 Overall performance evaluation

As aforementioned, both accuracy and speed are significant for UAV-based tracking. Thus, we make an overall performance evaluation of hand-crafted trackers on UAV123@10fps [38], DTB70 [31] and VisDrone2018-SOT-test-dev [56]. The comparative results of top-5 trackers are depicted in Table 3. One can see that the proposed ASTR-CF performs the best in terms of both success rate and precision. This is attributed to the fact that the proposed model can effectively estimate an object-aware spatial regularization and context-aware temporal regularization in an adaptive modality simultaneously. The adaptive temporal regularization enables the learned filters be more robust to occlusion while adapting well to large appearance variation. Meanwhile, the learned filters focus on the reliable features of the target, and they can alleviate the effects of unexpected noises within the object region by introducing adaptive spatial regularization. Meanwhile, the proposed tracker achieves the second fast speed (55.00fps) which is slightly slower than AutoTrack [33] (55.96fps). However, it is two times faster than the baseline STRCF [30] (25.00fps). This is thanks to the adaptive temporal regularization, which can reduce meaningless and detrimental training on contaminated samples.

4.4 Qualitative evaluations

The Qualitative results of the ASTR-CF tracker with 9 hand-crafted trackers and 9 deep-learning based trackers are depicted in Fig. 11 and Fig. 12, respectively. The

Table 2 Comparison of ASTR-CF and 19 deep trackers on UAVDT benchmark in terms of speed. The top-3 results are shown in red, green, and blue fonts, respectively

Tracker	CPU/GPU	FPS	Tracker	CPU/GPU	FPS
RSST	GPU	1.63	ASRCF	GPU	23.01
DeepSTRCF	GPU	3.76	UDT+	GPU	47.93
IBCCF	GPU	4.21	ADNet	GPU	7.55
C-COT	GPU	0.93	ECO	GPU	16.50
CFNet	GPU	41.05	CREST	GPU	4.34
HDT	GPU	9.02	MCPF	GPU	3.63
MDNet	GPU	0.96	PTAV	GPU	26.56
SiameseFC	GPU	37.87	STCT	GPU	1.85
GOTURN	GPU	16.50	CFWCR	GPU	14.27
MCCT	GPU	3.16	Ours	CPU	62.11

Table 3 The average accuracy and speed of the top-5 trackers on UAV123@10fps, DTB70 and VisDrone2018-SOT-test-dev. The top-3 results are shown in red, green, and blue fonts, respectively

	STRCF	LADCF-HC	ECO-HC	AutoTrack	Ours
Success rate	0.467	0.473	0.475	0.492	0.495
Precision	0.657	0.663	0.660	0.703	0.706
FPS	25.00	32.71	43.09	55.96	55.00

targets in these sequences undergo deformation (*e.g.*, BMX4 and SnowBoarding6), partial occlusion (*e.g.*, group2_2, group2_3, uav0000074_11915_s, S0305, S0309, S1401 and S1701), illumination variation (*e.g.*, group2_3, uav0000074_11915_s, S0305 and S1401), rotation (*e.g.*, BMX4 and ManRunning1), camera motion (*e.g.*, BMX4, ManRunning1, group2_2, group2_3 and S0309), small object(*e.g.*, S1401) etc. As shown in the plots, the proposed tracker achieves satisfying performance with fixed parameters for all these sequences.

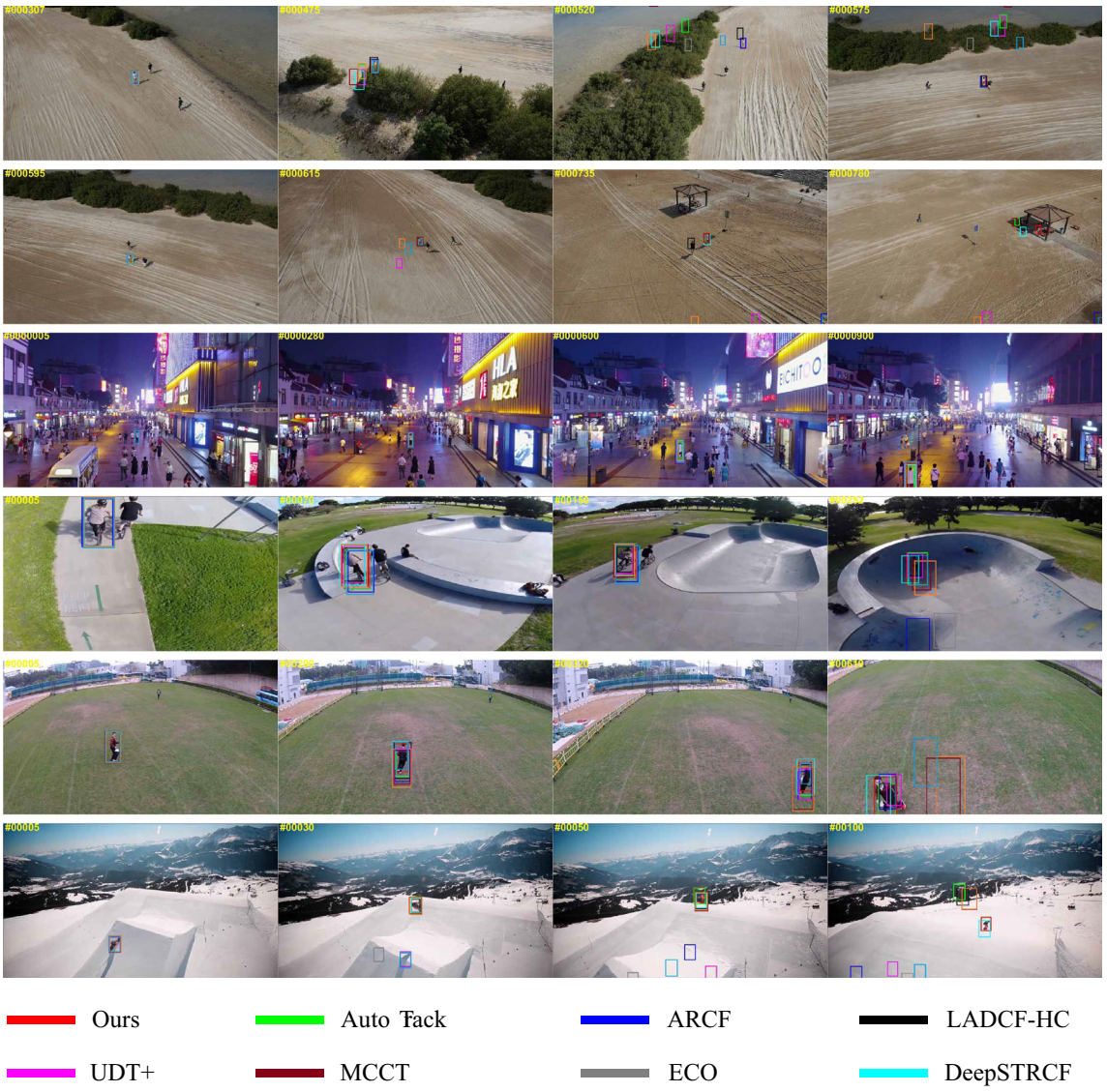

Fig. 11 Qualitative comparison of ASTR-CF tracker with the state-of-the-art hand-crafted trackers on 6 sequences from UAV123@10fps, DTB70 and VisDrone2018-SOT-test-dev benchmarks (*i.e.*, from top to bottom and from left to right, group2_2, group2_3, uav0000074_11915_s, BMX4, ManRunning1 and SnowBoarding6.) The indices of the frames are shown in the top-left of each figure. (Note: Zoom in for a better view.)

Fig. 12 Qualitative comparison of the ASTR-CF tracker with the state-of-the-art deep learning-based trackers on 4 sequences from UAVDT benchmark (*i.e.*, from top to bottom, S0305, S0309, S1401 and S1701). The indices of the frames are shown in the top-left of each figure. (Note: Zoom in for a better view)

4.5 Ablation study

The proposed ASTR-CF model combines ASR (adaptive spatial regularization) and ATR (adaptive temporal regularization) simultaneously. In this section, we demonstrate how much ASR or ATR module contributes to the overall tracking performance through the ablation experiments. More specifically, the proposed tracker is compared to its counterpart with different modules enabled. The overall evaluation is presented in Table 4. It shows that after the ASR module and ATR module being added to the baseline (STRCF [30]), the performance is improved gradually. *e.g.*, the final tracker improves the baseline method by 4.7% and 6.6% in terms of success rate and precision criterion, respectively, on DTB70 benchmark.

4.6 Failure cases

In some challenging sequences, the proposed ASTR-CF model fails in tracking the target. Fig. 13 shows some failure examples. In the sequences of bird1 and car2_s (from UAV123@10fps [38]), the targets undergo out-of-view situation and full occlusion. Even with the adaptive spatial-temporal regularized, the ASTR-CF can not handle such severe appearance variation of targets. For instance, in the car2_s sequence, the appearance of the target is greatly perturbed by full occlusion, which generates a disordered response map and yields the tracking failure. Besides the aforementioned reasons, the lack of enough motion information is another important factor for the tracking failures, as shown in Fig. 13.

Table 4 Ablation analysis of ASTR-CF on DTB70, UAV123@10fps and VisDrone2018 UAV tracking benchmarks

Benchmarks	DTB70		UAV123@10fps		VisDrone2018	
Metrics	Success rate	Precision	Success rate	Precision	Success rate	Precision
Baseline	0.437	0.649	0.457	0.627	0.567	0.778
Baseline+ATR	0.476	0.701	0.474	0.668	0.568	0.782
Baseline+ASR	0.480	0.704	0.470	0.670	0.570	0.786
Ours	0.484	0.715	0.479	0.678	0.574	0.789

Fig. 13 Failure cases of the ASTR-CF (bird1 and car2_s from up to bottom). The results of ASTR-CF are shown in green and the ground truth boxes are denoted in red

5 Conclusion

In this work, a novel Adaptive Spatio-Temporal Regularized Correlation Filters (ASTR-CF) model is proposed to alleviate the boundary effect and filter degradation for UAV-based tracking. An alternating direction method of multipliers (ADMM) algorithm is developed to optimize the ASTR-CF model efficiently. Comparative experiments on 4 UAV tracking benchmarks with more than 30 state-of-the-art trackers are carried out to validate the accuracy and efficiency of the proposed tracker. Experimental results demonstrate that the proposed ASTR-CF outperforms most state-of-the-art trackers, with a speed of exceeded 50fps running on a single CPU.

Acknowledgements The authors sincerely thank Prof. Zheng Liu from the University of British Columbia for his helps to check and revise the organization and language. Many thanks to the anonymous reviewers and editors for the valuable comments and suggestions. This work is supported by the National Natural Science Foundation of China (No. 61801272 and 61601266).

References

1. Bertinetto L, Valmadre J, Golodetz S, Miksik O, Torr PHS (2016) Staple: Complementary learners for real-time tracking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 1401–1409, https://doi.org/10.1109/CVPR.2016.156
2. Bertinetto L, Valmadre J, Henriques JF, Vedaldi A, Torr PHS (2016) Fully-convolutional siamese networks for object tracking. In: Proceedings of the European Conference on Computer Vision Workshops, pp 850–865, https://doi.org/10.1007/978-3-319-48881-3_56
3. Bibi A, Mueller M, Ghanem B (2016) Target response adaptation for correlation filter tracking. In: Proceedings of the European Conference on Computer Vision, pp 419–433, https://doi.org/10.1007/978-3-319-46466-4_25
4. Bolme DS, Beveridge JR, Draper BA, Lui YM (2010) Visual object tracking using adaptive correlation filters. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 2544–2550, https://doi.org/10.1109/CVPR.2010.5539960
5. Boyd S, Parikh N, Chu E (2011) Distributed optimization and statistical learning via the alternating direction method of multipliers. Now Publishers Inc, https://doi.org/10.1561/9781601984616
6. Dai K, Wang D, Lu H, Sun C, Li J (2019) Visual tracking via adaptive spatially-regularized correlation filters. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 466i5–4674, https://doi.org/10.1109/CVPR.2019.00480
7. Dalal N, Triggs B (2005) Histograms of oriented gradients for human detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, vol 1, pp 886–893, https://doi.org/10.1109/CVPR.2005.177
8. Danelljan M, Hager G, Khan FS, Felsberg M (2015) Learning spatially regularized correlation filters for visual tracking. In: Proceedings of the IEEE International Conference on Computer Vision, pp 4310–4318, https://doi.org/10.1109/ICCV.2015.490
9. Danelljan M, Hager G, Shahbaz Khan F, Felsberg M (2016) Adaptive decontamination of the training set: A unified formulation for discriminative visual tracking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 1430–1438, https://doi.org/10.1109/CVPR.2016.159
10. Danelljan M, Robinson A, Shahbaz Khan F, Felsberg M (2016) Beyond correlation filters: Learning continuous convolution operators for visual tracking. In: Proceedings of the European Conference on Computer Vision, pp 472–488, https://doi.org/10.1007/978-3-319-46454-1_29
11. Danelljan M, Bhat G, Khan FS, Felsberg M (2017) Eco: Efficient convolution operators for tracking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 9631–6939, https://doi.org/10.1109/CVPR.2017.733
12. Danelljan M, Häger G, Khan FS, Felsberg M (2017) Discriminative scale space tracking. IEEE Transactions on Pattern Analysis and Machine Intelligence 39(8):1561–1575, https://doi.org/10.1109/TPAMI.2016.2609928
13. De Moraes RS, de Freitas EP (2020) Multi-uav based crowd monitoring system. IEEE Transactions on Aerospace and Electronic Systems 56(2):1332–1345, https://doi.org/10.1109/TAES.2019.2952420
14. Du D, Qi Y, Yu H, Yang Y, Duan K, Li G, Zhang W, Huang Q, Tian Q (2018) The unmanned aerial vehicle benchmark: Object detection and tracking. In: Proceedings of

the European Conference on Computer Vision, pp 375–391, https://doi.org/10.1007/978-3-030-01249-6_23
15. Fan H, Ling H (2017) Parallel tracking and verifying: A framework for real-time and high accuracy visual tracking. In: Proceedings of the IEEE International Conference on Computer Vision, pp 5487–5495, https://doi.org/10.1109/ICCV.2017.585
16. Fan H, Lin L, Yang F, Chu P, Deng G, Yu S, Bai H, Xu Y, Liao C, Ling H (2019) Lasot: A high-quality benchmark for large-scale single object tracking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 5369–5378, https://doi.org/10.1109/CVPR.2019.00552
17. Fu C, Zhang Y, Huang Z, Duan R, Xie Z (2019) Part-based background-aware tracking for uav with convolutional features. IEEE Access 7:79997–80010, https://doi.org/10.1109/ACCESS.2019.2922703
18. Fu C, Xu J, Lin F, Guo F, Liu T, Zhang Z (2020) Object saliency-aware dual regularized correlation filter for real-time aerial tracking. IEEE Transactions on Geoscience and Remote Sensing 58(12):8940–8951, https://doi.org/10.1109/TGRS.2020.2992301
19. Galoogahi HK, Fagg A, Lucey S (2017) Learning background-aware correlation filters for visual tracking. In: Proceedings of the IEEE International Conference on Computer Vision, pp 1145–1152, https://doi.org/10.1109/ICCV.2017.129
20. Gao P, Yuan R, Wang F, Xiao L, Fujita H, Zhang Y (2020) Siamese attentional keypoint network for high performance visual tracking. Knowledge-Based Systems 193:105448, https://doi.org/10.1016/j.knosys.2019.105448
21. Gao P, Zhang Q, Wang F, Xiao L, Fujita H, Zhang Y (2020) Learning reinforced attentional representation for end-to-end visual tracking. Information Sciences 517:52–67, https://doi.org/10.1016/j.ins.2019.12.084
22. Han Z, Wang P, Ye Q (2020) Adaptive discriminative deep correlation filter for visual object tracking. IEEE Transactions on Circuits and Systems for Video Technology 30(1):155–166, https://doi.org/10.1109/TCSVT.2018.2888492
23. He Z, Fan Y, Zhuang J, Dong Y, Bai H (2017) Correlation filters with weighted convolution responses. In: Proceedings of the International Conference on Computer Vision Workshops, pp 1992–2000, https://doi.org/10.1109/ICCVW.2017.233
24. Held D, Thrun S, Savarese S (2016) Learning to track at 100 fps with deep regression networks. In: Proceedings of the European Conference on Computer Vision, pp 749–765, https://doi.org/10.1007/978-3-319-46448-0_45
25. Henriques JF, Caseiro R, Martins P, Batista J (2012) Exploiting the circulant structure of tracking-by-detection with kernels. In: Proceedings of the European Conference on Computer Vision, pp 702–715, https://doi.org/10.1007/978-3-642-33765-9_50
26. Henriques JF, Caseiro R, Martins P, Batista J (2015) High-speed tracking with kernelized correlation filters. IEEE Transactions on Pattern Analysis and Machine Intelligence 37(3):583–596, https://doi.org/10.1109/TPAMI.2014.2345390
27. Huang Z, Fu C, Li Y, Lin F, Lu P (2019) Learning aberrance repressed correlation filters for real-time uav tracking. In: Proceedings of the International Conference on Computer Vision, pp 2891–2900, https://doi.org/10.1109/ICCV.2019.00298
28. Kristan M, Matas J, Leonardis A, Vojir T, Pflugfelder R, Fernandez G, Nebehay G, Porikli F, ÚCCehovin L (2016) A novel performance evaluation methodology for single-target trackers. IEEE Transactions on Pattern Analysis and Machine Intelligence 38(11):2137–2155, https://doi.org/10.1109/TPAMI.2016.2516982
29. Li F, Yao Y, Li P, Zhang D, Zuo W, Yang M (2017) Integrating boundary and center correlation filters for visual tracking with aspect ratio variation. In: Proceedings of the IEEE International Conference on Computer Vision Workshop, pp 2001–2009, https://doi.org/10.1109/ICCVW.2017.234
30. Li F, Tian C, Zuo W, Zhang L, Yang M (2018) Learning spatial-temporal regularized correlation filters for visual tracking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 4904–4913, https://doi.org/10.1109/CVPR.2018.00515
31. Li S, Yeung DY (2017) Visual object tracking for unmanned aerial vehicles: A benchmark and new motion models. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp 4140–4146
32. Li X, Liu Q, He Z, Wang H, Zhang C, Chen WS (2016) A multi-view model for visual tracking via correlation filters. Knowledge-Based Systems 113:88–99, https://doi.org/10.1016/j.knosys.2016.09.014
33. Li Y, Fu C, Ding F, Huang Z, Lu G (2020) Autotrack: Towards high-performance visual tracking for uav with automatic spatio-temporal regularization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 11923–11932, https://doi.org/10.1109/CVPR42600.2020.01194
34. Li Y, Fu C, Huang Z, Zhang Y, Pan J (2021) Intermittent contextual learning for keyfilter-aware uav object tracking using deep convolutional feature. IEEE Transactions on Multimedia 23:810–822, https://doi.org/10.1109/TMM.2020.2990064
35. Liu T, Wang G, Yang Q (2015) Real-time part-based visual tracking via adaptive correlation filters. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 4902–4912, https://doi.org/10.1109/CVPR.2015.7299124
36. Lukezic A, Vojir T, CehovinZajc L, Matas J, Kristan M (2018) Discriminative correlation filter tracker with channel and spatial reliability. International Journal of Computer Vision 126(7):671–688, https://doi.org/10.1007/s11263-017-1061-3
37. Ma C, Huang JB, Yang X, Yang MH (2019) Robust visual tracking via hierarchical convolutional features. IEEE Transactions on Pattern Analysis and Machine Intelligence 41(11):2709–2723, https://doi.org/10.1109/TPAMI.2018.2865311
38. Mueller M, Smith N, Ghanem B (2016) A benchmark and simulator for uav tracking. In: Proceedings of the European Conference on Computer Vision, pp 445–461, https://doi.org/10.1007/978-3-319-46448-0_27
39. Mueller M, Smith N, Ghanem B (2017) Context-aware correlation filter tracking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 1387–1395, https://doi.org/10.1109/CVPR.2017.152
40. Nam H, Han B (2016) Learning multi-domain convolutional neural networks for visual tracking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 4293–4302, https://doi.org/10.1109/CVPR.2016.465
41. Padhy RP, Xia F, Choudhury SK, Sa PK, Bakshi S (2019) Monocular vision aided autonomous uav navigation in indoor corridor environments. IEEE Transactions on Sustainable Computing 4(1):96–108, https://doi.org/10.1109/TSUSC.2018.2810952
42. Qi Y, Zhang S, Qin L, Yao H, Huang Q, Lim J, Yang M (2016) Hedged deep tracking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 4303–4311, https://doi.org/10.1109/CVPR.2016.466
43. Rey N, Volpi M, Joost S, Tuia D (2017) Detecting animals in african savanna with uavs and the crowds. Remote Sensing of Environment 200:341–351, https://doi.org/10.1016/j.rse.2017.08. 026
44. Song Y, Ma C, Gong L, Zhang J, Lau RWH, Yang M (2017) Crest: Convolutional residual learning for visual tracking. In: Proceedings of the IEEE International Conference on Computer Vision, pp 2574–2583, https://doi.org/10.1109/ICCV.2017.279
45. Sun Y, Sun C, Wang D, He Y, Lu H (2019) Roi pooled correlation filters for visual tracking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 5776–5784, https://doi.org/10.1109/CVPR.2019.00593

46. Tang Y, Hu Y, Cui J, Liao F, Lao M, Lin F, Teo RSH (2019) Vision-aided multi-uav autonomous flocking in gps-denied environment. IEEE Transactions on Industrial Electronics 66(1):616–626, https://doi.org/10.1109/TIE.2018.2824766
47. Valmadre J, Bertinetto L, Henriques J, Vedaldi A, Torr PHS (2017) End-to-end representation learning for correlation filter based tracking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 5000–5008, https://doi.org/10.1109/CVPR.2017.531
48. van de Weijer J, Schmid C, Verbeek J, Larlus D (2009) Learning color names for real-world applications. IEEE Transactions on Image Processing 18(7):1512–1523, https://doi.org/10.1109/TIP.2009.2019809
49. Voigtlaender P, Luiten J, Torr PHS, Leibe B (2020) Siam r-cnn: Visual tracking by re-detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, https://doi.org/10.1109/CVPR42600.2020.00661
50. Wang C, Wang J, Shen Y, Zhang X (2019) Autonomous navigation of uavs in large-scale complex environments: A deep reinforcement learning approach. IEEE Transactions on Vehicular Technology 68(3):2124–2136, https://doi.org/10.1109/TVT.2018. 2890773
51. Wang L, Ouyang W, Wang X, Lu H (2016) Stct: Sequentially training convolutional networks for visual tracking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 1373–1381, https://doi.org/10.1109/CVPR.2016.153
52. Wang M, Liu Y, Huang Z (2017) Large margin object tracking with circulant feature maps. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 4800–4808, https://doi.org/10.1109/CVPR.2017.510
53. Wang N, Zhou W, Tian Q, Hong R, Wang M, Li H (2018) Multi-cue correlation filters for robust visual tracking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 4844–4853, https://doi.org/10.1109/CVPR.2018.00509
54. Wang N, Song Y, Ma C, Zhou W, Liu W, Li H (2019) Unsupervised deep tracking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 1308–1317, https://doi.org/10.1109/CVPR.2019.00140
55. Wang W, Zhang K, Lv M, Wang J (2020) Hierarchical spatiotemporal context-aware correlation filters for visual tracking. IEEE Transactions on Cybernetics pp 1–14, https://doi.org/10.1109/TCYB.2020.2964757
56. Wen L, Zhu P, Du D, et al (2019) Visdrone-sot2018: The vision meets drone single-object tracking challenge results. In: Proceedings of the European Conference on Computer Vision, pp 469–495, https://doi.org/10.1007/978-3-030-11021-5_28
57. Wu Y, Lim J, Yang MH (2015) Object tracking benchmark. IEEE Transactions on Pattern Analysis and Machine Intelligence 37(9):1834–1848, https://doi.org/10.1109/TPAMI.2014.2388226
58. Xu T, Feng ZH, Wu XJ, Kittler J (2019) Learning adaptive discriminative correlation filters via temporal consistency preserving spatial feature selection for robust visual object tracking. IEEE Transactions on Image Processing 28(11):5596–5609, https://doi.org/10.1109/TIP.2019.2919201
59. Xu Y, Wang Z, Li Z, Yuan Y, Yu G (2020) Siamfc++: Towards robust and accurate visual tracking with target estimation guidelines. In: Proceedings of the Association for the Advance of Artificial Intelligence, https://doi.org/10.1609/aaai.v34i07.6944
60. Yun S, Choi J, Yoo Y, Yun K, Choi JY (2017) Action-decision networks for visual tracking with deep reinforcement learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 1349–1358, https://doi.org/10.1109/CVPR.2017.148
61. Zahran S, Moussa AM, Sesay AB, El-Sheimy N (2019) A new velocity meter based on hall effect sensors for uav indoor navigation. IEEE Sensors Journal 19(8):3067–3076, https://doi.org/10.1109/JSEN.2018.2890094
62. Zhang L, Varadarajan J, Suganthan PN, Ahuja N, Moulin P (2017) Robust visual tracking using oblique random forests. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 5825–5834, https://doi.org/10.1109/CVPR.2017.617
63. Zhang T, Xu C, Yang M (2017) Multi-task correlation particle filter for robust object tracking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 4819–4827, https://doi.org/10.1109/CVPR.2017.512
64. Zhang T, Xu C, Yang M (2019) Robust structural sparse tracking. IEEE Transactions on Pattern Analysis and Machine Intelligence 41(2):473–486, https://doi.org/10.1109/TPAMI.2018.2797082
65. Zhang W, Song K, Rong X, Li Y (2019) Coarse-to-fine uav target tracking with deep reinforcement learning. IEEE Transactions on Automation Science and Engineering 16(4):1522–1530, https://doi.org/10.1109/TASE.2018.2877499

Libin Xu is pursuing the M.S. degree at the School of Electrical and Electronic Engineering, Shandong University of Technology, Zibo, China. His research interests include visual tracking and deep learning.

Mingliang Gao received his PhD in communication and information systems from Sichuan University, Chengdu, China, in 2013. He is currently an associate professor at the School of Electrical and Electronic Engineering, Shandong University of Technology, Zibo, China. His main research interests include computer vision and deep learning.

Qilei Li is currently a Ph.D student with the School of Electronic Engineering and Computer Science, Queen Mary University of London, London, United Kingdom. He received the M.S. degree in signal and information processing from Sichuan University. His research interests are computer vision and deep learning.

Guofeng Zou received the B.S. degree in electrical engineering and automation from College of Electrical and Electronic Engineering, Shandong University of Technology, Zibo, China, in 2007, and the Ph.D. degree in pattern recognition and intelligent system form the College of Automation, Harbin Engineering University, Harbin, in 2013. He is currently working as a lecturer in the College of Electrical and Electronic Engineering, Shandong University of Technology, Zibo, China. His current research interests include pattern recognition, digital image processing and analysis, machine learning.

Jinfeng Pan received her Ph.D in signal and information processing from University of Chinese Academy of Sciences, Xi'an, China, in 2016. She is currently an associate professor at the School of Electrical and Electronic Engineering, Shandong University of Technology, Zibo, China. Her main research interests include computer vision and deep learning.

Jun Jiang received his PhD in communication and information systems from Sichuan University, Chengdu, China, in 2015. Currently, he is currently an associate professor with the School of Computer Science, Southwest Petroleum University, Chengdu, China. His main research interests include machine vision and deep learning.