UAV Tracking Based on Correlation Filters With Dynamic Aberrance-Repressed Temporal Regularizations

Hong Zhang, Yan Li, Yifan Yang, Yachun Feng, Yawei Li, Chenwei Deng, and Ding Yuan, *Member, IEEE*

***Abstract*—As a significant research direction in remote sensing fields, unmanned aerial vehicles (UAVs) tracking has achieved rapid development in recent years. However, due to limited power and computation resources on aerial platforms, the tracking methods deployed on UAVs usually require high computational efficiency and performance. In addition, various challenges (i.e., similar object, background clutter, and occlusion) have inevitably occurred during the UAV tracking phase. Therefore, considering the above issues comprehensively, this article proposes a dynamic aberrance-repressed temporal regularized correlation filter (CF) to achieve stable tracking in UAV remote sensing videos. First, we have introduced the aberrance-repressed temporal regularizations into the discriminative CF framework. Second, a novel objective loss function is constructed to adjust the strength of each regularization for training the filter. Then, a new judgment mechanism based on the response variation is exploited to reflect the response fluctuation and applied to tune parameters of both regularizations. Finally, comprehensive experiments are done on three different UAV benchmarks, i.e., UAV123@10fps, UAVDT, and VisDrone2018, to verify the performance of our tracker and have demonstrated that our tracker achieves superior performance against other total 25 state-of-the-art trackers while reaching ∼35 FPS on a single CPU.**

***Index Terms*—Discriminative correlation filter (DCF), dynamic aberrance-repressed temporal regularizations, unmanned aerial vehicles (UAV) tracking.**

I. Introduction

VISUAL tracking based on unmanned aerial vehicle (UAV) remote sensing videos is an important research direction of remote sensing [1], [2], [3]. Generally, UAV remote sensing videos are shot by onboard cameras at a higher flight altitude, which contain a variety of ground-based observation objects. With the given target information from the first frame, trackers deployed on UAV platforms can realize specific target tracking in subsequent remote sensing videos. Simultaneously, UAVs with tracking capabilities also provide a new research idea for Earth observation and have been applied to the following fields, such as glacier observation [4], vehicle recognition [5], forest fire analysis [6], and object detection [7], [8], [9], etc. Discriminative correlation filter (DCF)-based trackers [10], [11], [12], [13] intend to learn a filter online that can distinguish the object in the foreground from the entire environment. Due to the circular correlation method and training samples generated by cyclic shift, DCF-based trackers can convert complex computation in the spatial domain into dot-product operation in the frequency domain by using fast Fourier transforms (FFTs). Compared with deep learning-based trackers, DCF-based trackers do not rely on high-performance GPUs and are suitable for UAV platforms that are usually equipped with CPUs. Nevertheless, UAV tracking generally encounters some challenges, including similar object, background clutter, and occlusion, which often cause DCF-based trackers to drift or lose targets.

Several DCF-based trackers have been proposed for the above challenging scenarios, mainly including response regularization-based and temporal regularization-based methods. In the response regularization-based methods, Mueller et al. [14] utilized contextual information to establish a background penalty term, which can oppress the response of background patches around the target. Zhang et al. [15] constructed a sparse response regularized term by using $l2$-norm constraint, which attempts to remove unexpected response peaks. Huang et al. [16] exploited a special aberrance-repressed regularization aiming to suppress background distractors by constraining responses in adjacent frames.

Although these above methods can alleviate the abnormal response variation caused by distractors in the background, the quality of response generated by the filter will still decline due to occlusion. In the temporal regularization-based methods, Li et al. [17] proposed a spatial-temporal regularized correlation filter (STRCF) that can maintain the consistency between adjacent filters. Hence, STRCF can prevent the sudden change of the filter in occlusion scenarios. However, Li et al. [17] did not consider the issue of distractors around the target, which results in the peak value of distractors being higher than the

Manuscript received 19 June 2023; revised 17 July 2023; accepted 11 August 2023. Date of publication 17 August 2023; date of current version 31 August 2023. This work was supported by the National Natural Science Foundation of China under Grant 62002005 and Grant 61972015. *(Corresponding author: Ding Yuan.)*

Hong Zhang, Yan Li, Yawei Li, and Ding Yuan are with the School of Astronautics, Beihang University, Beijing 100191, China (e-mail: dmrzhang@buaa.edu.cn; yanliz@buaa.edu.cn; lyw3074@buaa.edu.cn; dyuan@buaa.edu.cn).

Yifan Yang is with the Institute of Artificial Intelligence, Beihang University, Beijing 100191, China (e-mail: stephenyoung@buaa.edu.cn).

Yachun Feng is with the School of Mechanical Engineering and Automation, Beihang University, Beijing 100191, China (e-mail: zb2007103@buaa.edu.cn).

Chenwei Deng is with the School of Information and Electronics, Beijing Institute of Technology, Beijing 100081, China (e-mail: cwdeng@bit.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2023.3306273

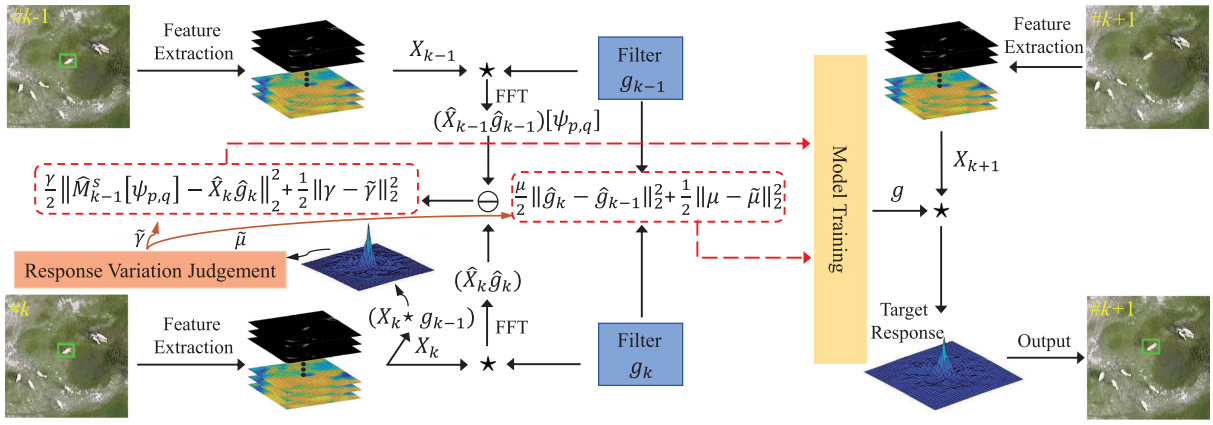

Fig. 1. Flowchart of the proposed method. The aberrance-repressed and temporal regularizations are all combined into the training model. The dynamic parameters (γ and μ) are tuned under the guidance of $\tilde{\gamma}$ and $\tilde{\mu}$ that are developed by the judgment mechanism based on response variation. Furthermore, the target position can be obtained with the target response generated by the learned filter in the upcoming frame.

target's so that the target position mistakenly locates on one of these distractors. Therefore, an effective DCF-based tracking method that can simultaneously suppress distractors in complex scenarios and deal with occlusion challenges is the research goal of this article.

In our work, we propose a dynamic aberrance-repressed temporal regularized correlation filter (DARTCF), as shown in Fig. 1, which introduces aberrance-repressed and temporal regularizations into the DCF framework. Then, we propose a new judgment mechanism based on response variation, which reflects the fluctuation degree of response and participates in the automatic tuning process of regularization parameters. By virtue of tuned parameters, we can control the respective strength of each regularization during the filter training phase. Finally, we employ the alternating direction method of multipliers (ADMM) [18] to find the closed-form solution of our method. Experimental results evaluated on three public UAV benchmarks have proved that the proposed tracker can achieve excellent tracking performance.

The main contributions of this article are as follows.

1) A novel DCF-based framework is constructed for dynamically learning aberrance-repressed and temporal regularizations and adjusting the strengths of both regularizations online for UAV tracking.
2) A novel judgment mechanism based on response variation is proposed to reflect the fluctuation degree of response and participates in the parameter tuning process of the aberrance-repressed and temporal regularizations.
3) Extensive experiments have been undertaken on three UAV benchmarks (i.e., UAV123@10fps [19], UAVDT [20], and VisDrone2018 [21]), and have demonstrated that the proposed DARTCF can achieve superior tracking performance against other total 25 advanced trackers while reaching 35.37 FPS on a single CPU.

II. Related Works

A. Tracking With DCF-Based Methods

The DCF-based method has attracted researchers' attention in the UAV field because of its high computational efficiency and tracking accuracy. Bolme et al. [22] first proposed the correlation filter (CF) method for visual object tracking, which derives the minimum output sum of squared error between the ideal given response and the correlation response. In [23], a kernelized CF method is proposed to realize better performance, which is equipped with the histogram of oriented gradient (HOG) [24] feature. In addition, considering the unwanted boundary effect existed in DCF-based methods, a novel spatial weight term in [25] has been applied to the filter, which can diminish the filter's learning for background information and keep the filter focusing on the target for alleviating the boundary effect. Galoogahi et al. [26] proposed a background-aware CF (BACF), which utilizes a binary matrix to crop real training samples aiming to strengthen the filter's discriminative ability and reduce the boundary effect. Lukezic et al. [27] proposed a DCF-based tracker with channel and spatial reliability (CSR-DCF), which dynamically generates a spatial reliability map suitable for the object to restrict the filter. Han et al. [11] proposed a state-aware antidrift tracker (SAT), in which a novel color-based mask is developed to segment the reliable target region and prevent the filter from background interferences. Dai et al. [28] proposed an adaptive spatially regularized CF, which can generate reliable filter coefficients and combine shallow and deep features to obtain the accurate object's position. Furthermore, Xing et al. [29] introduced a multipeak-redetection mechanism into the DCF-based method to track the target stably. Han et al. [30] incorporated a robust feature representation into the DCF framework by utilizing $l1$-norm to automatically select significant feature for improving the tracking performance.

B. Tracking With Aberrance-Repressed Regularization

The tracking methods with aberrance-repressed regularization aim to suppress response aberrance and maintain the reliability of response under scenarios (i.e., similar object and background clutter). In [14] and [15], an independent response regularization is introduced into the DCF framework to suppress background interferences and remove abnormal peaks surrounded by the target. In [16], a novel regularized term has been developed, which restricts adjacent response

variations to achieve aberrance repression. Elayaperumal and Joo [31] proposed an aberrance suppressed spatio-temporal CF, which introduces the spatial-temporal regularization for reducing the boundary effect and avoiding the filter's mutation. The aberrance-repressed regularization with a fixed control coefficient in [31] is employed to alleviate abnormal responses. Furthermore, the authors in [32], [33], [34], and [35] also introduced the aberrance-repressed regularization into the DCF-based model and use constant regularization parameters to participate in model training.

C. Tracking With Temporal Regularization

To address the issue of filter degradation and mitigate sudden changes of the filter, several methods are proposed by keeping the filter similar to the historical filter. Li et al. [17] exploited a temporal regularization by restricting two adjacent filters to be consistent, which can prevent the learned filter from degradation and cope with occlusion better. Han et al. [36] integrated the spatial-temporal constrain into the DCF framework, which effectively deals with sudden appearance variations during the tracking phase. In [37], an automatic spatio-temporal regularized DCF (AutoTrack) is proposed, which uses local and global response variations to adaptively tune hyperparameters of both regularizations.

To sum up, the authors in [14], [15], and [16] usually employed an independent response regularization to address the aberrance issue without considering the filter degradation. In addition, the authors in [31], [32], [33], [34], and [35] have also chosen the fixed-parameter aberrance-repressed regularization to deal with response interferences for improving the tracking performance. However, the authors in [31], [32], [33], [34], and [35] could not automatically adjust the parameter that belongs to the aberrance-repressed regularization based on response variation. In [17], [36], and [37], although these methods have utilized the temporal regularization to avoid filter degradation, they did not consider the problem of response interferences caused by similar object and background clutter, etc. Thus, in this article, we propose a novel DCF-based method that is not only a combination of aberrance-repressed and temporal regularizations but also can automatically tune both regularized parameters by utilizing the global response variation. Since the strengths of both regularizations in the training model can be adaptively changed, our method has a more advanced ability to brace for different scenarios.

III. Proposed Method

A. Loss Function of DARTCF

The proposed tracker DARTCF is based on the aberrance-repressed correlation filter (ARCF) [16] that has introduced an aberrance-repressed regularization into the DCF-based framework. The overall objective of ARCF can be constructed as follows:

$$E_{\text{ARCF}}(w_k) = \frac{1}{2}\|y - X_k(I_D \otimes B^\top)w_k\|_2^2 + \frac{\lambda}{2}\|w_k\|_2^2 + \frac{\gamma}{2}\|M_{k-1}[\psi_{p,q}] - X_k(I_D \otimes B^\top)w_k\|_2^2 \quad (1)$$

where subscripts k and $k-1$ indicate the kth and $(k-1)$th frame in the tracking sequence. Subscript D denotes the number of feature channels. X_k is the matrix form of the kth input feature sample $x_k^d \in R^N (d = 1, 2, 3 \ldots D)$. I_D signifies a $D \times D$ identity matrix. The size of X_k is $N \times DN$. $B \in R^{M \times N}$ is a binary matrix that crops the central elements of x_k^d. Operator $\otimes$ and $\top$ represent the Kronecker production and conjugate transpose. $w_k^d \in R^M$ is the CF to be trained in the kth frame. $M \ll N$. w_k is the matrix form of w_k^d and its size is $DM \times 1$. M_{k-1} is the response map calculated by $X_{k-1}(I_D \otimes B^\top)w_{k-1}$ from the previous frame. $M_{k-1} \in R^N$. p and q stand for the location difference between two peaks of response maps in two adjacent frames. Besides, $[\psi_{p,q}]$ denotes the shifting operation that makes these two peaks coincide with each other. λ indicates a common regularization parameter. γ denotes the aberrance-repressed regularization parameter. $y \in R^N$ is a Gaussian label as the desired response map. Based on the above, we construct the objective loss function in our work as follows:

$$E(w_k) = E_{\text{ARCF}}(w_k) + \frac{\mu}{2}\|g_k - g_{k-1}\|_2^2 + \frac{1}{2}\|\gamma - \tilde{\gamma}\|_2^2 + \frac{1}{2}\|\mu - \tilde{\mu}\|_2^2 \quad (2)$$

where $g_k = (I_D \otimes B^\top)w_k$ and $g_{k-1} = (I_D \otimes B^\top)w_{k-1}$. The sizes of g_k and g_{k-1} are $DN \times 1$. μ represents the temporal regularization parameter. Here, $\|g_k - g_{k-1}\|_2^2$ is the temporal regularization term. $\tilde{\gamma}$ and $\tilde{\mu}$ are two guiding values belonging to γ and μ, respectively. Although the objective loss function can be converted to matrix form as (2), it could still perform massive correlation operations in the spatial domain [16]. Therefore, (2) is transformed into frequency domain to speed up the computing efficiency as follows:

$$\begin{aligned}\hat{E}(w_k, \hat{g}_k) = &\frac{1}{2}\|\hat{y} - \hat{X}_k\hat{g}_k\|_2^2 + \frac{\lambda}{2}\|w_k\|_2^2 + \frac{\gamma}{2}\|\hat{M}_{k-1}^s - \hat{X}_k\hat{g}_k\|_2^2 \\ &+ \frac{\mu}{2}\|\hat{g}_k - \hat{g}_{k-1}\|_2^2 + \frac{1}{2}\|\gamma - \tilde{\gamma}\|_2^2 + \frac{1}{2}\|\mu - \tilde{\mu}\|_2^2 \\ s.t \quad &\hat{g}_k = \sqrt{N}(I_D \otimes FB^\top)w_k \end{aligned} \quad (3)$$

where subscript $\hat{}$ indicates the discrete Fourier transformation (DFT), i.e., $\hat{\alpha} = \sqrt{N}F\alpha$, where F is an orthonormal $N \times N$ matrix. $\hat{g}_k$ denotes an auxiliary variable that is used for subsequent optimization of (3). $\hat{M}_{k-1}^s$ is the form of $M_{k-1}[\psi_{p,q}]$ after the DFT. Besides, in order to simplify the calculation, $\hat{M}_{k-1}^s$ can be considered as a constant in the current frame that has been calculated in the previous frame.

B. Optimization

Considering the convexity of (3), it can be minimized by the ADMM method and obtain the global optimal solution. Therefore, we can convert it into the augmented Lagrangian form as follows:

$$\hat{E}(w_k, \hat{g}_k, \hat{\zeta}) = \frac{1}{2}\|\hat{y} - \hat{X}_k\hat{g}_k\|_2^2 + \frac{\lambda}{2}\|w_k\|_2^2$$

$$+\frac{\gamma}{2}\|\hat{M}^s_{k-1}-\hat{X}_k\hat{g}_k\|_2^2 +\frac{\mu}{2}\|\hat{g}_k-\hat{g}_{k-1}\|_2^2+\frac{1}{2}\|\gamma-\tilde{\gamma}\|_2^2+\frac{1}{2}\|\mu-\tilde{\mu}\|_2^2 +\hat{\zeta}^\top(\hat{g}_k-\sqrt{N}(I_D\otimes FB^\top)w_k) +\frac{\eta}{2}\|\hat{g}_k-\sqrt{N}(I_D\otimes FB^\top)w_k\|_2^2 \tag{4}$$

where η is a fixed penalty factor. $\hat{\zeta}=[\hat{\zeta}_1^\top,\ldots,\hat{\zeta}_D^\top]$ is an auxiliary variable as the Lagrangian vector whose size is $DN\times 1$. With the ADMM method, the solution of (4) can be obtained by alternatingly solving the following subproblems as follows:

$$\begin{aligned} w_k^* &= \arg\min_{w_k}\left\{\frac{\lambda}{2}\|w_k\|_2^2+\hat{\zeta}^\top(\hat{g}_k-\sqrt{N}(I_D\otimes FB^\top)w_k) +\frac{\eta}{2}\|\hat{g}_k-\sqrt{N}(I_D\otimes FB^\top)w_k\|_2^2\right\}\\ \hat{g}_k^* &= \arg\min_{g_k}\left\{\frac{1}{2}\|\hat{y}-\hat{X}_k\hat{g}_k\|_2^2+\frac{\gamma}{2}\|\hat{M}^s_{k-1}-\hat{X}_k\hat{g}_k\|_2^2 +\frac{\mu}{2}\|\hat{g}_k-\hat{g}_{k-1}\|_2^2+\hat{\zeta}^\top(\hat{g}_k-\sqrt{N}(I_D\otimes FB^\top)w_k) +\frac{\eta}{2}\|\hat{g}_k-\sqrt{N}\left(I_D\otimes FB^\top\right)w_k\|_2^2\right\}\\ \gamma^* &= \arg\min_{\gamma}\left\{\frac{\gamma}{2}\|\hat{M}^s_{k-1}-\hat{X}_k\hat{g}_k\|_2^2+\frac{1}{2}\|\gamma-\tilde{\gamma}\|_2^2\right\}\\ \mu^* &= \arg\min_{\mu}\left\{\frac{\mu}{2}\|\hat{g}_k-\hat{g}_{k-1}\|_2^2+\frac{1}{2}\|\mu-\tilde{\mu}\|_2^2\right\}. \end{aligned} \tag{5}$$

Then, the closed-form solution of each subproblem has been given in detail as follows.

Subproblem for w_k^*: With the simplification of w_k^*, we can obtain the following:

$$w_k^*=\left(\frac{\lambda}{N}+\eta\right)^{-1}(\zeta+\eta g_k). \tag{6}$$

Here, by means of the inverse fast Fourier transform (IFFT), we could obtain two equations that are $g_k=\frac{1}{\sqrt{N}}(I_D\otimes BF^\top)\hat{g}_k$ and $\zeta=\frac{1}{\sqrt{N}}(I_D\otimes BF^\top)\hat{\zeta}_k$.

Subproblem for $\hat{g}_k^*$: Since $\hat{X}_k\hat{g}_k$ is included in $\hat{g}_k^*$, the computation for solving $\hat{g}_k^*$ consumes much time during the process of every ADMM iteration. However, considering that $\hat{X}_k$ is sparse banded [26], each element of $\hat{y}(\hat{y}(n),n=1,2,\ldots,N)$ is merely related to each $\hat{x}_k(n)=[\hat{x}_k^1(n),\hat{x}_k^2(n),\ldots,\hat{x}_k^D(n)]^\top$ and $\hat{g}_k(n)=[\text{conj}(\hat{g}_k^1(n)),\ldots,\text{conj}(\hat{g}_k^D(n))]^\top$ [38]. Operator conj(.) signifies the complex conjugate that is applied on complex vector.

Therefore, solving subproblem $\hat{g}_k^*$ in (5) can be considered to solve N smaller independent problems. Solving for $\hat{g}_k^*(n)$, $n=[1,2,3,\ldots,N]$ can be seen as follows:

$$\hat{g}_k^*(n)=\arg\min_{g_k(n)}\left\{\frac{1}{2}\|\hat{y}(n)-\hat{x}_k^\top\hat{g}_k(n)\|_2^2+\frac{\mu}{2}\|\hat{g}_k-\hat{g}_{k-1}\|_2^2 +\frac{\gamma}{2}\|\hat{M}^s_{k-1}-\hat{x}_k^\top\hat{g}_k(n)\|_2^2+\hat{\zeta}^\top\left(\hat{g}_k(n)-\hat{w}_k(n)\right) +\frac{\eta}{2}\|\hat{g}_k(n)-\hat{w}_k(n)\|_2^2\right\} \tag{7}$$

where $\hat{w}_k(n)=[\hat{w}_k^1(n),\hat{w}_k^2(n)\ldots,\hat{w}_k^D(n)]$ and $\hat{w}_k^d(n)=\sqrt{D}\text{FB}^\top w_k^d$, $d=[1,2,\ldots,D]$. Actually, $\hat{w}_k^d$ can be obtained by performing an FFT on w_k^d. Solving for each $\hat{g}_k^*$ will appear the inverse operation. Thus, we convert the inverse operation into another form by introducing the Sherman–Morrison formula, i.e., $(A+uv^\top)^{-1}=A^{-1}-A^{-1}u(I+v^\top A^{-1}u)^{-1}v^\top A^{-1}$. Furthermore, (7) can be rewritten as

$$\begin{aligned} &\hat{g}_k^*(n)\\ &=\alpha\left(\hat{S}_{xy}(n)+\gamma\hat{x}_k(n)\hat{M}^s_{k-1}-\hat{\zeta}(n)+\eta\hat{w}_k(n)+\mu\hat{g}_{k-1}(n)\right)\\ &\quad-\alpha\frac{\hat{x}_k}{\beta}\left(\hat{S}_{xk}\hat{y}(n)+\gamma\hat{S}_{xk}\hat{M}^s_{k-1}-\hat{S}_\zeta+\eta\hat{S}_{wk}+\mu\hat{S}_{xg}\right) \end{aligned} \tag{8}$$

where $\alpha=\frac{1}{\eta+\mu}$, $\beta=\frac{\eta+\mu}{1+\gamma}+\hat{S}_{xk}(n)$, $\hat{S}_{xy}(n)=\hat{x}_k(n)\hat{y}(n)$, $\hat{S}_{xk}(n)=\hat{x}_k^\top(n)\hat{x}_k(n)$, $\hat{S}_\zeta(n)=\hat{x}_k^\top(n)\hat{\zeta}$, $\hat{S}_{wk}=\hat{x}_k^\top(n)\hat{w}_k(n)$, and $\hat{S}_{xg}=\hat{x}_k^\top(n)\hat{g}_{k-1}(n)$.

Subproblem for γ^*: Given the variables, i.e., $\hat{M}^s_{k-1}$, $\hat{X}_k$, $\hat{g}_k$, and $\tilde{\gamma}$, the derivation for optimal γ can be seen as follows:

$$\gamma^*=\tilde{\gamma}-\frac{\|\hat{M}^s_{k-1}-\hat{X}_k\hat{g}_k\|_2^2}{2}. \tag{9}$$

Subproblem for μ^*: Given the variables, i.e., $\hat{g}_k$, $\hat{g}_{k-1}$, and $\tilde{\mu}$, we can obtain the optimal solution of μ as follows:

$$\mu^*=\tilde{\mu}-\frac{\|\hat{g}_k-\hat{g}_{k-1}\|_2^2}{2}. \tag{10}$$

Lagrangian Update: After solving w_k^* and $\hat{g}_k^*$ in turn, we update the Lagrangian parameter as follows:

$$\begin{aligned} \hat{\zeta}_k^{j+1}&=\hat{\zeta}_k^j+\eta^{j+1}\left(\hat{g}_k^{*(j+1)}-\hat{w}_k^{*(j+1)}\right)\\ \eta^{j+1}&=\min(\eta_{\max},\phi\eta^j) \end{aligned} \tag{11}$$

where subscripts j and $(j+1)$ express the jth and $(j+1)$th iterations in the computation process of ADMM, respectively. $\hat{g}_k^{*(j+1)}$ and $\hat{w}_k^{*(j+1)}$ are the $(j+1)$th solutions of above subproblems $\hat{g}_k^*$ and $\hat{w}_k^*$ in iterations. Besides, $\hat{w}_k^{*(j+1)}=(I_D\otimes FB^\top)w_k^{*(j+1)}$. $\eta_{\max}$ is the maximum of η. ϕ is a fixed scale factor.

Appearance Model: The appearance model $\hat{x}_k^{\text{model}}$ is constructed as follows:

$$\hat{x}_k^{\text{model}}=(1-\delta)\,\hat{x}_{k-1}^{\text{model}}+\delta\hat{x}_k \tag{12}$$

where $\hat{x}_{k-1}^{\text{model}}$ represents the previous appearance model in the $(k-1)$th frame, $\hat{x}_k$ is the extracted feature in the kth frame, and δ denotes the fixed learning rate. Generally, δ ranges from 0 to 0.1 (i.e., $0<\delta\leq 0.1$), which can make $\hat{x}_k^{\text{model}}$ maintain more historical appearance information and keep the tracker free from the severe appearance change in the kth frame.

Tracking Position: To locate the target's position in the $(k+1)$th frame, we need to obtain the response map M_{k+1} generated

by $\hat{x}_{k+1}^d$ and $\hat{g}_k$

$$M_{k+1} = \mathcal{F}^{-1}\left(\sum_{d=1}^{D} \hat{x}_{k+1}^d \odot \hat{g}_k^d\right) \tag{13}$$

where $\mathcal{F}^{-1}$ represents the inverse DFT transformation, $\odot$ denotes elementwise operation. By searching the maximum value of M_{k+1}, we can get the target's position in the $(k+1)$th frame.

C. Judgment Mechanism Based on Response Variation

In the tracking phase, the degree of response variation can serve as a reflection of whether the tracking results are reliable. When the target is tracked stably, there is less fluctuation in the response map. However, when the target encounters occlusion, etc., the response will fluctuate intensely. To identify abnormal response fluctuation, we develop a novel judging mechanism by optimizing the average peak-to-correlation energy (APCE) criteria [39]. First, we calculate the fluctuation score S_k related to the response in the kth frame as follows:

$$\text{APCE} = \frac{|R_{\max}^k - R_{\min}^k|^2}{\text{mean}\left(\sum_{w,h}(R_{w,h}^k - R_{\min}^k)^2\right)}$$

$$S_k = \frac{\left|\frac{R_{\max}^k - R_{\text{ref}}}{R_{\max}^k}\right|}{\text{APCE}} \tag{14}$$

where k denotes the kth frame and $|\cdot|$ means to take the absolute value. $R_{\max}^k$, $R_{\min}^k$, and $R_{w,h}^k$ represent the maximum, minimum, and the w-row h-column element of the response in the kth frame. R_{ref} is the maximum of the response in the first frame. In addition, $\left|\frac{R_{\max}^k - R_{\text{ref}}}{R_{\max}^k}\right|$ in S_k can usually be simplified as $\left|1 - \frac{R_{\text{ref}}}{R_{\max}^k}\right|$.

For the response with sharp peak and few noise, $R_{\max}^k$ is close to R_{ref} and the value of APCE is large [39]. Accordingly, $\left|1 - \frac{R_{\text{ref}}}{R_{\max}^k}\right|$ becomes small, and then S_k gets small. For the response with no sharp peak and more noise, the value of APCE and $R_{\max}^k$ all become small, and the gap between $R_{\max}^k$ and R_{ref} becomes large. Thus, $\left|1 - \frac{R_{\text{ref}}}{R_{\max}^k}\right|$ correspondingly grows due to this large gap, and then S_k becomes large.

In summary, when S_k becomes smaller, the tracking process is more robust. Conversely, when S_k becomes larger, the tracking result becomes unreliable. To utilize the continuous variation of S_k in the tracking sequence, we formulate L_k based on S_k, to represent the degree of response variation

$$L_k = \frac{S_k}{\text{mean}\left(\sum_{f=1}^{k-1} S_f\right)} \tag{15}$$

where $\text{mean}(\cdot)$ represents the operation for finding the average value. Compared with the value of APCE, the varying range of L_k is effectively narrowed, which reflects the trend of response variation, as shown in Fig. 2.

To dynamically adjust γ and μ, we take advantage of the variation of L_k. That is, when the target is tracked steadily, L_k remains around a small value. At this time, γ and μ can maintain appropriate values close to $\tilde{\gamma}$ and $\tilde{\mu}$ for keeping the response and

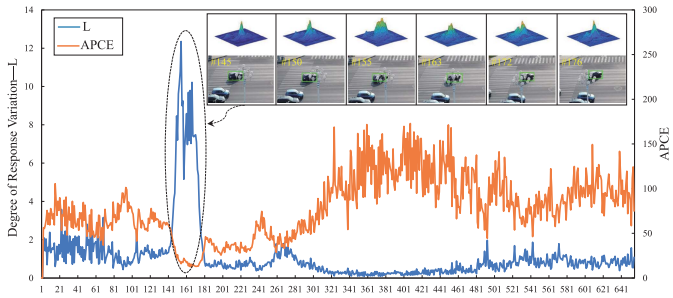

Fig. 2. Visualization of the degree of response variation L and the variation of APCE. The first row in the upper right box lists response maps of given frames. The second row in the upper right box shows the target tracking state in the corresponding image frame.

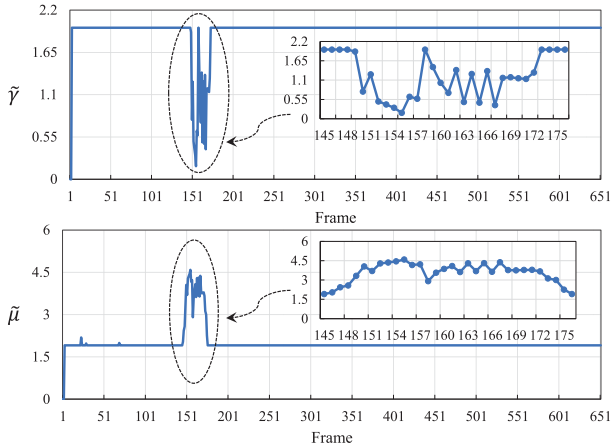

Fig. 3. Visualization of the aberrance-repressed regularization parameter $\tilde{\gamma}$ and the temporal regularization parameter $\tilde{\mu}$.

the filter in the current frame consistent with their states in the previous frame. On the contrary, when the target encounters occlusion scenarios, γ and μ can be decreased and increased, respectively, to enhance the strength of temporal regularization, which assists in restraining the sudden change of the filter. Thus, we exploit $\tilde{\gamma}$ and $\tilde{\mu}$ as the guiding value of γ and μ. Here, $\tilde{\gamma}$ and $\tilde{\mu}$ are defined as follows:

$$\tilde{\gamma} = \min(\gamma_{\text{ref}}, \exp(\nu * L_k) * \gamma_0)$$

$$\tilde{\mu} = \max\left(\mu_{\text{ref}}, \frac{1}{(\exp(\nu * L_k) + \mu_0)}\right) \tag{16}$$

where γ_{ref} and μ_{ref} denote fixed reference parameters, which control the variation magnitude of $\tilde{\gamma}$ and $\tilde{\mu}$. γ_{ref} limits the maximum range of $\tilde{\gamma}$. μ_{ref} limits the minimum range of $\tilde{\mu}$. $\exp(.)$ represents the exponential function with a natural constant e. In addition, ν, γ_0, and μ_0 are fixed parameters. Generally, the trend of $\exp(\nu * L_k)$ is opposite to L_k. To avoid $\exp(\nu * L_k)$ being too small, we use γ_0 as an amplified factor. Contrary to γ_0, μ_0 is a bias factor that avoids $\frac{1}{\exp(\nu * L_k)}$ being too large. The details of the above parameters can be found in Section IV-B. $\min(A, B)$ means obtaining the minimum value between A and B. $\max(A, B)$ means obtaining the maximum value between A and B.

Figs. 2 and 3 illustrate the variations of L, $\tilde{\gamma}$, and $\tilde{\mu}$ during the tracking phase. When the target encounters occlusion from the #145 frame to the #176 frame, the value of L deviates from the normal change level in the nonocclusion period and varies due to the response fluctuation. Simultaneously, $\tilde{\mu}$ and $\tilde{\gamma}$ are dynamically enlarging and reducing, aiming to maintain the continuity of the filter and diminish the constraints on the current and previous responses. Instead, when the target is tracked stably without occlusion, L varies in a relatively small range while $\tilde{\gamma}$ and $\tilde{\mu}$ are mostly equal to the reference values, i.e., γ_{ref} and μ_{ref}, respectively. Note that the values of $\tilde{\gamma}$ and $\tilde{\mu}$ are zero by default in the first frame, which can be seen in the plots of $\tilde{\gamma}$ and $\tilde{\mu}$ in Fig. 3. In other words, the aberrance-repressed and temporal regularizations do not work in the first frame.

D. Filter and Appearance Model Updating

Since the target inevitably encounters several situations (e.g., severe appearance change or occlusion) during the tracking phase, selecting an appropriate time to update the filter and appearance model is essential. Considering that the above situations generally cause abnormal response fluctuation, L_k is also regarded as a judgment standard for the filter and appearance model updating. Here, we introduce a threshold θ to determine when to update. When $L_k \leq \theta$ is satisfied, we allow to update. On the contrary, we stop updating when the following formulation is met:

$$L_k > \theta \tag{17}$$

where θ is a fixed parameter. It is worth noting that although we can determine when to update the filter and appearance model by setting the threshold θ, we should also consider when to exit from the nonupdating situation and start updating again. Here, a simple method is adopted. That is, if the values of $L_k(k, k-1, \ldots, k-10)$ are less than θ_L, we deem that the abnormal response fluctuation has disappeared. At this time, the proposed DARTCF exits from the nonupdating situation, and the filter and appearance model can be allowed to update. $(k, k-1, \ldots, k-10)$ represents the frame number in a continuous image sequence. $L_k(k, k-1, \ldots, k-10)$ denotes the L value from the kth frame to the $(k-1)$th frame. θ_L is a fixed threshold selected empirically.

Note that the optimization process for solving the filter $\hat{g}_k$ in (4) and the appearance model $\hat{x}_k^{\text{model}}$ in (12) are not performed during $L_k > \theta$.

In this work, the tracking pipeline of the proposed DARTCF is illustrated in Algorithm 1.

Algorithm 1: Proposed tracker (DARTCF).

Input: Image sequences captured by onboard cameras and initial state information about the target in the first frame.

Output: Predict the target location in the $k > 1$ frame.

for frame $k = 1$ to end **do**
  if frame $k > 1$ **then**
    Extract the feature $\hat{x}_k$ from the searching region.
    Calculate M_k with $\hat{x}_k$ and $\hat{g}_{k-1}$ in (13).
    Find the position of peak value in M_k and Set it as the target's position.
    Calculate parameters S_k, L_k, $\tilde{\gamma}$, and $\tilde{\mu}$.
    if $L_k > \theta$ and $U_{\text{flag}} == 1$ **then**
      Set $U_{\text{flag}} = 0$.
    end if
    if $U_{\text{flag}} == 0$ **then**
      if $L_k(k, k-1, \ldots, k-10) < \theta_L$ **then**
        Set $U_{\text{flag}} = 1$.
      end if
    end if
    if $U_{flag} == 0$ **then**
      Remain the correlation filter $\hat{g}_k$ and appearance model $\hat{x}_k^{\text{model}}$.
    else
      Update the appearance model $\hat{x}_k^{\text{model}}$ in (12).
      Learn the correlation filter $\hat{g}_k$ in (8).
    end if
  else
    Extract the feature $\hat{x}_k$ of the searching region in the first frame.
    Set the appearance model $\hat{x}_k^{\text{model}}$ equal to $\hat{x}_k$ and Set the update flag $U_{\text{flag}} = 1$.
    Learn the correlation filter $\hat{g}_k$.
  end if
end for

IV. Experiments

In this work, we have evaluated the proposed DARTCF on UAV123@10fps [19], UAVDT [20], and VisDrone2018 [21] benchmarks. In addition, we have compared experimental results of DARTCF with other total 25 state-of-the-art trackers proposed in recent years, i.e., 13 handcrafted features-based trackers (DR2Track [40], AutoTrack [37], AMCF [41], ARCF [16], STRCF [17], KCC [42], ECO _ HC [43], CSR _ DCF [27], STAPLE _ CA [14], BACF [26], fDSST [44], KCF [23], and SAMF [45]) and 12 deep features-based trackers (LUDT [46], LUDT+ [46], fECO [47], fDeepSTRCF [47], KAOT [48], ASRCF [28], UDT [49], UDT+ [49], TRACA [50], IBCCF [51], DSiam [52] and SiamFC [53]).

A. Evaluation Metrics

In this work, the one-pass evaluation protocol [19] is employed to evaluate the tracking performance of all trackers on three UAV benchmarks. Simultaneously, we use two metrics, i.e., precision rate and success rate, as the evaluation criteria. The precision rate is obtained by measuring the center location error (CLE) pixels between the predicted object box and the ground truth box. The precision rate plot illustrates the ratio of frames, in which the CLE is below a given threshold varied from 0 to 50 pixels, on the whole sequence frames. The success rate is measured by calculating the intersection over union (IOU) between the predicted object box and the ground truth box. The success rate plot illustrates the ratio of frames, in which the IOU is larger than a given threshold varied from 0 to 1, on the whole

sequence frames. To visually express the tracking results, the default ranking protocols are adopted. That is, the ratio of frames whose CLE is within 20 pixels ranks trackers in the perspective of precision rate evaluation and the area under the curve (AUC) with the condition of IOU ≥ 0.5 ranks trackers in the perspective of success rate evaluation.

B. Implementation Details

The proposed tracker DARTCF adopts handcrafted features [31-channel HOG, 10-channel color names (CN), and single-channel GrayScale]. All experiments are implemented on MATLAB R2019a with an Intel i7-10875H CPU (2.3 GHz), 16 G RAM, and a single Nvidia RTX2060 GPU. As for the parameters, we set $\lambda = 0.01$, $\gamma_{\text{ref}} = 1.97$, and $\mu_{\text{ref}} = 1.91$. The parameter ν is set to -0.39. The amplified factor γ_0 and bias factor μ_0 are 21 and 0.21, respectively. The fixed learning rate δ is 0.0229. The number of ADMM iterations is set to 3. Following the settings of AutoTrack [37], the initial penalty factor η is 1, the maximum penalty factor $\eta_{\max}$, and the scale factor ϕ are 10 000 and 10, respectively. The threshold θ and θ_L are set to 20 and 1.5, respectively. The source code is available at https://github.com/YanLiVision/DARTCF.

The operation $\exp(\nu * L)$ with an empirical parameter ν aims to map L to the interval (0,1). Also, $\exp(\nu * L)$ has a tendency opposite to the change of L. Then, by setting the amplified factor $\gamma_0 = 21$, we can obtain $0 < \exp(\nu * L) * \gamma_0 < 21$. With the function $\min(\gamma_{\text{ref}}, .)$, the value of $\tilde{\gamma}$ varies from $(\exp(\nu * L) * \gamma_0)$ to γ_{ref} during the severe response fluctuation. The γ_{ref} is the upper limit of $\tilde{\gamma}$. In addition, by setting the bias factor $\mu_0 = 0.21$, we can obtain $0.8 < \frac{1}{(\exp(\nu * L_k) + \mu_0)} < 5$. With the function $\max(\mu_{\text{ref}}, .)$, the value of $\tilde{\mu}$ varies from μ_{ref} to $(\frac{1}{(\exp(\nu * L_k) + \mu_0)})$ during the severe response fluctuation. The μ_{ref} is the lower limit of $\tilde{\mu}$. Note that when the target is tracked stably without the severe response fluctuation, L is usually small, and the values of $\tilde{\gamma}$ and $\tilde{\mu}$ are usually γ_{ref} and μ_{ref}, respectively.

C. Evaluation Datasets

The experiments utilize three challenging UAV benchmarks to evaluate the tracking performance of the proposed DARTCF.

UAV123@10fps: The UAV dataset in [19] includes 123 video sequences and over 110-K image frames captured at 30 FPS, which is then downsampled to a 10-FPS (UAV123@10fps) benchmark. Since the camera platform is located on the UAV, UAV123@10fps contains several attributes (e.g., aspect ratio change, camera motion, similar object, background clutter, partial/full occlusion, and viewpoint change) that increase the difficulty for target tracking.

UAVDT: A public UAV detection and tracking (UAVDT) [20] benchmark contains 50 video sequences captured by a UAV in different flying altitude for single object tracking, in which there are in total eight main attributes, including background clutter, large occlusion, illumination variation, object blur, scale variation, camera rotation, small object, and object rotation. Compared with [19] and [21], the interest targets of UAVDT are vehicles.

TABLE I
OVERALL PERFORMANCE OF DARTCF AND OTHER 13 STATE-OF-THE-ART HANDCRAFTED-BASED TRACKERS, WHICH ARE EVALUATED BY AVERAGE SUCCESS AND PRECISION RATES, AS WELL AS AVERAGE FPS (FRAMES PER SECOND) ON ALL UAV BENCHMARKS

Trackers	Venue	Avg. Succ.	Avg. Prec.	Avg. FPS
DARTCF	**Ours**	0.517	0.755	35.37
DR2Track [40]	21'EAAI	0.470	0.675	49.90
AutoTrack [37]	20'CVPR	0.503	0.731	36.35
AMCF [41]	20'IROS	0.478	0.686	43.22
ARCF [16]	19'ICCV	0.509	0.734	30.25
STRCF [17]	18'CVPR	0.480	0.680	29.65
KCC [42]	18'AAAI	0.434	0.660	48.41
ECO_HC [43]	17'CVPR	0.485	0.711	76.04
CSR_DCF [27]	17'CVPR	0.468	0.705	17.71
STAPLE_CA [14]	17'CVPR	0.458	0.695	53.22
BACF [26]	17'ICCV	0.474	0.686	46.85
fDSST [44]	16'TPAMI	0.427	0.633	187.42
KCF [23]	15'TPAMI	0.324	0.555	669.57
SAMF [45]	14'ECCV	0.399	0.609	15.64

The top three performances are highlighted by using red, green, and blue fonts, respectively.

VisDrone2018: The vision meets drone single-object-tracking (VisDrone2018) [21] testing set consists of 35 sequences with 29 367 frames under different weather and lighting conditions. In addition, VisDrone2018 contains a total of 12 attributes, e.g., background clutter, occlusion, fast motion, viewpoint change, and similar object, etc. The tracking objects in VisDrone2018 include pedestrians, vehicles, and animals.

D. Comparison With Handcrafted-Based Trackers

We have compared the proposed DARTCF with other 13 handcrafted-based trackers, i.e., DR2Track [40], AutoTrack [37], AMCF [41], ARCF [16], STRCF [17], KCC [42], ECO _ HC [43], CSR _ DCF [27], STAPLE _ CA [14], BACF [26], fDSST [44], KCF [23], and SAMF [45].

1) Overall Analysis: We have completed the overall analysis of the proposed DARTCF on all UAV benchmarks.

UAV123@10fps: In Fig. 4, our tracker DARTCF has attained the top precision and success rates that are 68.0% and 48.0%, respectively, and exceeds other 13 advanced handcrafted-based trackers.

UAVDT: As shown in Fig. 4, the tracking results of DARTCF is superior to other compared trackers. The precision and success rates of DARTCF exceed ARCF ranked second by 2.0% and 0.8%, respectively. Simultaneously, AutoTrack ranked third is 2.2% and 1.7% lower than DARTCF in precision and success rates.

VisDrone2018: In Fig. 4, DARTCF has surpassed other total 13 advanced trackers. Besides, DARTCF has 2.8% and 3.7% advantages over the tracker ARCF (0.797) and the tracker AutoTrack (0.788) in the precision rate, as well as advantages of 1.1% and 2.1% over them in the success rate.

Table I intuitively expresses the performance of DARTCF with the average success and precision rates. Besides, DARTCF attains the highest scores against other competitive trackers and also exceeds ARCF and AutoTrack by 0.8% and 1.4% in the success rate, 2.1% and 2.4% in the precision rate, respectively.

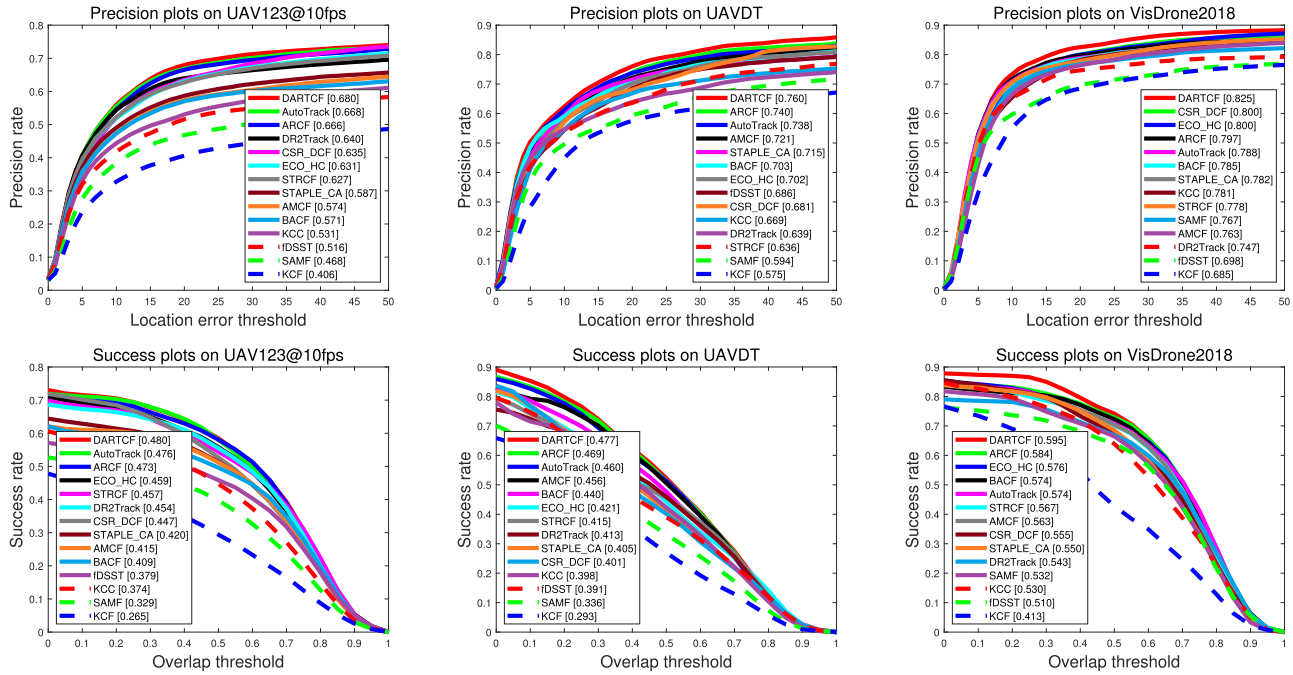

Fig. 4. Precision and success plots of DARTCF and other 13 state-of-the-art handcrafted-based trackers on UAV123@10fps, UAVDT, and VisDrone2018 benchmarks.

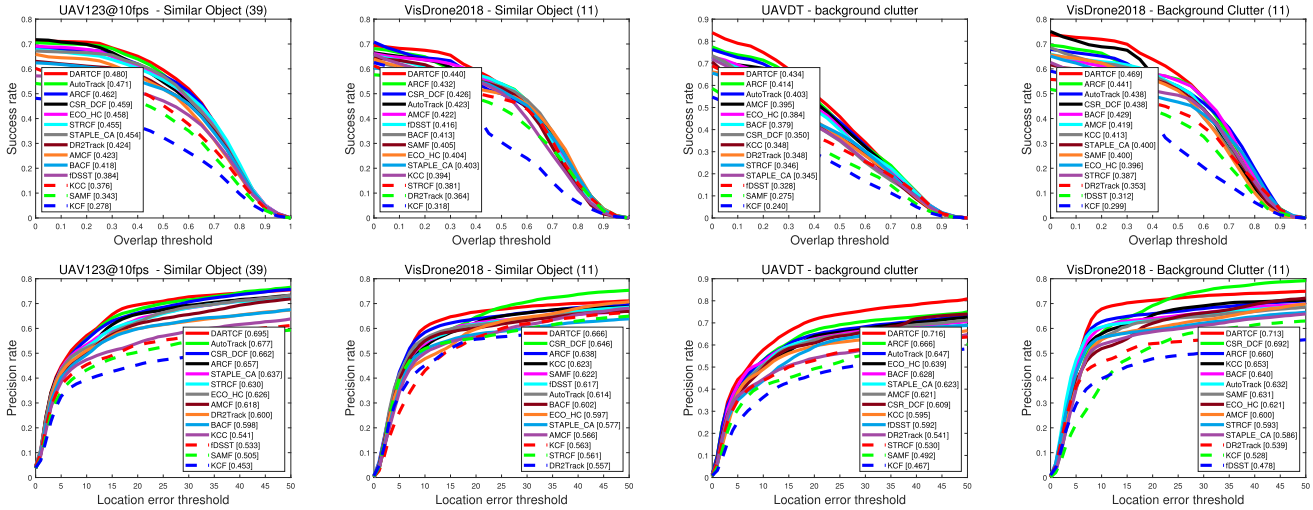

Fig. 5. Attribute analysis of DARTCF and other 13 handcrafted-based trackers in scenarios with similar object and background clutter.

2) Attribute Analysis: We have completed the attribute analysis of DARTCF and other 13 excellent trackers in scenarios, i.e., similar object, background clutter, and occlusion.

Fig. 5 shows the comparison results when all trackers work in scenarios that possess attributes of similar object and background clutter. Our tracker DARTCF has achieved the highest scores in the attribute of similar object on UAV123@10fps and VisDrone2018 benchmarks. Simultaneously, we have validated the performance of all trackers on the attribute with background clutter and DARTCF has improved the success rate by 2.0% and the precision rate by 5.0% compared with ARCF on UAVDT benchmark. Moreover, DARTCF ranks first on VisDrone2018 benchmark and outperforms ARCF by 2.8% in the success rate and 5.3% in the precision rate under the background clutter attribute.

In Fig. 6, we have also conducted the attribute analysis of all trackers in scenarios with occlusion. It shows that DARTCF has achieved the best performance compared with other trackers on all UAV benchmarks. Moreover, DARTCF has 1.5% (success rate) and 3.3% (precision rate) advantages over ARCF under the attribute of partial occlusion on UAV123@10fps benchmark. DARTCF exceeds ARCF by 2.9% and 7.7% in success and precision rates on UAVDT benchmark. On VisDrone2018 benchmark, DARTCF outperforms ARCF by 3.1% and 5.0% in success and precision rates under the attribute of partial occlusion. In summary, when encountering occlusion challenges,

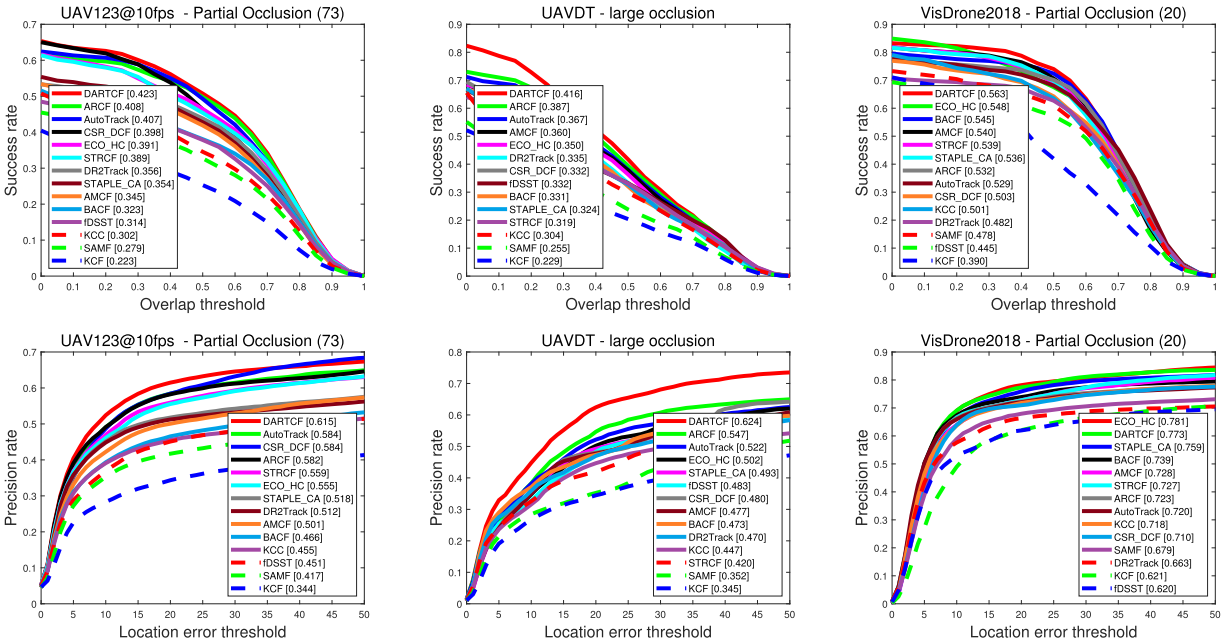

Fig. 6. Attribute analysis of our tracker DARTCF and other 13 handcrafted-based trackers in scenarios with occlusion.

Fig. 7. Visualization of qualitative comparisons between DARTCF and other four advanced trackers. From first to six rows, the image sequences are group2_2_1 and truck4_2_1 in UAV123@10fps [19], S0103 and S1606 in UAVDT [20], uav0000093_00000_s _1 and uav0000294_00000_ s_1 in VisDrone2018 [21].

DARTCF can achieve more stable tracking than other compared trackers.

3) Qualitative Evaluation: For the sake of qualitative evaluation, we have compared our tracker DARTCF with other four advanced trackers, including AutoTrack [37], ARCF [16], ECO _ HC [43], and BACF [26], on six tracking sequences.

In Fig. 7, our tracker is marked with a red bounding box, which has an excellent performance in scenarios with similar object, background clutter, and occlusion attributes. For example, in the group2_2_1 (first row) sequence, the target is vulnerable to similar distractors and it is partially occluded by plants in the #169 frame. But, in #181 and #289 frames, compared

with ARCF, ECO _ HC, and BACF, the target is still captured by our tracker. In the truck4_2_1 (second row) sequence, the target with small size is partially occluded by a tree in the #39 frame. But, in the following frames, our tracker can remain steadily tracking. In the S0103 (third row) sequence, our tracker and other trackers have accurately located the target in the #1 frame under the scenario with background clutter and occlusion. However, in the #14 frame, only our tracker has tracked the target stably instead of other compared trackers. Besides, despite the viewpoint change occurring in #130 and #307 frames, our tracker also captures the target stably. In the S1606 (fourth row) and uav0000294_00000_s _1 (sixth row) sequences, both targets are partially occluded in #155 and #39 frames, respectively, and surrounded by distractors, our tracker keeps accurate tracking until targets completely return to the UAV's view. In the uav0000093_00000_ s_1 (fifth row) sequence, the small target is in a scenario containing many similar targets. At the beginning, the target is distinguished by all five trackers in the #176 frame. However, when the tracked target is clustered with other similar targets, it is difficult to acquire the real target's position. In #556, #829, and #1460 frames, ECO _ HC, AutoTrack, and BACF have gradually lost the target. Until the #1609 frame, only DARTCF and ARCF can locate the target. In total, it has demonstrated that DARTCF can attain more advanced performance in scenarios with similar object, background clutter, and occlusion attributes.

4) Tracking Speed Analysis: Real-time tracking based on UAV platforms requires the tracker's speed to exceed 30 FPS. As shown in Table I, the proposed DARTCF has achieved an average speed of 35.37 FPS on all UAV benchmarks, which satisfies the real-time requirement of UAV tracking. Moreover, compared with the top three fast trackers (i.e., KCF [23], fDSST [44], and ECO _ HC [43]), the proposed DARTCF has significant advantages in precision and success rates. Furthermore, in actual UAV applications, tracking methods are typically deployed in embedded devices, e.g., multiprocessing system-on-chips. Thus, by rebuilding tracking codes and utilizing multithreading technology, the proposed DARTCF can achieve acceleration in embedded devices and make real-time UAV tracking tasks to be possible.

E. Comparison With Deep-Based Trackers

In this section, we have compared our tracker with other 12 deep-features-based trackers, i.e., LUDT [46], LUDT+ [46], fECO [47], fDeepSTRCF [47], KAOT [48], ASRCF [28], UDT [49], UDT+ [49], TRACA [50], IBCCF [51], DSiam [52], and SiamFC [53]. As shown in Table II, the proposed DARTCF reaches the best tracking performance and outperforms ASRCF by 1.0% and 1.4% in success and precision rates, respectively. As for the tracking speed, although DARTCF is not superior to several deep trackers, including TRACA, LUDT, UDT, SiamFC, and LUDT, which rely on a single GPU, DARTCF has still reached 35.37 FPS with a CPU and met the real-time requirement of UAV tracking.

TABLE II
OVERALL PERFORMANCE OF DARTCF AND OTHER 12 ADVANCED DEEP-BASED TRACKERS (I.E., LUDT [46], LUDT+ [46], FECO [47], FDEEPSTRCF [47], KAOT [48], ASRCF [28], UDT [49], UDT+ [49], TRACA [50], IBCCF [51], DSIAM [52], AND SIAMFC [53]), WHICH ARE EVALUATED BY AVERAGE SUCCESS AND PRECISION RATES ON ALL UAV BECHMARKS

Trackers	Venue	Avg. Succ.	Avg. Prec.	Avg. FPS
DARTCF	**Ours**	0.517	0.755	35.37
LUDT [46]	21'IJCV	0.466	0.668	80.04*
LUDT+ [46]	21'IJCV	0.486	0.719	47.33*
fECO [47]	20'TIP	0.495	0.715	31.32*
fDeepSTRCF [47]	20'TIP	0.509	0.721	20.49*
KAOT [48]	20'ICRA	0.433	0.712	11.99*
ASRCF [28]	19'CVPR	0.507	0.741	22.02*
UDT [49]	19'CVPR	0.473	0.682	78.47*
UDT+ [49]	19'CVPR	0.492	0.725	47.33*
TRACA [50]	18'CVPR	0.451	0.652	88.03*
IBCCF [51]	17'ICCV	0.479	0.677	3.82*
DSiam [52]	17'ICCV	0.470	0.688	24.26*
SiamFC [53]	16'ECCV	0.485	0.704	65.42*

The top three performances are highlighted by using red, green, and blue fonts, respectively. The subscript * means GPU speed.

TABLE III
ABLATION STUDY OF THE PROPOSED TRACKER ON VISDRONE2018 [21] BENCHMARK

Tracker	Prec.	Succ.	FPS
Baseline	0.777	0.566	36.2
Baseline + AR	0.793	0.578	32.5
Baseline + AR + U	0.794	0.578	32.6
Baseline + AR + TE	0.817	0.589	31.1
Baseline + FAR + FTE	0.807	0.584	31.2
Baseline + FAR + FTE + U	0.808	0.585	31.5
Baseline + AR + TE + U (Final)	0.825	0.595	31.7

F. Ablation Study

In this section, aiming to complete the ablation analysis, we have constructed seven trackers, i.e., Baseline, Baseline+AR, Baseline+AR+U, Baseline+AR+TE, Baseline+FAR+FTE, Baseline+FAR+FTE+U, and Baseline+AR+TE+U, which are evaluated on the VisDrone2018 [21] benchmark. The Baseline tracker is DARTCF without the dynamic aberrance-repressed and temporal regularizations. In other words, the Baseline tracker is only equipped with HOG, CN, and GrayScales features and utilizes the scale pyramid with 33 scales based on BACF. The tracker named Baseline+AR integrates the dynamic aberrance-repressed regularization based on the Baseline tracker. The tracker named Baseline+AR+TE introduces the dynamic temproal regularization into the tracker (Baseline+AR). The trackers (i.e., Baseline+AR+U and Baseline+AR+TE+U) add the update mechanism in Section III-D on the basis of Baseline+AR and Baseline+AR+TE trackers, respectively. In addition, the Baseline+FAR+FTE tracker integrates fixed parameters-based aberrance-repressed and temporal regularizations into the Baseline tracker, which is different from the dynamic parameters-based Baseline+AR+TE tracker. The Baseline+FAR+FTE+U tracker equips with the update mechanism based on the Baseline+FAR+FTE tracker.

In Table III, the Baseline tracker achieves 77.7% precision rate and 56.6% success rate. Compared with the Baseline tracker,

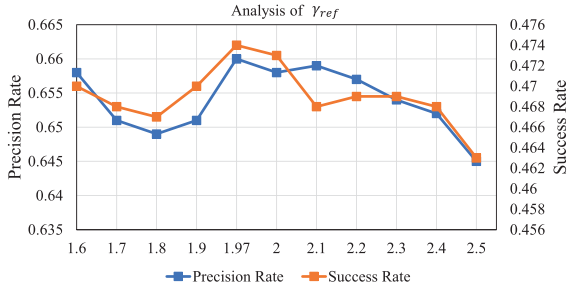

Fig. 8. Analysis of parameter γ_{ref} on UAV123@10fps benchmark.

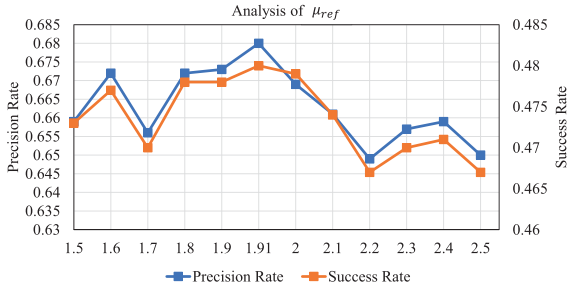

Fig. 9. Analysis of parameter μ_{ref} on UAV123@10fps benchmark.

the Baseline+AR tracker improves 1.6% and 1.2% in precision and success rates, respectively. Compared with Baseline+AR, the Baseline+AR+TE tracker increases the precision and success rates by 2.4% and 1.1%, respectively. As for the update mechanism in Section III-D, the Baseline+AR+U and Baseline+AR+TE+U trackers have been improved slightly compared with Baseline+AR and Baseline+AR+TE. The reason is that the setting of threshold θ makes the proposed tracker unable to update the filter and appearance model easily and frequently, which aims to track stably. Furthermore, compared with fixed parameters-based trackers (i.e., Baseline+FAR+FTE and Baseline+FAR+FTE+U), the Baseline+AR+TE tracker surpasses Baseline+FAR+FTE by 1% and 0.5% in precision and success rates, respectively. The Baseline+AR+TE+U tracker outperforms Baseline+FAR+FTE+U by 1.7% and 1.0% in precision and success rates, respectively. That is, the trackers equipped with dynamic-parameters regularizations (i.e., Baseline+AR+TE and Baseline+AR+TE+U) have achieved a superior tracking performance against fixed parameters-based trackers. To sum up, the results of this ablation study indicates that dynamic aberrance-repressed and temporal regularizations based on response variation can improve the tracking performance.

G. Parametric Sensitivity

γ_{ref} and μ_{ref} are two significant reference parameters in this work. These parameters mainly affect the tracking performance of DARTCF in the situation without severe response fluctuations. As shown in Fig. 3, when DARTCF does not encounter occlusion, parameters $\tilde{\gamma}$ and $\tilde{\mu}$ are usually equal to γ_{ref} and μ_{ref}, respectively. In addition, γ_{ref} and μ_{ref} also serve as the upper limit of $\tilde{\gamma}$ and the lower limit of $\tilde{\mu}$ when DARTCF is in a situation with severe response fluctuations. Here, to analyze the impact of γ_{ref} and μ_{ref} on the tracking performance, we evaluate one of them on UAV123@10fps benchmark while fixing the other one.

1) Analysis of γ_{ref}: The reference parameter γ_{ref} aims to constrain the maximum range of $\tilde{\gamma}$. That is, the value of $\tilde{\gamma}$ does not exceed γ_{ref}. To obtain the optimal value of γ_{ref}, we change γ_{ref} from 1.6 to 2.5 while ensuring that the value of μ is set to 0, which corresponds to the Baseline+AR+U tracker. As shown in Fig. 8, when γ_{ref} reaches 1.97, the Baseline+AR+U tracker has achieved the best success rate (0.474) and precision rate (0.66).

2) Analysis of μ_{ref}: The reference parameter μ_{ref} restricts the minimum value of $\tilde{\mu}$. In other words, the value of $\tilde{\mu}$ is not lower than μ_{ref}. Here, considering that the Baseline+AR+U tracker has achieved the best performance by setting $\gamma_{\text{ref}} = 1.97$, we vary μ_{ref} from 1.5 to 2.5 for selecting the optimal value of μ_{ref} based on the Baseline+AR+TE+U tracker by fixing $\gamma_{\text{ref}} = 1.97$. As shown in Fig. 9, the Baseline+AR+TE+U tracker obtains the best performance by setting $\mu_{\text{ref}} = 1.91$ and the precision and success rates reach 0.68 and 0.48, respectively.

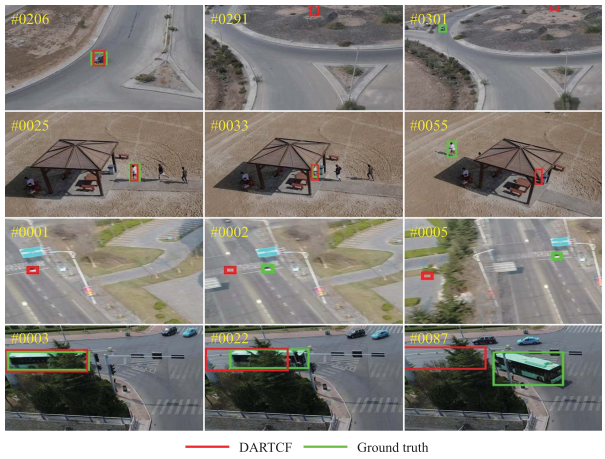

Fig. 10. Failure cases of the proposed DARTCF. The visualized sequences (from first to four rows) are car14 and group2_1_1 in UAV123@10fps [19], as well as S1605 and S1501 in UAVDT [20]. The red and green boxes show the tracking status of DARTCF and the ground truth, respectively.

H. Failure Cases and Limitation

1) Failure Cases: As shown in Fig. 10, there are four image sequences (car14 and group2_1_1 in UAV123@10fps [19], as well as S1605 and S1501 in UAVDT [20]) from first to four rows, which visualizes four situations where the proposed DARTCF fails to track the target.

In the car14 sequence, the target is not within the UAV's view and its tracking box has drifted in the # 291 frame. When the target reappears in the UAV's view in the #301 frame, DARTCF cannot recapture it. In the group2_1_1 sequence, there is a situation where the target is about to be obstructed in the #33 frame. Then, when the target reappears in the #55 frame, DARTCF fails to locate the target accurately. Thus, the tracking results that exist in car14 and group2_1_1 sequences show that the proposed DARTCF cannot handle the challenge, i.e., out of view, and does not possess the ability of recapturing target. Furthermore, in the S1605 sequence, the target encounters serious camera motion in the #2 frame, which causes the tracking box drifting

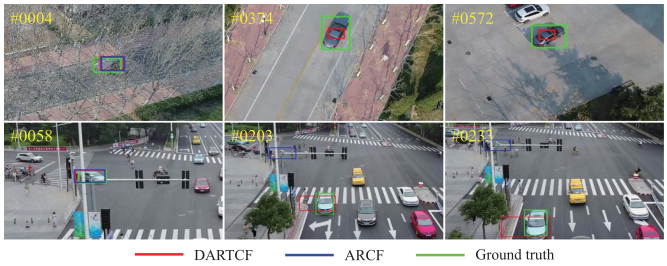

Fig. 11. Limitation of the proposed DARTCF. The visualized sequences (from first to second rows) are S0103 in UAVDT [20], and uav0000294_00000_s in VisDrone2018 [21].

rapidly. In the S1501 sequence, partial occlusion has occurred in the initial tracking phase, as shown in the #3 frame. Thus, the discrimination ability of the initial filter declines and the target is inaccurately tracked in the following frames.

2) Limitation: Through extensive experimental evaluation in Section IV-D, the proposed DARTCF has achieved an excellent tracking performance. However, compared with the obvious improvement in precision rate, the success rate of the proposed DARTCF is not significantly higher than that of ARCF. After analyzing the tracking state of DARTCF, we deem that this situation is caused by the limitation of our method, i.e., inaccurate estimation of target scale.

As shown in Fig. 11, we select two sequences (S0103 from UAVDT and uav0000294_00000_s from VisDrone2018) to validate this limitation. In S0103 sequence, the IOU between DARTCF and ground truth is greater than 0.5 in the initial frame, but in the subsequent image frames, such as #374 and #572 frames, the IOU between DARTCF and ground truth markedly decreases and is less than 0.5, which makes the success rate value of DARTCF unable to increase effectively. As for the precision rate, since the CLE between DARTCF and ground truth generally remains within 20 pixels, and the ARCF cannot track the target stably, the tracking accuracy of DARTCF has got a notable promotion compared with ARCF. Similarly, in uav0000294_00000_s sequence, the IOU between DARTCF and ground truth has not achieved a satisfactory performance from the #203 frame to the #233 frame, which leads to a slow growth in success rate compared with ARCF that has failed to track the target. To sum up, although the proposed DARTCF can significantly improve the tracking accuracy (i.e., precision rate), its performance on estimating the target scale still has the limitation, which causes a lack of obvious improvement in the success rate.

V. Conclusion

In this article, a novel aberrance-repressed temporal CF method is developed, which can dynamically adjust the strengths of both regularizations. In addition, a novel response variation-based judgment mechanism is exploited, which participates in the parameter tuning process of regularizations and determines when to update the filter and appearance model. Furthermore, the proposed DARTCF and other total 25 excellent trackers have undergone validation experiments on three UAV benchmarks. Experimental results have proved that the proposed DARTCF has the competitive performance compared with other 25 trackers. Moreover, the running speed of DARTCF can reach 35.37 FPS on a single CPU, which satisfies the real-time tracking standard.

Nevertheless, the proposed DARTCF still has shortcomings. When encountering the scenarios (i.e., out of view, serious camera motion, and polluted appearance model in initial frames), the proposed DARTCF cannot capture the target accurately in the follow-up tracking process. In addition, the proposed DARTCF has the limitation in the aspect of accurately estimating target scale. Thus, in the future, a real-time tracker with a robust redetection mechanism, a strong appearance expression ability, and an effective target scale estimation will be the research focus of our work.

References

[1] X. Xu et al., "STN-track: Multiobject tracking of unmanned aerial vehicles by swin transformer neck and new data association method," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 8734–8743, 2022.

[2] Y. Han, H. Liu, Y. Wang, and C. Liu, "A comprehensive review for typical applications based upon unmanned aerial vehicle platform," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 9654–9666, 2022.

[3] Y. Li, H. Zhang, Y. Yang, H. Liu, and D. Yuan, "Ristrack: Learning response interference suppression correlation filters for UAV tracking," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, 2023, Art. no. 8000705.

[4] W. W. Immerzeel et al., "High-resolution monitoring of himalayan glacier dynamics using unmanned aerial vehicles," *Remote Sens. Environ.*, vol. 150, pp. 93–103, 2014.

[5] A. Hamedpour and F. Farnood Ahmadi, "Recognition and tracking of moving objects in the images captured by UAV intelligently in Earth observation operations," *Arabian J. Geosci.*, vol. 11, pp. 1–8, 2018.

[6] C. Yuan, Z. Liu, and Y. Zhang, "UAV-based forest fire detection and tracking using image processing techniques," in *Proc. Int. Conf. Unmanned Aircr. Syst.*, 2015, pp. 639–643.

[7] H. Shi, Z. Fang, Y. Wang, and L. Chen, "An adaptive sample assignment strategy based on feature enhancement for ship detection in SAR images," *Remote Sens.*, vol. 14, no. 9, 2022, Art. no. 2238.

[8] H. Shi, B. Chai, Y. Wang, and L. Chen, "A local-sparse-information-aggregation transformer with explicit contour guidance for SAR ship detection," *Remote Sens.*, vol. 14, no. 20, 2022, Art. no. 5247.

[9] H. Shi, C. He, J. Li, L. Chen, and Y. Wang, "An improved anchor-free SAR ship detection algorithm based on brain-inspired attention mechanism," *Front. Neurosci.*, vol. 16, 2022, Art. no. 1074706.

[10] C. Deng, S. He, Y. Han, and B. Zhao, "Learning dynamic spatial-temporal regularization for UAV object tracking," *IEEE Signal Process. Lett.*, vol. 28, pp. 1230–1234, 2021.

[11] Y. Han, C. Deng, B. Zhao, and D. Tao, "State-aware anti-drift object tracking," *IEEE Trans. Image Process.*, vol. 28, no. 8, pp. 4075–4086, Aug. 2019.

[12] Y. Han, H. Wang, Z. Zhang, and W. Wang, "Boundary-aware vehicle tracking upon UAV," *Electron. Lett.*, vol. 56, no. 17, pp. 873–876, 2020.

[13] W. Xing, H. Zhang, Y. Wu, Y. Li, and D. Yuan, "Redefined target sample-based background-aware correlation filters for object tracking," *Appl. Intell.*, vol. 53, pp. 11120–11141, 2023.

[14] M. Mueller, N. Smith, and B. Ghanem, "Context-aware correlation filter tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1387–1395.

[15] W. Zhang, L. Jiao, Y. Li, and J. Liu, "Sparse learning-based correlation filter for robust tracking," *IEEE Trans. Image Process.*, vol. 30, pp. 878–891, 2021.

[16] Z. Huang, C. Fu, Y. Li, F. Lin, and P. Lu, "Learning aberrance repressed correlation filters for real-time UAV tracking," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 2891–2900.

[17] F. Li, C. Tian, W. Zuo, L. Zhang, and M.-H. Yang, "Learning spatial-temporal regularized correlation filters for visual tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 4904–4913.

[18] S. Boyd et al., "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations Trends Mach. Learn.*, vol. 3, no. 1, pp. 1–122, 2011.
[19] M. Mueller, N. Smith, and B. Ghanem, "A benchmark and simulator for UAV tracking," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 445–461.
[20] D. Du et al., "The unmanned aerial vehicle benchmark: Object detection and tracking," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 375–391.
[21] L. Wen et al., "Visdrone-sot2018: The vision meets drone single-object tracking challenge results," in *Proc. Eur. Conf. Comput. Vis. Workshops*, 2018, pp. 469–495.
[22] D. S. Bolme, J. R. Beveridge, B. A. Draper, and Y. M. Lui, "Visual object tracking using adaptive correlation filters," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2010, pp. 2544–2550.
[23] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista, "High-speed tracking with kernelized correlation filters," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 3, pp. 583–596, Mar. 2015.
[24] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2005, pp. 886–893.
[25] M. Danelljan, G. Häger, F. S. Khan, and M. Felsberg, "Learning spatially regularized correlation filters for visual tracking," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 4310–4318.
[26] H. K. Galoogahi, A. Fagg, and S. Lucey, "Learning background-aware correlation filters for visual tracking," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 1144–1152.
[27] A. Lukezic, T. Vojir, L. C. Zajc, J. Matas, and M. Kristan, "Discriminative correlation filter with channel and spatial reliability," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4847–4856.
[28] K. Dai, D. Wang, H. Lu, C. Sun, and J. Li, "Visual tracking via adaptive spatially-regularized correlation filters," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 4670–4679.
[29] W. Xing, H. Zhang, H. Chen, Y. Yang, and D. Yuan, "Feature adaptation-based multipeak-redetection spatial-aware correlation filter for object tracking," *Neurocomputing*, vol. 488, pp. 299–314, 2022.
[30] Y. Han, C. Deng, Z. Zhang, J. Li, and B. Zhao, "Adaptive feature representation for visual tracking," in *Proc. IEEE Int. Conf. Inf. Process.*, 2017, pp. 1867–1870.
[31] D. Elayaperumal and Y. H. Joo, "Aberrance suppressed spatio-temporal correlation filters for visual object tracking," *Pattern Recognit.*, vol. 115, 2021, Art. no. 107922.
[32] L. Xu, P. Kim, M. Wang, J. Pan, X. Yang, and M. Gao, "Spatio-temporal joint aberrance suppressed correlation filter for visual tracking," *Complex Intell. Syst.*, vol. 8, pp. 3765–3777, 2022.
[33] X. Wang and B. Fan, "Learning aberrance repressed and temporal regularized correlation filters for visual tracking," in *Proc. Chin. Automat. Congr.*, 2020, pp. 2604–2609.
[34] Y. Ji, J. He, X. Sun, Y. Bai, Z. Wei, and K. H. B. Ghazali, "Learning augmented memory joint aberrance repressed correlation filters for visual tracking," *Symmetry*, vol. 14, no. 8, 2022, Art. no. 1502.
[35] T. Li, F. Ding, and W. Yang, "UAV object tracking by background cues and aberrances response suppression mechanism," *Neural Comput. Appl.*, vol. 33, no. 8, pp. 3347–3361, 2021.
[36] Y. Han, C. Deng, B. Zhao, and B. Zhao, "Spatial-temporal context-aware tracking," *IEEE Signal Process. Lett.*, vol. 26, no. 3, pp. 500–504, Mar. 2019.
[37] Y. Li, C. Fu, F. Ding, Z. Huang, and G. Lu, "Autotrack: Towards high-performance visual tracking for UAV with automatic spatio-temporal regularization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 11920–11929.
[38] H. K. Galoogahi, T. Sim, and S. Lucey, "Multi-channel correlation filters," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 3072–3079.
[39] M. Wang, Y. Liu, and Z. Huang, "Large margin object tracking with circulant feature maps," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4800–4808.
[40] C. Fu, F. Ding, Y. Li, J. Jin, and C. Feng, "Learning dynamic regression with automatic distractor repression for real-time UAV tracking," *Eng. Appl. Artif. Intell.*, vol. 98, 2021, Art. no. 104116.
[41] Y. Li, C. Fu, F. Ding, Z. Huang, and J. Pan, "Augmented memory for correlation filters in real-time UAV tracking," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2020, pp. 1559–1566.
[42] C. Wang, L. Zhang, L. Xie, and J. Yuan, "Kernel cross-correlator," in *Proc. AAAI Conf. Artif. Intell.*, 2018, pp. 4179–4186.
[43] M. Danelljan, G. Bhat, F. S. Khan, and M. Felsberg, "ECO: Efficient convolution operators for tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6931–6939.
[44] M. Danelljan, G. Häger, F. S. Khan, and M. Felsberg, "Discriminative scale space tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 8, pp. 1561–1575, Aug. 2017.
[45] Y. Li and J. Zhu, "A scale adaptive kernel correlation filter tracker with feature integration," in *Proc. Eur. Conf. Comput. Vis. Workshops*, 2014, pp. 254–265.
[46] N. Wang, W. Zhou, Y. Song, C. Ma, W. Liu, and H. Li, "Unsupervised deep representation learning for real-time tracking," *Int. J. Comput. Vis.*, vol. 129, no. 2, pp. 400–418, 2021.
[47] N. Wang, W. Zhou, Y. Song, C. Ma, and H. Li, "Real-time correlation tracking via joint model compression and transfer," *IEEE Trans. Image Process.*, vol. 29, pp. 6123–6135, 2020.
[48] Y. Li, C. Fu, Z. Huang, Y. Zhang, and J. Pan, "Keyfilter-aware real-time UAV object tracking," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2020, pp. 193–199.
[49] N. Wang, Y. Song, C. Ma, W. Zhou, W. Liu, and H. Li, "Unsupervised deep tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 1308–1317.
[50] J. Choi et al., "Context-aware deep feature compression for high-speed visual tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 479–488.
[51] F. Li, Y. Yao, P. Li, D. Zhang, W. Zuo, and M.-H. Yang, "Integrating boundary and center correlation filters for visual tracking with aspect ratio variation," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, 2017, pp. 2001–2009.
[52] Q. Guo, W. Feng, C. Zhou, R. Huang, L. Wan, and S. Wang, "Learning dynamic Siamese network for visual object tracking," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 1781–1789.
[53] L. Bertinetto, J. Valmadre, J. F. Henriques, A. Vedaldi, and P. H. Torr, "Fully-convolutional Siamese networks for object tracking," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 850–865.

Hong Zhang received the B.S. degree from the Hebei University of Technology, Tianjin, China, in 1988, the M.S. degree from the Harbin University of Science and Technology, Harbin, China, 1993, and the Ph.D. degree from the Beijing Institute of Technology, Beijing, China, in 2002, all in electrical engineering.

She was in the Department of Neurosurgery, University of Pittsburgh, Pittsburgh, PA, USA, as a Visiting Scholar, from 2007 to 2008. She is currently a Professor with the School of Astronautics, Beihang University, Beijing. Her research interests include target recognition, image restoration, image indexing, object detection, and stereovision.

Yan Li is currently working toward the Ph.D. degree in pattern recognition and intelligent system with the School of Astronautics, Beihang University, Beijing, China.

His research interests include visual tracking, object detection, and pattern recognition.

Yifan Yang received the M.S. degree in pattern recognition and intelligent system from the School of Automation and Electrical Engineering, University of Science and Technology Beijing, Beijing, China, in 2011, and the Ph.D. degree in pattern recognition and intelligent system from the School of Astronautics, Beihang University, Beijing, in 2018.

He is currently an Associate Professor with the Institute of Artificial Intelligence, Beihang University. His research interests include pattern recognition and intelligent system, object tracking and detection, real-time image processing, and embedded systems.

Yachun Feng received the M.S.E. degree in pattern recognition and intelligent system in 2016, from Beihang University, Beijing, China, where he is currently working toward the Ph.D. degree in mechanics with the School of Mechanical Engineering and Automation.

His research interests include computer vision and machine learning and in particular on target tracking.

Yawei Li received the B.S. degree in automation from Xidian University, Xi'an, China, in 2013, and the Ph.D degree in pattern recognition and intelligent system from the School of Astronautics, Beihang University, Beijing, China, in 2022.

He is currently a Postdoc with Beihang University. His research interests include computer vision and machine learning and in particular on image restoration and image deblurring.

Chenwei Deng received the Ph.D. degree in signal and information processing from the Beijing Institute of Technology, Beijing, China, in 2009.

He was a Postdoctoral Research Fellow with the School of Computer Engineering, Nanyang Technological University, Singapore. Since 2012, he has been an Associate Professor and then a Full Professor with the School of Information and Electronics, Beijing Institute of Technology. He has authored or coauthored over 50 technical papers in refereed international journals and conferences. He has coedited two books. His research interests include video coding, quality assessment, perceptual modeling, feature representation, object recognition, and tracking.

Ding Yuan (Member, IEEE) received the Ph.D. degree in mechanical and automation engineering from the Chinese University of Hong Kong, Hong Kong, in 2008.

She is currently an Associate Professor with the Image Processing Center, School of Astronautics, Beihang University, Beijing, China, and has worked in the field of computer vision for over 15 years. Her research interests include multiview stereo, tracking, camera motion estimation, semantic segmentation.