

Learning Aberrance Repressed Correlation Filters for Real-Time UAV Tracking

Ziyuan Huang¹, Changhong Fu^{2,*}, Yiming Li², Fuling Lin² and Peng Lu³

¹School of Automotive Studies, ²School of Mechanical Engineering, Tongji University, China

³Adaptive Robotic Controls Lab, Hong Kong Polytechnic University, Hong Kong, China

tjhuangziyuan@gmail.com, changhongfu@tongji.edu.cn, peng.lu@polyu.edu.hk

Abstract

Traditional framework of discriminative correlation filters (DCF) is often subject to undesired boundary effects. Several approaches to enlarge search regions have been already proposed in the past years to make up for this shortcoming. However, with excessive background information, more background noises are also introduced and the discriminative filter is prone to learn from the ambience rather than the object. This situation, along with appearance changes of objects caused by full/partial occlusion, illumination variation, and other reasons has made it more likely to have aberrances in the detection process, which could substantially degrade the credibility of its result. Therefore, in this work, a novel approach to repress the aberrances happening during the detection process is proposed, i.e., aberrance repressed correlation filter (ARCF). By enforcing restriction to the rate of alteration in response maps generated in the detection phase, the ARCF tracker can evidently suppress aberrances and is thus more robust and accurate to track objects. Considerable experiments are conducted on different UAV datasets to perform object tracking from an aerial view, i.e., UAV123, UAVDT, and DTB70, with 243 challenging image sequences containing over 90K frames to verify the performance of the ARCF tracker and it has proven itself to have outperformed other 20 state-of-the-art trackers based on DCF and deep-based frameworks with sufficient speed for real-time applications.

1. Introduction

Visual object tracking has been widely applied in numerous fields, especially in unmanned aerial vehicle (UAV) applications, where it has been used for target following [3], mid-air aircraft tracking [11] and aerial refueling [28]. Due to fast motion of both UAV and tracked object, occlusion, deformation, illumination variation, and other challenges, robust and accurate tracking has remained a demanding task.

In recent years, discriminative correlation filter (DCF)

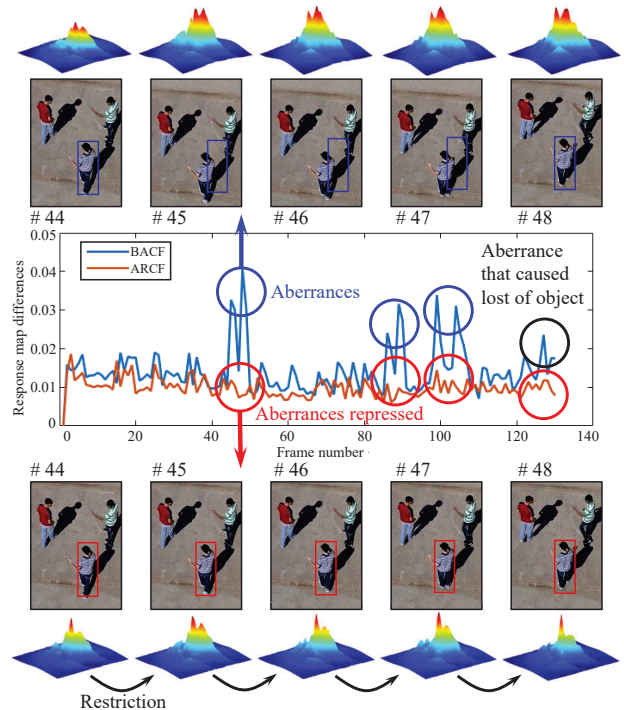


Figure 1. Comparison between background-aware correlation filter (BACF) and the proposed ARCF tracker. The central figure is to demonstrate the differences between previous response map and current response map on *group1_1* from UAV123@10fps. Sudden changes of response maps indicate aberrances. When aberrances take place, BACF is tend to lose track of the object while the proposed ARCF can repress aberrances so that this kind of drifting can be avoided.

has contributed tremendously to the field of visual tracking because of its high computational efficiency. It utilizes a property of circulant matrices to carry out the otherwise complicated calculation in the frequency domain rather than in the spatial domain to raise computing speed. Unfortunately, utilization of this property creates artificial samples, leading to undesired boundary effects, which severely degrades tracking performances.

In detection process, traditional DCF framework generates a response map and the object is believed to be located where its value is the largest. Information hidden in response map is crucial as its quality to some extent reflects the similarity between object appearance model learned in previous frames and the actual object detected in current frame. Aberrances are omnipresent in occlusion, in/out of the plan rotation and many other challenging scenarios. However, traditional DCF framework fails to utilize this information and when aberrances take place, no action can be further taken and the tracked object is simply lost.

In UAV object tracking, these two problems are especially crucial. There are relatively more cases of fast motion or low resolution and lack of search region can thus easily result in drift or lost of object. Objects also go through more out-of-the-plane rotations and thus aberrances are more likely to take place in aerial tracking scenarios. In addition, with restricted calculate capability, a tracker that can cope with these two problems and perform efficiently is especially needed.

1.1. Main contributions

This work proposes a novel tracking approach that resolves both aforementioned problems, i.e., ARCF tracker. A cropping matrix and a regularization term are introduced respectively for search region enlargement and for aberrance repression. An efficient convex optimization method is applied in order to ensure sufficient computing efficiency.

Contributions of this work can be listed as follows:

- A novel tracking method capable of effectively and efficiently suppressing aberrances while solving boundary effects is proposed. Background patches are fed into both learning and detection process to act as negative training samples and to enlarge search areas. A regularization term to restrict the change rate of response maps is added so that abrupt alteration of response maps can be avoided.
- The proposed ARCF tracker is exhaustively tested on 243 challenging image sequences captured by UAV. Both hand-crafted based trackers, i.e., histogram of oriented gradient (HOG) and color names (CN), and deep trackers are compared in the extensive experiments with the proposed ARCF tracker. Thorough evaluations have demonstrated that ARCF tracker performs favorably against other 20 state-of-the-art trackers.

To the best of our knowledge, this is the first time aberrance repression formulation has been applied in DCF framework. It can raise the robustness of DCF based trackers and improve their performances in UAV tracking tasks.

2. Related work

2.1. Discriminative correlation filter

Discriminative correlation filter based framework has been broadly applied to visual tracking since it was first introduced by Bolme *et al.* [2] who proposed a method called minimum output sum of squared error (MOSSE) filter. Kernel trick was introduced to DCF framework by Henriques *et al.* [13] to achieve better performance. Introduction of scale estimation has further improved the framework [18]. Context and background information are also exploited to have more negative samples so that learned correlation filters can have more discriminative power [7, 15, 21]. Besides hand-crafted features used in [7, 13, 15, 18], the application of deep features is also investigated for more precise and comprehensive object appearance representation [6, 12, 19]. Some trackers combine hand-crafted features with deep ones to better describe the tracked objects from multiple aspects [5, 16]. DCF based trackers have achieved state-of-the-art performance in multiple datasets specified for UAV object tracking [10, 17, 20].

2.2. Prior solution to boundary effects

As was stated before, traditional DCF based framework usually suffers from boundary effects due to the limited search region originating from its periodic shifting of the area near original object. Some measures are already taken to mitigate this effect [7, 12, 15]. Spatially regularized DCF (SRDCF) was proposed to introduce punishment for background in training correlation filters so that they can be learned in larger search regions [7]. Unfortunately, this method has high computational costs. Background-aware correlation filter (BACF) extracts patches densely from background using cropping matrix [15], which expands search region with lower computational cost. Background effect-aware visual tracker (BEVT) merges these two methods, thus achieving a better performance [12].

2.3. Prior solution to aberrances

There is few attention paid to information revealed in response maps. Wang *et al.* proposed a method called LMCF where the quality of response maps is verified in the learning phase and used to perform high-confidence update of appearance models [26], which reduces the learning rate to zero in low-confidence situations. Attentional correlation filter network (ACFN) integrates a few trackers as a network and generates a validation score for response maps from each frame. A neural network is trained based on that score to choose a suitable tracker in the next frame [4]. However, both methods take measures after the possible aberrances, which can only have limited influence in suppressing those aberrances compared to the proposed ARCF tracker which tries to repress aberrances during the training phase.

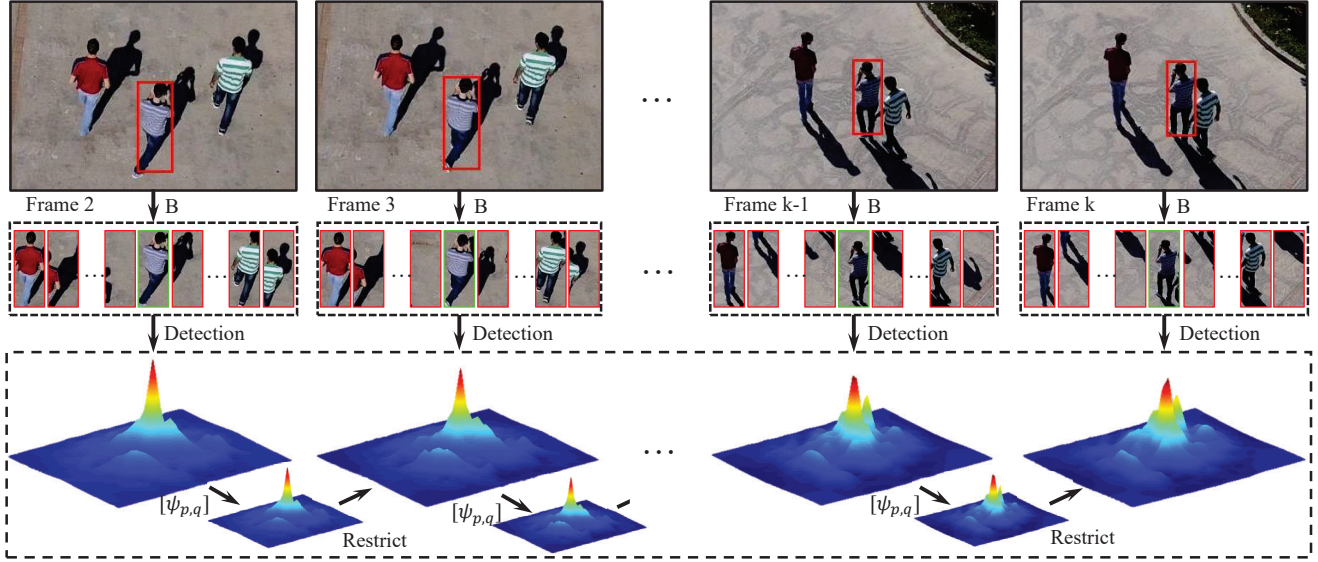


Figure 2. Main work-flow of the proposed ARCF tracker. It learns both positive sample (green samples) of the object and negative samples (red samples) extracted from the background and the response map restriction is integrated in the learning process so that aberrances in response maps can be repressed. $[\psi_{p,q}]$ serves to shift the generated response map so that the peak position in the previous frame is the same as that of the current frame and thus the position of the detected object will not affect the restriction.

3. Background-aware correlation filter

In this section, background-aware correlation filter (BACF) [15], on which our method is based, is reviewed.

Given the vectorized sample \mathbf{x} with D channels of $\mathbf{x}^d \in \mathbb{R}^N$ ($d = 1, 2, \dots, D$) and the vectorized ideal response $\mathbf{y} \in \mathbb{R}^N$, the overall objective of BACF is to minimize the objective $\mathcal{E}(\mathbf{w})$, i.e.,

$$\mathcal{E}(\mathbf{w}) = \frac{1}{2} \|\mathbf{y} - \sum_{d=1}^D \mathbf{B}\mathbf{x}^d \star \mathbf{w}^d\|_2^2 + \sum_{d=1}^D \|\mathbf{w}^d\|_2^2, \quad (1)$$

where $\mathbf{B} \in \mathbb{R}^{M \times N}$ is a cropping matrix to select central M elements of the each channel \mathbf{x}^d of input vectorized sample, and $\mathbf{w}^d \in \mathbb{R}^M$ is the correlation filter to be learned in d -th channel. Usually, $M \ll N$. The operator \star is a correlation operator.

By introducing the cropping matrix, BACF is able to utilize not only objects but also real background information instead of shifted patches in training process of correlation filters. Due to expanded search region, it is capable of tracking an object with relatively high relative speed to the camera or UAV. Unfortunately, with excessive background information, more background clutter is introduced and similar objects in the contexts are more likely than prior DCF frameworks to be detected and recognized as the original object being tracked. When this problem is observed in the response map, it can be clearly seen that BACF does not handle well when aberrances take place.

4. Aberrance repressed correlation filter

As stated in 3, BACF, just as other DCF based trackers, is vulnerable when aberrance happens. In this work, an aberrance repressed correlation filter, i.e., ARCF, is proposed to suppress sudden changes of response maps. The main structure can be seen in Fig. 2.

4.1. Overall objective of ARCF

Compared to other measures taken after occurrences of aberrances in LMCF and ACFN, the proposed ARCF tracker aims to integrate the suppression of their occurrences to the training process of correlation filters. In order to repress aberrances, they should be firstly identified. Euclidean norm is introduced to define difference level of two response maps \mathbf{M}_1 and \mathbf{M}_2 as follows:

$$\|\mathbf{M}_1[\psi_{p,q}] - \mathbf{M}_2\|_2^2, \quad (2)$$

where p and q denote the location difference of two peaks in both response maps in two-dimensional space and $[\psi_{p,q}]$ indicates the shifting operation in order for two peaks to coincide with each other. Usually when an aberrance takes place, the similarity would suddenly drop and thus the value of Eq. 2 will be high. By judging the value of Eq. 2, the aberrances can easily be identified.

In order to repress aberrances in the training process, the training objective is optimized to minimize the loss function

as follows:

$$\begin{aligned} \mathcal{E}(\mathbf{w}_k) = & \frac{1}{2} \|\mathbf{y} - \sum_{d=1}^D \mathbf{B}\mathbf{x}_k^d \star \mathbf{w}_k^d\|_2^2 + \frac{\lambda}{2} \sum_{d=1}^D \|\mathbf{w}_k^d\|_2^2 \\ & + \frac{\gamma}{2} \left\| \sum_{d=1}^D (\mathbf{B}\mathbf{x}_{k-1}^d \star \mathbf{w}_{k-1}^d) [\psi_{p,q}] - \sum_{d=1}^D \mathbf{B}\mathbf{x}_k^d \star \mathbf{w}_k^d \right\|_2^2 \end{aligned} \quad (3)$$

where subscript k and $k-1$ denote the k th and $(k-1)$ th frame respectively. The third term of Eq. 3 is a regularization term to restrict the aberrances mentioned before. Parameter γ is introduced as the aberrance penalty. In the following transformation and optimization part, the restriction will be transformed into frequency domain and optimized so that the repression can be carried out in the training process of correlation filters.

Here the cropping matrix \mathbf{B} is retained from BACF to ensure sufficient search region. Meanwhile, the regularization term is introduced to counteract the aberrances that background information has brought by expanding search area.

In order for the overall objective to be more easily transformed into frequency domain, it is firstly expressed in matrix form as follows:

$$\begin{aligned} \mathcal{E}(\mathbf{w}_k) = & \frac{1}{2} \|\mathbf{y} - \mathbf{X}_k (\mathbf{I}_D \otimes \mathbf{B}^\top) \mathbf{w}_k\|_2^2 + \frac{\lambda}{2} \|\mathbf{w}_k\|_2^2 \\ & + \frac{\gamma}{2} \|\mathbf{M}_{k-1} [\psi_{p,q}] - \mathbf{X}_k (\mathbf{I}_D \otimes \mathbf{B}^\top) \mathbf{w}_k\|_2^2 \end{aligned} \quad (4)$$

where \mathbf{X}_k is the matrix form of input sample \mathbf{x}_k . \mathbf{I}_D is an identity matrix whose size is $D \times D$. Operator \otimes and superscript \top indicates respectively Kronecker production and conjugate transpose operation. \mathbf{M}_{k-1} denotes the response map from previous frame and its value is equivalent to $\mathbf{X}_{k-1} (\mathbf{I}_D \otimes \mathbf{B}^\top) \mathbf{w}_{k-1}$.

4.2. Transformation into frequency domain

Although the overall loss function can be expressed in matrix form as Eq. 4, essentially it is still carrying out convolution operation. Therefore, to minimize the overall objective, Eq. 4 is also transformed into frequency domain as follows to ensure sufficient computing efficiency:

$$\begin{aligned} \hat{\mathcal{E}}(\mathbf{w}_k, \hat{\mathbf{g}}_k) = & \frac{1}{2} \|\hat{\mathbf{y}} - \hat{\mathbf{X}}_k \hat{\mathbf{g}}_k\|_2^2 + \frac{\lambda}{2} \|\mathbf{w}_k\|_2^2 \\ & + \frac{\gamma}{2} \|\hat{\mathbf{M}}_{k-1}^s - \hat{\mathbf{X}}_k \hat{\mathbf{g}}_k\|_2^2 \end{aligned} \quad (5)$$

s.t. $\hat{\mathbf{g}}_k = \sqrt{N} (\mathbf{I}_D \otimes \mathbf{F}\mathbf{B}^\top) \mathbf{w}_k$

where the superscript $\hat{\cdot}$ denotes a signal that has been performed discrete Fourier transformation (DFT), i.e., $\hat{\alpha} = \sqrt{N} \mathbf{F} \alpha$. A new parameter $\hat{\mathbf{g}}_k \in \mathbb{C}^{DN \times 1}$ is introduced in preparation for further optimization. $\hat{\mathbf{M}}_{k-1}^s$ denotes the discrete Fourier transformation of shifted signal $\mathbf{M}_{k-1} [\psi_{p,q}]$. Since in the current frame, the response map in the former frame is already generated, $\hat{\mathbf{M}}_{k-1}^s$ can be treated as a constant signal, which can simplify the further calculation.

4.3. Optimization through ADMM

Similar to BACF tracker, alternative direction method of multipliers (ADMM) is applied to speed up calculation. Due to the convexity of equation 5, it can be minimized using ADMM to achieve a global optimal solution. Therefore, Eq. 5 is first required to be written in augmented Lagrangian form as follows:

$$\begin{aligned} \hat{\mathcal{E}}(\mathbf{w}_k, \hat{\mathbf{g}}_k, \hat{\boldsymbol{\zeta}}) = & \frac{1}{2} \|\hat{\mathbf{y}} - \hat{\mathbf{X}}_k \hat{\mathbf{g}}_k\|_2^2 + \frac{\lambda}{2} \|\mathbf{w}_k\|_2^2 \\ & + \frac{\gamma}{2} \|\hat{\mathbf{M}}_{k-1}^s - \hat{\mathbf{X}}_k \hat{\mathbf{g}}_k\|_2^2 \\ & + \hat{\boldsymbol{\zeta}}^\top (\hat{\mathbf{g}}_k - \sqrt{N} (\mathbf{I}_D \otimes \mathbf{F}\mathbf{B}^\top) \mathbf{w}_k) \\ & + \frac{\mu}{2} \|\hat{\mathbf{g}}_k - \sqrt{N} (\mathbf{I}_D \otimes \mathbf{F}\mathbf{B}^\top) \mathbf{w}_k\|_2^2 \end{aligned} \quad (6)$$

where μ is introduced as a penalty factor and the Lagrangian vector in the Fourier domain $\hat{\boldsymbol{\zeta}} = [\hat{\zeta}^1, \dots, \hat{\zeta}^{DN}]^\top$ is introduced as auxiliary variable that has a size of $DN \times 1$.

Employing ADMM in the k th frame means that the augmented Lagrangian form can be solved by solving two subproblems, respectively the following \mathbf{w}_{k+1}^* and $\hat{\mathbf{g}}_{k+1}^*$ to calculate correlation filters for the $(k+1)$ th frame:

$$\begin{cases} \mathbf{w}_{k+1}^* = \arg \min_{\mathbf{w}_k} \left\{ \frac{\lambda}{2} \|\mathbf{w}_k\|_2^2 \right. \\ \quad \left. + \hat{\boldsymbol{\zeta}}^\top (\hat{\mathbf{g}}_k - \sqrt{N} (\mathbf{I}_D \otimes \mathbf{F}\mathbf{B}^\top) \mathbf{w}_k) \right. \\ \quad \left. + \frac{\mu}{2} \|\hat{\mathbf{g}}_k - \sqrt{N} (\mathbf{I}_D \otimes \mathbf{F}\mathbf{B}^\top) \mathbf{w}_k\|_2^2 \right\} \\ \hat{\mathbf{g}}_{k+1}^* = \arg \min_{\hat{\mathbf{g}}_k} \left\{ \frac{1}{2} \|\hat{\mathbf{y}} - \hat{\mathbf{X}}_k \hat{\mathbf{g}}_k\|_2^2 \right. \\ \quad \left. + \frac{\gamma}{2} \|\hat{\mathbf{M}}_{k-1}^s - \hat{\mathbf{X}}_k \hat{\mathbf{g}}_k\|_2^2 \right. \\ \quad \left. + \hat{\boldsymbol{\zeta}}^\top (\hat{\mathbf{g}}_k - \sqrt{N} (\mathbf{I}_D \otimes \mathbf{F}\mathbf{B}^\top) \mathbf{w}_k) \right. \\ \quad \left. + \frac{\mu}{2} \|\hat{\mathbf{g}}_k - \sqrt{N} (\mathbf{I}_D \otimes \mathbf{F}\mathbf{B}^\top) \mathbf{w}_k\|_2^2 \right\} \end{cases} \quad (7)$$

Both of these two subproblems have closed-form solutions.

4.3.1 Solution to subproblem \mathbf{w}_k^*

The solution to subproblem \mathbf{w}_k^* can be easily obtained as follows:

$$\begin{aligned} \mathbf{w}_{k+1}^* = & (\lambda + \mu N)^{-1} \\ & \left(\sqrt{N} (\mathbf{I}_D \otimes \mathbf{B}\mathbf{F}^\top) \hat{\boldsymbol{\zeta}} + \mu \sqrt{N} (\mathbf{I}_D \otimes \mathbf{B}\mathbf{F}^\top) \hat{\mathbf{g}}_k \right), \quad (8) \\ = & \left(\frac{\lambda}{N} + \mu \right)^{-1} (\zeta + \mu \mathbf{g}_k) \end{aligned}$$

where \mathbf{g}_k and ζ can be obtained respectively through following inverse fast Fourier transformation operations:

$$\begin{cases} \mathbf{g}_k = \frac{1}{\sqrt{N}} (\mathbf{I}_D \otimes \mathbf{B}\mathbf{F}^\top) \hat{\mathbf{g}}_k \\ \zeta = \frac{1}{\sqrt{N}} (\mathbf{I}_D \otimes \mathbf{B}\mathbf{F}^\top) \hat{\boldsymbol{\zeta}} \end{cases} \quad (9)$$

4.3.2 Solution to subproblem $\hat{\mathbf{g}}_k^*$

Unfortunately, unlike subproblem \mathbf{w}_k^* , solving subproblem $\hat{\mathbf{g}}_k^*$ containing $\hat{\mathbf{X}}_k \hat{\mathbf{g}}_k$ can be highly time consuming and the calculation needs to be carried out in every ADMM iteration. Therefore, the sparsity of $\hat{\mathbf{X}}_k$ is exploited. Each element of $\hat{\mathbf{y}}$, i.e., $\hat{\mathbf{y}}(n)$, $n = 1, 2, \dots, N$, is solely dependent on each $\hat{\mathbf{x}}_k(n) = [\hat{\mathbf{x}}_k^1(n), \hat{\mathbf{x}}_k^2(n), \dots, \hat{\mathbf{x}}_k^D(n)]^\top$ and $\hat{\mathbf{g}}_k(n) = [\text{conj}(\hat{\mathbf{g}}_k^1(n)), \dots, \text{conj}(\hat{\mathbf{g}}_k^D(n))]^\top$. Operator $\text{conj}(\cdot)$ denotes the complex conjugate operation.

The subproblem $\hat{\mathbf{g}}_k^*$ can be thus further divided into N smaller problems as follows solved over $n = [1, 2, \dots, N]$:

$$\begin{aligned} \hat{\mathbf{g}}_{k+1}(n)^* = \arg \min_{\mathbf{g}_k(n)} \left\{ \frac{1}{2} \|\hat{\mathbf{y}}(n) - \hat{\mathbf{x}}_k^\top(n) \hat{\mathbf{g}}_k(n)\|_2^2 \right. \\ \left. + \frac{\gamma}{2} \|\hat{\mathbf{M}}_{k-1}^s - \hat{\mathbf{x}}_k^\top(n) \hat{\mathbf{g}}_k(n)\|_2^2 \right. \\ \left. + \hat{\zeta}^\top (\hat{\mathbf{g}}_k(n) - \hat{\mathbf{w}}_k(n)) \right. \\ \left. + \frac{\mu}{2} \|\hat{\mathbf{g}}_k(n) - \hat{\mathbf{w}}_k(n)\|_2^2 \right\}, \quad (10) \end{aligned}$$

where $\hat{\mathbf{w}}_k(n) = [\hat{\mathbf{w}}_k^1(n), \dots, \hat{\mathbf{w}}_k^D(n)]$ and $\hat{\mathbf{w}}_k^d$ is the DFT of \mathbf{w}_k^d , i.e., $\hat{\mathbf{w}}_k^d = \sqrt{D} \mathbf{F} \mathbf{B}^\top \mathbf{w}_k^d$. Each smaller problem can be efficiently calculated and solution is presented below:

$$\begin{aligned} \hat{\mathbf{g}}_{k+1}(n)^* = \frac{1}{1+\gamma} \left(\hat{\mathbf{x}}_k(n) \hat{\mathbf{x}}_k^\top(n) + \frac{\mu}{1+\gamma} \mathbf{I}_D \right)^{-1} \\ \left(\hat{\mathbf{x}}_k(n) \hat{\mathbf{y}}(n) + \gamma \hat{\mathbf{x}}_k(n) \hat{\mathbf{M}}_{k-1}^s - \hat{\zeta}(n) + \mu \hat{\mathbf{w}}_k(n) \right) \end{aligned} \quad (11)$$

Still, with inverse operation, the calculation can be further optimized and accelerated by applying the Sherman-Morrison formula, i.e., $(\mathbf{A} + \mathbf{u}\mathbf{v}^\top)^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1} \mathbf{u} (\mathbf{I}_m + \mathbf{v}^\top \mathbf{A}^{-1} \mathbf{u})^{-1} \mathbf{v}^\top \mathbf{A}^{-1}$, where \mathbf{u} is an $a \times m$ matrix, \mathbf{v} is and $m \times a$ matrix and \mathbf{A} is an $a \times a$ matrix. In this case, $\mathbf{A} = \frac{\mu}{1+\gamma} \mathbf{I}_D$, and $\mathbf{u} = \mathbf{v} = \hat{\mathbf{x}}_k(n)$. Eq. 11 is equivalent to the following equation:

$$\begin{aligned} \hat{\mathbf{g}}_{k+1}(n)^* \\ = \gamma^* \left(\hat{\mathbf{x}}_k(n) \hat{\mathbf{y}}(n) + \gamma \hat{\mathbf{x}}_k(n) \hat{\mathbf{M}}_{k-1}^s - \hat{\zeta}(n) + \mu \hat{\mathbf{w}}_k(n) \right) \\ - \gamma^* \frac{\hat{\mathbf{x}}_k(n)}{b} \left(\hat{\mathbf{S}}_{\mathbf{x}k}(n) \hat{\mathbf{y}}(n) + \gamma \hat{\mathbf{S}}_{\mathbf{x}k}(n) \hat{\mathbf{M}}_{k-1}^s \hat{\mathbf{S}}_{\zeta}(n) + \mu \hat{\mathbf{S}}_{\mathbf{w}k}(n) \right) \end{aligned} \quad (12)$$

where $\gamma^* = \frac{\mu}{(1+\gamma)^2}$, $\hat{\mathbf{S}}_{\mathbf{x}k}(n) = \hat{\mathbf{x}}_k(n)^\top \hat{\mathbf{x}}_k(n)$, $\hat{\mathbf{S}}_{\zeta}(n) = \hat{\mathbf{x}}_k(n)^\top \hat{\zeta}$, $\hat{\mathbf{S}}_{\mathbf{w}k}(n) = \hat{\mathbf{x}}_k(n)^\top \hat{\mathbf{w}}_k$ and $b = \hat{\mathbf{x}}_k(n)^\top \hat{\mathbf{x}}_k(n) + \frac{\mu}{1+\gamma}$. Thus far, the subproblems \mathbf{w}_{k+1}^* and $\hat{\mathbf{g}}_{k+1}^*$ are both solved.

4.3.3 Update of Lagrangian parameter

The Lagrangian parameter is updated according to the following equation:

$$\hat{\zeta}_{k+1}^{(j+1)} = \hat{\zeta}_{k+1}^j + \mu \left(\hat{\mathbf{g}}_{k+1}^{*(j+1)} - \hat{\mathbf{w}}_{k+1}^{*(j+1)} \right), \quad (13)$$

where the subscript j and $j+1$ denotes the j th and the $(j+1)$ th iteration respectively. $\hat{\mathbf{g}}_{k+1}^{*(j+1)}$ indicates the solution to the $\hat{\mathbf{g}}_{k+1}^*$ subproblem and $\hat{\mathbf{w}}_{k+1}^{*(j+1)}$ indicates the solution to the \mathbf{w}_{k+1}^* subproblem, both in the $(j+1)$ th iteration. Here $\hat{\mathbf{w}}_{k+1}^{*(j+1)} = (\mathbf{I}_D \otimes \mathbf{F} \mathbf{B}^\top) \mathbf{w}_{k+1}^{*(j+1)}$.

4.4. Update of appearance model

The appearance model $\hat{\mathbf{x}}^{\text{model}}$ is updated as follows:

$$\hat{\mathbf{x}}_k^{\text{model}} = (1 - \eta) \hat{\mathbf{x}}_{k-1}^{\text{model}} + \eta \hat{\mathbf{x}}_k, \quad (14)$$

where k and $k-1$ denote k th and $(k-1)$ th frame respectively. η is the learning rate of the appearance model.

5. Experiments

In this section, the proposed ARCF tracker is exhaustively evaluated on 243 challenging image sequences with altogether over 90,000 frames from three widely applied benchmarks captured by UAV for tracking, respectively UAV123@10fps [20], DTB70 [17] and UAVDT [10]. The results are compared with 20 state-of-the-art trackers with both hand-crafted based trackers and deep-based trackers, i.e., KCF [13], DSST [8], SAMF [18], MUSTER [14], BACF [15], SRDCF [7], STAPLE_CA [21], MCCT-H [27], STRCF [16], ECO-HC (with gray-scale) [5], ECO [5], C-COT [9], HCF [19], ADNet [29], CFNet [25], CREST [23], MCPF [30], SINT [24], SiamFC [1], and HDT [22]. All evaluation criteria are according to the original protocol defined in three benchmarks respectively [10, 17, 20].

5.1. Implementation details

Two versions of ARCF tracker, respectively ARCF-H (with only HOG feature) and ARCF-HC (with HOG, CN and gray-scale features) are developed in the experiment to achieve comprehensive comparison with all trackers using HOG, both HOG and CN, as well as deep features. The value of γ is set to 0.71, ADMM iteration is set to 5 and the learning rate η is 0.0192. All experiments of all 21 trackers are carried out by MATLAB R2017a on a computer with an i7-8700K processor (3.7GHz), 48GB RAM and NVIDIA Quadro P2000 GPU. Tracking code is available here: <https://github.com/vision4robotics/ARCF-tracker>.

5.2. Comparison with hand-crafted based trackers

5.2.1 Quantitive evaluation

Overall performance evaluation: Figure 3 demonstrates the overall performance of ARCF-H and ARCF-HC with other state-of-the-art hand-crafted feature-based trackers on UAV123@10fps, DTB70 and UAVDT datasets. The proposed ARCF-HC tracker has outperformed all other trackers based on hand-crafted features on all three datasets.

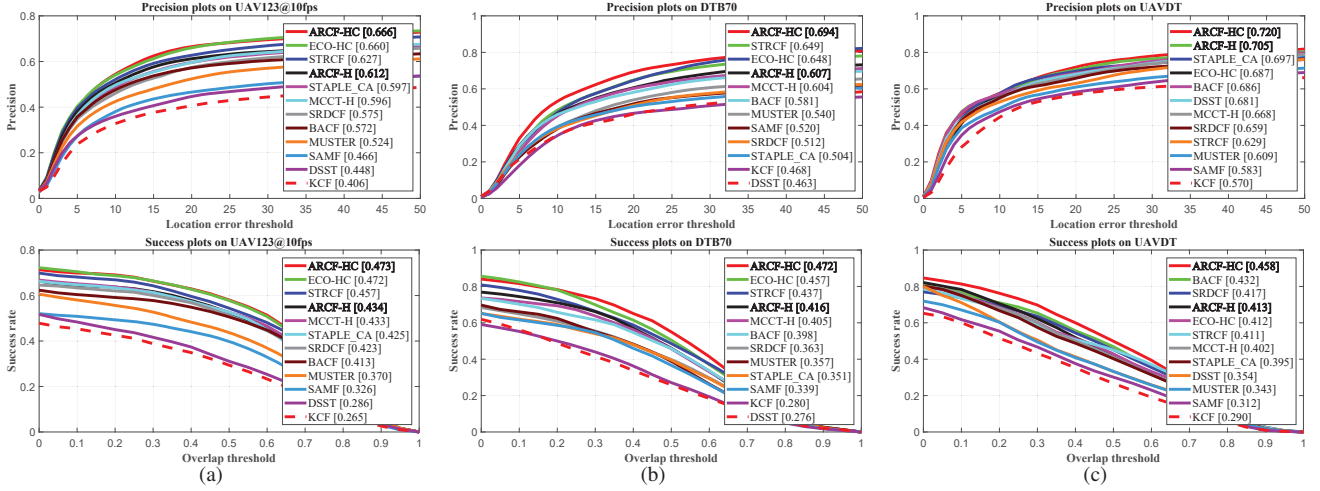


Figure 3. Precision and success plots of ARCF-HC, ARCF-H as well as other hand-crafted feature-based trackers on (a) UAV123, (b) DTB70 and (c) UAVDT. Precision and AUC are marked in the precision plots and success plots respectively.

Table 1. Average frame per second (FPS) and millisecond per frame (MSPF) of top hand-crafted based trackers on 243 image sequences. Red, green and blue fonts indicate the first, second and third place, respectively. All results are generated solely by CPU.

	ARCF-H	ARCF-HC	ECO-HC	STRCF	MCCT-H	STAPLE_CA	SRDCF	BACF	MUSTER	SAMF	DSST	KCF
FPS	51.2	15.3	41.1	22.6	32.1	37.2	11.7	52.5	2.1	9.9	100.7	326.1
MSPF	19.53	65.36	24.33	44.25	31.15	26.88	85.47	19.05	476.19	101.01	9.93	3.07

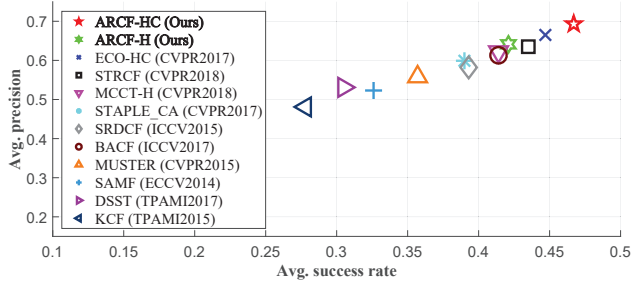


Figure 4. Comparison of different state-of-the-art trackers based on hand-crafted features. Value of average precision and average success rate is calculated by averaging OPE results from three datasets.

More specifically, on UAV123@10fps dataset, ARCF-HC (0.666) has an advantage of 0.6% and 3.9% over the second and third best tracker ECO-HC (0.660), STRCF (0.627) respectively in precision, as well as an advantage of 0.1% and 1.6% over the second (ECO-HC, 0.472) and third best tracker (STRCF, 0.457) respectively in AUC. On DTB70 dataset, ARCF-HC (0.694, 0.472) also achieved the best performance, followed by ECO-HC (0.648, 0.457) and STRCF (0.649, 0.437). On UAVDT, ARCF-HC tracker (0.720, 0.458) is closely followed by ARCF-H (0.705) and BACF (0.432) in precision and AUC respectively. Overall evaluation of performance on all three datasets in terms of precision and AUC is demonstrated in Fig. 4. Against

the baseline BACF, ARCF-H has an advancement of 2.77% in precision and 0.69% in AUC. ARCF-HC has made a progress of 7.98% and 5.32% in precision and AUC respectively. Besides satisfactory tracking results, the speed of ARCF-H and ARCF-HC is adequate for real-time UAV tracking applications, as shown in Table 1.

Attribute based comparison: In this section, quantitative analysis of different attributes in three benchmarks are performed. The proposed ARCF-HC tracker has performed favorably against other top hand-crafted based trackers in most attributes defined respectively in three benchmarks. Examples of overlap success plots are demonstrated in Fig. 5. In partial or full occlusion cases, ARCF-H and ARCF-HC demonstrated a huge improvement from its baseline BACF, and have achieved state-of-the-art performance in this aspect on all three benchmarks. Usually, in occlusion cases, CF learns appearance model of both the tracked object and irrelevant objects that caused occlusions. ARCF is able to restrict the learning of irrelevant objects by restricting the response map variations, thus achieving a better performance in occlusion cases. More specifically, ARCF-HC has achieved an advancement of 8.1% (UAV123@10fps), 9.8% (DTB70) and 5.2% (UAVDT) respectively from BACF in AUC in occlusion cases. In other attributes, ARCF-H and ARCF-HC have also shown a great improvement from BACF and achieved a performance with a high ranking. More complete results of attribute evaluation can be found on supplementary materials.

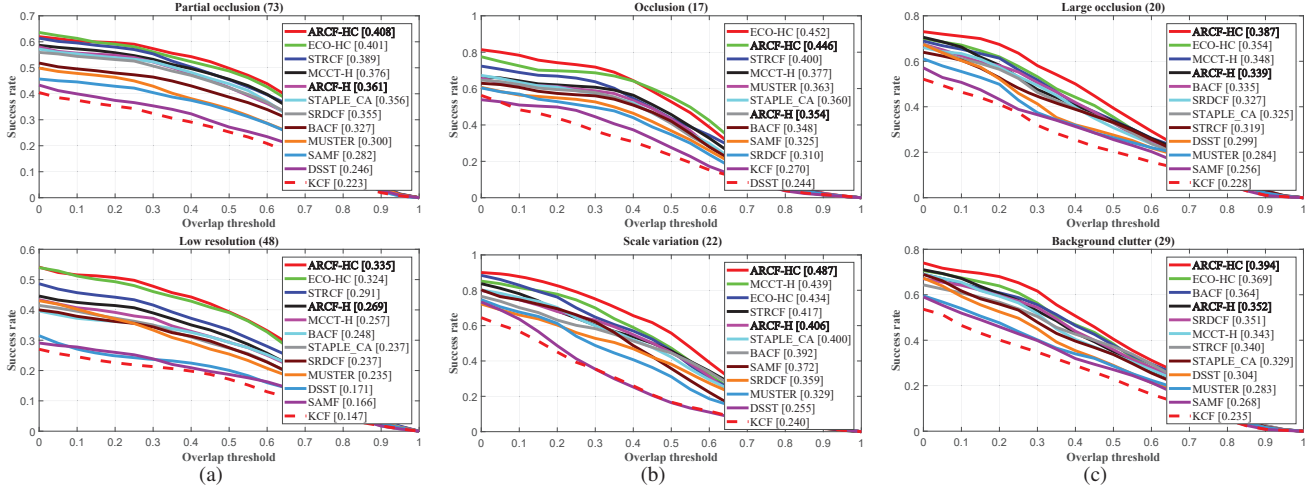


Figure 5. Attribute based evaluation. Success plots of attributes comparing ARCF-HC and ARCF-H with other state-of-the-art hand-crafted based trackers on (a) UAV123@10fps, (b) DTB70 and (c) UAVDT. AUC is used to rank different trackers. Detailed definitions and descriptions of these attributes can be seen in [10, 17, 20].

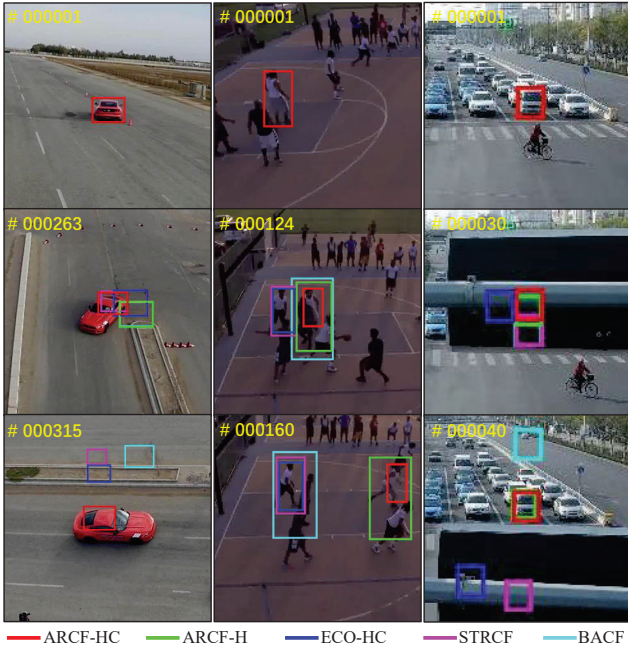


Figure 6. Qualitative performance evaluation of ARCF-H and ARCF-HC tracker on *Car16.2* from UAV123 dataset, *StreetBasketball3* from DTB70 dataset and *S0601* from UAVDT dataset.

5.2.2 Qualitative evaluation

Some qualitative tracking results of ARCF and other top trackers are shown in Fig. 6. It can be proven that ARCF is competent in dealing with both partial as well as full occlusions and performs satisfactorily in other aspects defined in three benchmarks as well.

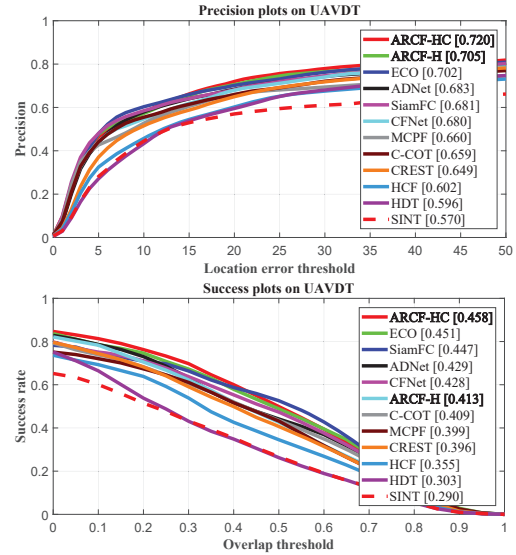


Figure 7. Comparison between ARCF tracker and different state-of-the-art deep-based trackers. Value of average precision and average success rate is calculated by averaging OPE results from three datasets.

5.3. Comparison with deep-based trackers

To achieve a more comprehensive evaluation of the proposed trackers ARCF-H and ARCF-HC, these two trackers are also compared to ones using deep features or even deep trackers. In terms of precision and success rate, ARCF-HC has also performed favorably against other state-of-the-art deep-based trackers. Fig. 7 has shown the quantitative comparison on UAVDT dataset.

Table 2. Average map difference comparison of BACF and ARCF-H on different datasets. Map difference is evaluated by Eq. 2. **Bold font indicates lower average difference.**

	UAV123@10fps	DTB70	UAVDT
ARCF-H	0.0106	0.0098	0.0074
BACF	0.0133	0.0129	0.0087

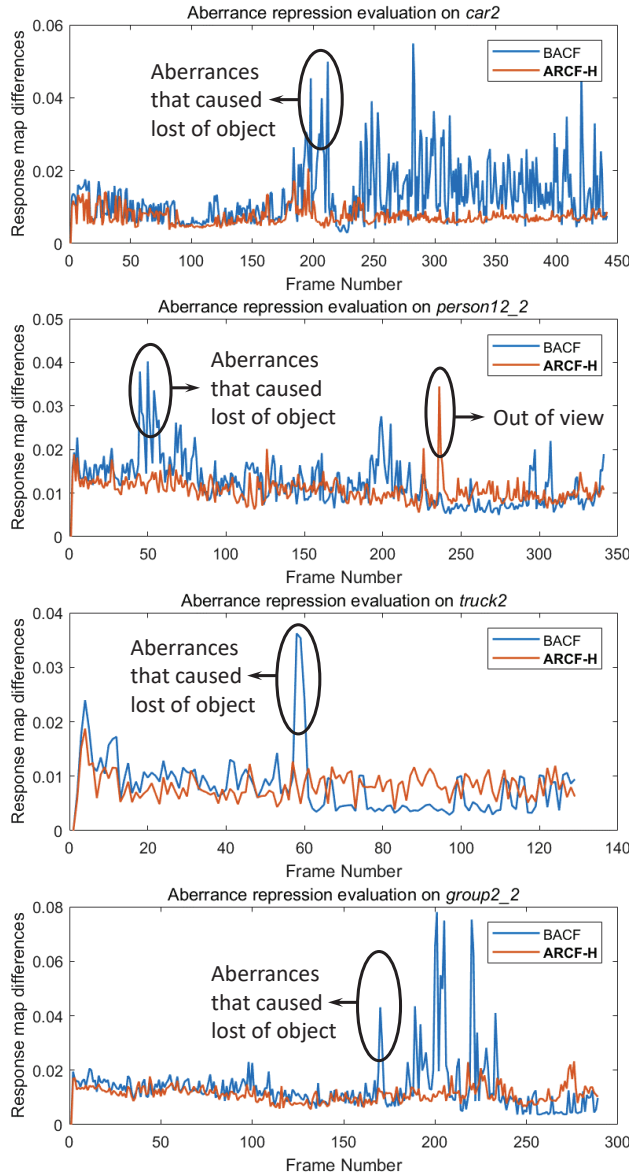


Figure 8. Comparison of response map differences between BACF tracker and ARCF tracker on UAV123@10fps dataset, specifically on *car2*, *person12_2*, *truck_2* and *group2_2*. The proposed ARCF tracker has remarkably repressed aberrances that can possibly cause lost of object. Note that after the out-of-view on *person12_2*, ARCF rapidly recapture the original tracked object.

5.4. Aberrance repression evaluation

In order to illustrate the effect of aberrance repression, this section investigates the difference between tracking performance of BACF and ARCF-H trackers. It can be clearly seen from Table 2 that ARCF-H tracker has significantly repressed the average map difference compared to BACF for respectively 20%, 24%, and 15% on UAV123@10fps, DTB70 and UAVDT dataset. Response map differences are visualized in Fig. 8 to demonstrate the performance of aberrance repression method. When objects go through relatively big appearance changes due to sudden illumination variation, partial or full occlusion and other reasons, response map tends to fluctuate and aberrances are very likely to happen, as denoted in Fig. 8. Although it is possible in cases like out-of-view and full occlusion that aberrances happen in ARCF tracker, ARCF is able to suppress most undesired fluctuations so that the tracker can be more robust against these appearance changes. It should be brought to attention that this kind of fluctuation is omnipresent in tracking scenarios of various image sequences. More examples of visualization of response map differences can be seen in the supplementary material.

6. Conclusion and future work

In this work, aberrance repressed correlation filters have been proposed for UAV visual tracking. By introducing a regularization term to restrict the response map variations to BACF, ARCF is capable of suppressing aberrances that is caused by both background noise information introduced by BACF and appearance changes of the tracked objects. After careful and exhaustive evaluation on three prevalent tracking benchmarks captured by UAVs, ARCF has proved itself to have achieved a big advancement from BACF and have state-of-the-art performance in terms of precision and success rate. Its speed is also more than sufficient for real-time UAV tracking. In conclusion, the proposed method i.e., aberrance repression correlation filters (ARCF), is able to raise the performance of DCF trackers without sacrificing much speed. Out of consideration for computing efficiency due to application of UAV tracking, the proposed ARCF has only used HOG and CN as extracted feature. In cases with low demand for real-time application, more comprehensive features such as convolutional ones can be applied to ARCF for better precision and success rate. Also, the framework of aberrance repression can be extended to other trackers like ECO [5] and SRDCF [7]. We believe, with our proposed aberrance repression method, DCF framework and the performances of DCF based trackers can be further improved.

Acknowledgment: This work is supported by the National Natural Science Foundation of China (No. 61806148) and the Fundamental Research Funds for the Central Universities (No. 22120180009).

References

- [1] Luca Bertinetto, Jack Valmadre, Joao F Henriques, Andrea Vedaldi, and Philip HS Torr. Fully-convolutional siamese networks for object tracking. In *European conference on computer vision*, pages 850–865. Springer, 2016.
- [2] David S Bolme, J Ross Beveridge, Bruce A Draper, and Yui Man Lui. Visual object tracking using adaptive correlation filters. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2544–2550. IEEE.
- [3] Hui Cheng, Lishan Lin, Zhuoqi Zheng, Yuwei Guan, and Zhongchang Liu. An autonomous vision-based target tracking system for rotorcraft unmanned aerial vehicles. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1732–1738, Sep. 2017.
- [4] Jongwon Choi, Hyung Jin Chang, Sangdoo Yun, Tobias Fischer, Yiannis Demiris, and Jin Young Choi. Attentional correlation filter network for adaptive visual tracking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4807–4816, 2017.
- [5] Martin Danelljan, Goutam Bhat, Fahad Shahbaz Khan, and Michael Felsberg. Eco: Efficient convolution operators for tracking. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6931–6939, 2017.
- [6] Martin Danelljan, Gustav Hager, Fahad Shahbaz Khan, and Michael Felsberg. Convolutional features for correlation filter based visual tracking. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 58–66, 2015.
- [7] Martin Danelljan, Gustav Hager, Fahad Shahbaz Khan, and Michael Felsberg. Learning spatially regularized correlation filters for visual tracking. In *Proceedings of the IEEE international conference on computer vision*, pages 4310–4318, 2015.
- [8] Martin Danelljan, Gustav Häger, Fahad Shahbaz Khan, and Michael Felsberg. Discriminative scale space tracking. *IEEE transactions on pattern analysis and machine intelligence*, 39(8):1561–1575, 2017.
- [9] Martin Danelljan, Andreas Robinson, Fahad Shahbaz Khan, and Michael Felsberg. Beyond correlation filters: Learning continuous convolution operators for visual tracking. In *European Conference on Computer Vision*, pages 472–488. Springer, 2016.
- [10] Dawei Du, Yuankai Qi, Hongyang Yu, Yifan Yang, Kaiwen Duan, Guorong Li, Weigang Zhang, Qingming Huang, and Qi Tian. The unmanned aerial vehicle benchmark: object detection and tracking. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 370–386, 2018.
- [11] Changhong Fu, Adrian Carrio, Miguel A Olivares-Mendez, Ramon Suarez-Fernandez, and Pascual Campoy. Robust real-time vision-based aircraft tracking from unmanned aerial vehicles. In *2014 IEEE international conference on robotics and automation (ICRA)*, pages 5441–5446. IEEE, 2014.
- [12] Changhong Fu, Ziyuan Huang, Yiming Li, Ran Duan, and Peng Lu. Boundary effect-aware visual tracking for uav with online enhanced background learning and multi-frame consensus verification. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2019.
- [13] João F Henriques, Rui Caseiro, Pedro Martins, and Jorge Batista. High-speed tracking with kernelized correlation filters. *IEEE Trans Pattern Analysis and Machine Intelligence*, 37(3):583–96, 2015.
- [14] Zhibin Hong, Zhe Chen, Chaohui Wang, Xue Mei, Danil Prokhorov, and Dacheng Tao. Multi-store tracker (muster): A cognitive psychology inspired approach to object tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 749–758, 2015.
- [15] Hamed Kiani Galoogahi, Ashton Fagg, and Simon Lucey. Learning background-aware correlation filters for visual tracking. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1135–1143, 2017.
- [16] Feng Li, Cheng Tian, Wangmeng Zuo, Lei Zhang, and Ming-Hsuan Yang. Learning spatial-temporal regularized correlation filters for visual tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4904–4913, 2018.
- [17] Siyi Li and Dit-Yan Yeung. Visual object tracking for unmanned aerial vehicles: A benchmark and new motion models. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.
- [18] Yang Li and Jianke Zhu. A scale adaptive kernel correlation filter tracker with feature integration. In *European conference on computer vision*, pages 254–265. Springer, 2014.
- [19] Chao Ma, Jia-Bin Huang, Xiaokang Yang, and Ming-Hsuan Yang. Hierarchical convolutional features for visual tracking. In *Proceedings of the IEEE international conference on computer vision*, pages 3074–3082, 2015.
- [20] Matthias Mueller, Neil Smith, and Bernard Ghanem. A benchmark and simulator for uav tracking. In *European conference on computer vision*, pages 445–461. Springer, 2016.
- [21] Matthias Mueller, Neil Smith, and Bernard Ghanem. Context-aware correlation filter tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1396–1404, 2017.
- [22] Yuankai Qi, Shengping Zhang, Lei Qin, Hongxun Yao, Qingming Huang, Jongwoo Lim, and Ming-Hsuan Yang. Hedged deep tracking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4303–4311, 2016.
- [23] Yibing Song, Chao Ma, Lijun Gong, Jiawei Zhang, Rynson WH Lau, and Ming-Hsuan Yang. Crest: Convolutional residual learning for visual tracking. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2555–2564, 2017.
- [24] Ran Tao, Efstratios Gavves, and Arnold WM Smeulders. Siamese instance search for tracking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1420–1429, 2016.
- [25] Jack Valmadre, Luca Bertinetto, João Henriques, Andrea Vedaldi, and Philip HS Torr. End-to-end representation learning for correlation filter based tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2805–2813, 2017.

- [26] Mengmeng Wang, Yong Liu, and Zeyi Huang. Large margin object tracking with circulant feature maps. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4021–4029, 2017.
- [27] Ning Wang, Wengang Zhou, Qi Tian, Richang Hong, Meng Wang, and Houqiang Li. Multi-cue correlation filters for robust visual tracking. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4844–4853, 2018.
- [28] Yingjie Yin, Xingang Wang, De Xu, Fangfang Liu, Yinglu Wang, and Wenqi Wu. Robust Visual Detection–Learning–Tracking Framework for Autonomous Aerial Refueling of UAVs. *IEEE Transactions on Instrumentation and Measurement*, 65(3):510–521, March 2016.
- [29] Sangdoo Yun, Jongwon Choi, Youngjoon Yoo, Kimin Yun, and Jin Young Choi. Action-decision networks for visual tracking with deep reinforcement learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2711–2720, 2017.
- [30] Tianzhu Zhang, Changsheng Xu, and Ming-Hsuan Yang. Multi-task correlation particle filter for robust object tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4335–4343, 2017.