BiCF: Learning Bidirectional Incongruity-Aware Correlation Filter for Efficient UAV Object Tracking

Fuling Lin[1], Changhong Fu[1,*], Yujie He[1], Fuyu Guo[2], and Qian Tang[2]

***Abstract*— Correlation filters (CFs) have shown excellent performance in unmanned aerial vehicle (UAV) tracking scenarios due to their high computational efficiency. During the UAV tracking process, viewpoint variations are usually accompanied by changes in the object and background appearance, which poses a unique challenge to CF-based trackers. Since the appearance is gradually changing over time, an ideal tracker can not only forward predict the object position but also backtrack to locate its position in the previous frame. There exist response-based errors in the reversibility of the tracking process containing the information on the changes in appearance. However, some existing methods do not consider the forward and backward errors based on while using only the current training sample to learn the filter. For other ones, the applicants of considerable historical training samples impose a computational burden on the UAV. In this work, a novel bidirectional incongruity-aware correlation filter (BiCF) is proposed. By integrating the response-based bidirectional incongruity error into the CF, BiCF can efficiently learn the changes in appearance and suppress the inconsistent error. Extensive experiments on 243 challenging sequences from three UAV datasets (UAV123, UAVDT, and DTB70) are conducted to demonstrate that BiCF favorably outperforms other 25 state-of-the-art trackers and achieves a real-time speed of 45.4 FPS on a single CPU, which can be applied in UAV efficiently.**

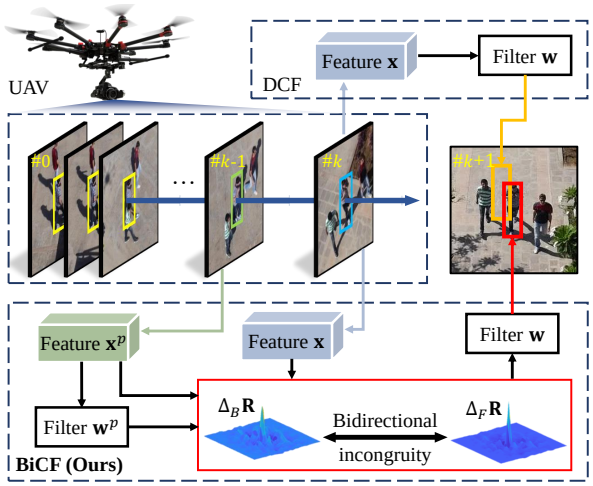


Fig. 1. Comparison between discriminative correlation filter (DCF) and the proposed BiCF. During the filter training phase in frame #k, BiCF uses both the sample information and filter of the previous frame to help construct the bidirectional incongruity error, aiming to utilize the inter-frame information better. The DCF only uses the sample information of the current frame, which makes the filter susceptible to the appearance changes.

I. Introduction

Unmanned aerial vehicles (UAVs) have been widely used recently, especially those equipped with intelligent vision-based technology. The UAV senses the environment through the visual system on the onboard computer to detect and track the specified target autonomously. In the literature, UAV tracking has been applied for human-computer interaction [1], autonomous landing [2], and object following [3].

Although many visual tracking methods have been proposed in recent years, there remain considerable challenges in visual object tracking, especially in UAV tracking scenarios. Factors that have a significant impact on UAV tracking include object deformation, the fast motion of the UAV/object, a wide range of viewpoint changes, to name a few [4]. Comparing with general tracking cases, tracking the object from the UAV perspective faces more particular difficulties. Currently, the correlation filter (CF)-based tracking method has received widespread attention with high computational efficiency, which is suitable for UAV real-time tracking applications. This method learns a CF-based tracker that can distinguish the specified target from the background and update the appearance model online. All computation benefits from the characteristic of the CF-based method and can be performed efficiently in the Fourier domain. However, the change of the UAV viewpoint usually makes the appearance model susceptible to interference. The change in CF's response to the object reflects the information about the object appearance changes, which is ignored by most traditional CF-based methods.

In UAV tracking scenarios, viewpoint changes often occur. To overcome similar challenges and improve the UAV tracking performance in robustness, using powerful features are essential for the object expression. Moreover, in the case of limited computing resources of the UAV, we consider the information about variations in the object/background appearance between frames, instead of only using the current sample for filter learning, as shown in Fig. 1.

In this work, multiple features are applied to express the object/background appearance including histogram of oriented gradient (HOG) [5] and color names (CN) [6]. Moreover, we dissect the tracking process with a view of forward detection and backward relocation. It can be found that there exists error reflecting the same appearance changes in the forward and backward tracking process. The error is referred to as bidirectional incongruity error in this work, and it can be measured by the response maps. By suppressing this response-based error, the filter can be more robust to the

[1]Fuling Lin, Changhong Fu, and Yujie He are with the School of Mechanical Engineering, Tongji University, 201804 Shanghai, China changhongfu@tongji.edu.cn

[2]Fuyu Guo and Qian Tang are with the College of Mechanical Engineering, Chongqing University, 400044 Chongqing, China

appearance variations. Therefore, a bidirectional incongruity-aware CF-based tracker (BiCF) is proposed in this work, which incorporates this error into the CF-based framework. The contributions are summarized as follows:

- A novel tracker is proposed, which can learn the object/background appearance changes more efficiently. By using inter-frame information to analyze and resolve the response-based bidirectional incongruity error, it helps to encode the filter with high accuracy and robustness. Moreover, multiple features (HOG and CN) are used to assist in the object/background expression.
- Considerable experiments on three UAV datasets with 243 challenging image sequences are conducted to verify the performance of the proposed BiCF tracker. Experimental results demonstrate the BiCF tracker performs favorably comparing with other 25 state-of-the-art trackers in accuracy, robustness, and efficiency.

II. Related Works

CF-based trackers have achieved significant progress recently. In [7], the minimum output sum of squared error, i.e., MOSSE tracker, was applied to learn a correlation filter to distinguish objects. Upon the MOSSE method, several works [8], [9] were proposed and show notable improvement by using multi-channel features in CF learning. To better encode the tracking model, J. F. Henriques et al. utilized kernel tricks to improve CF-based tracking performance [10]. The SAMF [11], DSST [12], and IBCCF [13] trackers were proposed to address the adaptive scale change problem. The BACF [14], SRDCF [15], and BEVT [16] trackers were designed to alleviate the boundary effects.

During the tracking process, the trackers need to maintain a robust appearance and filter model as the object and background changes over time. It is a challenge to CF-based trackers due to the limited training samples. KCF [10] was provided with some memory by updating both appearance and filter model, which makes the tracker more robust to object appearance variations. In [15], [17], all historical training samples are considered for current filter learning, but better precision is achieved at the expense of a high computational burden. To better use the inter-frame information and reduce the training set size, M. Danelljan et al. applied the Gaussian mixture model to generate sample space with a smaller number of samples [18]. F. Li et al. proposed the temporal regularization to learn and update the CF simultaneously without using the large training set [19]. The temporal regularization makes the filter similar to the previous one. However, the changes within two consecutive frames are not only reflected in the filter but also reflected in the variations of the appearance. In [20], the responses within two consecutive frames were used for aberrance repression, which only focused on the forward tracking process.

Therefore, we analyze the tracking process from the forward and backward perspective and propose the bidirectional incongruity error. By incorporating the error into the CF learning, the BiCF tracker is presented to utilize the inter-frame information more comprehensively and efficiently.

III. Proposed Method

A. Bidirectional incongruity modeling

Figure 2 illustrates the bidirectional incongruity error, which contains the forward tracking error $\Delta_F\mathbf{R}$ and historical backtrace error $\Delta_B\mathbf{R}$. Given the frame #k, we extract the sample feature $\mathbf{x}$ and train the filter $\mathbf{w}$ with the label function $\mathbf{y}$ [10]. It is expected that the response $\mathbf{R}^{\mathbf{x}}_{\mathbf{w}}$ to the current feature $\mathbf{x}$ is close to the desirable response distribution $\mathbf{y}$. Intuitively, the current response $\mathbf{R}^{\mathbf{x}}_{\mathbf{w}}$ should be more consistent with the detection response $\mathbf{R}^{\mathbf{x}}_{\mathbf{w}^p}$ obtained by the previous filter $\mathbf{w}^p$, which is applied to search the object position in the frame #k. However, in actual situation, the responses both $\mathbf{R}^{\mathbf{x}}_{\mathbf{w}}$ and $\mathbf{R}^{\mathbf{x}}_{\mathbf{w}^p}$ tend to be different due to viewpoint variations, illumination changes, and other factors. In this work, the error between these two responses is defined as the forward tracking error $\Delta_F\mathbf{R}$:

$$\Delta_F\mathbf{R} = \mathbf{R}^{\mathbf{x}}_{\mathbf{w}} - \mathbf{R}^{\mathbf{x}}_{\mathbf{w}^p} \ . \tag{1}$$

Moreover, an ideal filter can trackback to the object's original position. That is, the backtrace response $\mathbf{R}^{\mathbf{x}^p}_{\mathbf{w}}$ of the filter $\mathbf{w}$ to the feature $\mathbf{x}^p$ is similar to the response $\mathbf{R}^{\mathbf{x}^p}_{\mathbf{w}^p}$. However, there still exists a difference between the two responses in real-world scenarios, which is denoted as the backward trace error $\Delta_B\mathbf{R}$:

$$\Delta_B\mathbf{R} = \mathbf{R}^{\mathbf{x}^p}_{\mathbf{w}^p} - \mathbf{R}^{\mathbf{x}^p}_{\mathbf{w}} \ . \tag{2}$$

By analyzing the tracking process from different angles, it can be found that there are inconsistent errors for the same appearance changes in consecutive frames. In other words, the forward and backward tracking errors are consistent in theory, but in real scenes, there exists a difference in the reversibility of the tracking process. Thus the difference between the two errors ($\Delta_F\mathbf{R}$ and $\Delta_B\mathbf{R}$) reflects the bidirectional incongruity during the tracking process and contains information about the variations in appearance. Therefore, the bidirectional incongruity error ϵ is introduced so that the filter can sense the changes within two consecutive frames:

$$\epsilon = \|\Delta_F\mathbf{R} - \Delta_B\mathbf{R}\|_2^2 \ . \tag{3}$$

B. Objective function of BiCF

In this work, the bidirectional incongruity error is incorporated into the CF learning. The optimization problem are formulated as:

$$\mathbf{W} = \arg\min_{\mathbf{W}} \mathcal{E}(\mathbf{W}; \mathbf{X}, \mathbf{W}^p, \mathbf{X}^p) \ , \tag{4}$$

where $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_D]$ and $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_D]$ respectively denotes the filter and feature with all D channels concatenated together in the k-th frame. $\mathbf{W}^p = [\mathbf{w}_1^p, \ldots, \mathbf{w}_D^p]$ and $\mathbf{X}^p = [\mathbf{x}_1^p, \ldots, \mathbf{x}_D^p]$ respectively represents the filter and feature obtained from the previous training sample, i.e., the sample in $(k-1)$-th frame. For clarity, $\mathbf{x}_d \in \mathbb{R}^{N\times 1}$ is assumed as a one-dimensional signal of length N in the following derivation, which can be directly generalized to the two-dimensional image. The objective function $\mathcal{E}(\mathbf{W}; \mathbf{X}, \mathbf{W}^p, \mathbf{X}^p)$ is defined as follows:

$$\mathcal{E}(\mathbf{W}; \mathbf{X}, \mathbf{W}^p, \mathbf{X}^p) = \mathcal{E}_1(\mathbf{W}; \mathbf{X}) + \mathcal{E}_2(\mathbf{W}; \mathbf{X}, \mathbf{W}^p, \mathbf{X}^p) \ . \tag{5}$$

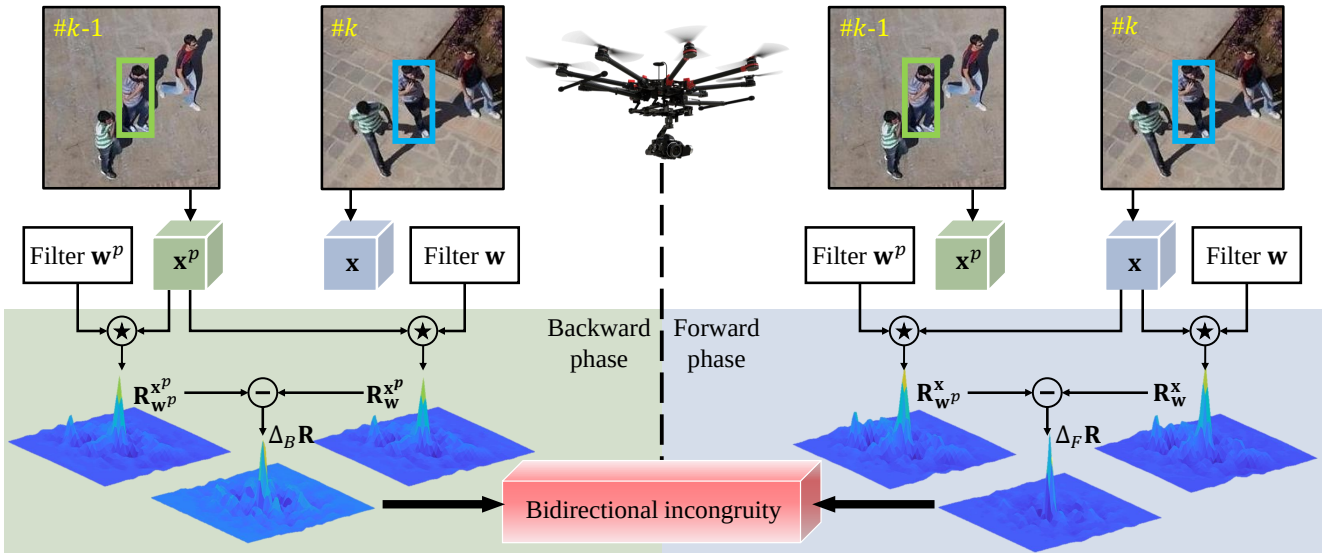


Fig. 2. A flowchart of the proposed BiCF tracker. In the forward phase, the existence of the forward tracking error is due to the difference between the current and detected responses, $\mathbf{R}^{\mathbf{x}}_{\mathbf{w}}$ and $\mathbf{R}^{\mathbf{x}}_{\mathbf{w}^p}$. In the backward phase, the tracking error is also caused by the inconsistency of the responses. For the same appearance changes in consecutive frames, the forward and backward tracking errors are consistent in theory, but in real scenes, there exists a difference in the reversibility of the tracking process. Therefore, the bidirectional incongruity during the tracking process is reflected by the two errors, i.e., $\Delta_F\mathbf{R}$ and $\Delta_B\mathbf{R}$. Note that the channel index $(\cdot)_d$ is ignored here for clarity of the notations.

The first term $\mathcal{E}_1$ is the error of the ridge regression between the label function $\mathbf{y}$ and the response reflected by the feature $\mathbf{X}$ extracted from the current training sample:

$$\mathcal{E}_1(\mathbf{W};\mathbf{X}) = \sum_{d=1}^{D} \left\|\mathbf{y} - \mathbf{R}^{\mathbf{x}_d}_{\mathbf{w}_d}\right\|_2^2 + \lambda \sum_{d=1}^{D} \|\mathbf{s}\odot\mathbf{w}_d\|_2^2 \ , \quad (6)$$

where $\mathbf{s}$ is the spatial regularizer [15] and λ denotes the regularization parameter. The response in the d-th channel of the filter $\mathbf{w}_d$ to the feature $\mathbf{x}_d$ is computed by:

$$\mathbf{R}^{\mathbf{x}_d}_{\mathbf{w}_d} = \mathbf{w}_d \star \mathbf{x}_d \ , \quad (7)$$

where $\star$ denotes circular correlation operator.

The second term $\mathcal{E}_2$ introduces the bidirectional incongruity error ϵ to help construct the filter that can learn the changes in successive frames:

$$\mathcal{E}_2(\mathbf{W};\mathbf{X},\mathbf{W}^p,\mathbf{X}^p) = \gamma \sum_{d=1}^{D} \epsilon_d \ , \quad (8)$$

where γ is a regularization term. ϵ_d represents the bidirectional inconsistency error in d-th channel and can be calculated by:

$$\begin{aligned}\epsilon_d &= \|\Delta_F\mathbf{R}_d - \Delta_B\mathbf{R}_d\|_2^2 \\ &= \left\|(\mathbf{R}^{\mathbf{x}_d}_{\mathbf{w}_d} - \mathbf{R}^{\mathbf{x}_d}_{\mathbf{w}^p_d}) - (\mathbf{R}^{\mathbf{x}^p_d}_{\mathbf{w}^p_d} - \mathbf{R}^{\mathbf{x}^p_d}_{\mathbf{w}_d})\right\|_2^2 \\ &= \|(\mathbf{w}_d\star\mathbf{x}_d - \mathbf{w}^p_d\star\mathbf{x}_d) - (\mathbf{w}^p_d\star\mathbf{x}^p_d - \mathbf{w}_d\star\mathbf{x}^p_d)\|_2^2 \ .\end{aligned} \quad (9)$$

Therefore $\mathcal{E}_2$ is reformulated as:

$$\mathcal{E}_2(\mathbf{W};\mathbf{X},\mathbf{W}^p,\mathbf{X}^p) = \gamma \sum_{d=1}^{D} \|(\mathbf{w}_d - \mathbf{w}^p_d)\star(\mathbf{x}_d + \mathbf{x}^p_d)\|_2^2 \ . \quad (10)$$

Note that $\mathcal{E}(\mathbf{W};\mathbf{X},\mathbf{W}^p,\mathbf{X}^p)$ can be decomposed into D error terms $\mathcal{E}_d$ $(d = 1,\ldots,D)$ for optimization, since the filter is trained independently on each channel. In this work, the d-th channel is chosen for the following model derivation.

C. BiCF learning

To obtain the optimal filter $\mathbf{w}_d$ minimizing $\mathcal{E}_d$, we introduce an auxiliary variable $\mathbf{h}_d \in \mathbb{R}^{N\times 1}$ and requiring $\mathbf{w}_d = \mathbf{h}_d$ so that the optimization problem can be decomposed into several subproblems and solved iteratively by the ADMM technique [21]. Thus $\mathcal{E}_d$ can be equivalently written as the equality constraint form:

$$\begin{aligned}\mathcal{E}_d(\mathbf{w}_d,\mathbf{h}_d) = &\|\mathbf{y} - \mathbf{w}_d\star\mathbf{x}_d\|_2^2 + \lambda\|\mathbf{S}\mathbf{h}_d\|_2^2 \\ &+ \gamma\|(\mathbf{w}_d - \mathbf{w}^p_d)\star(\mathbf{x}_d + \mathbf{x}^p_d)\|_2^2 \ , \\ s.t.\quad &\mathbf{w}_d = \mathbf{h}_d,\ d = 1,\ldots,D\end{aligned} \quad (11)$$

where $\mathbf{S} = \mathrm{diag}(\mathbf{s})$ denotes the diagonal matrix. For computational efficiency, Eq. (11) can be expressed in the Fourier domain by Parseval's theorem:

$$\begin{aligned}\mathcal{E}_d(\hat{\mathbf{w}}_d,\mathbf{h}_d) = &\|\hat{\mathbf{y}} - \hat{\mathbf{w}}^*_d\odot\hat{\mathbf{x}}_d\|_2^2 + \lambda\|\mathbf{S}\mathbf{h}_d\|_2^2 \\ &+ \gamma\left\|(\hat{\mathbf{w}}^*_d - \hat{\mathbf{w}}^{p*}_d)\odot(\hat{\mathbf{x}}_d + \hat{\mathbf{x}}^p_d)\right\|_2^2 \ , \\ s.t.\quad &\hat{\mathbf{w}}_d = \sqrt{N}\mathbf{F}\mathbf{h}_d,\ d = 1,\ldots,D\end{aligned} \quad (12)$$

where $\odot$ stands for the Hadamard product. The superscript $\hat{}$ and * is the discrete Fourier transform (DFT) of a signal and the conjugate of a complex vector, respectively. $\mathbf{F}\in\mathbb{C}^{N\times N}$ is the DFT matrix that transforms a signal $\mathbf{v}\in\mathbb{R}^{N\times 1}$ into the frequency domain, such that $\hat{\mathbf{v}} = \sqrt{N}\mathbf{F}\mathbf{v}$. Eq. (12) can be formulated as the augmented Lagrangian form:

$$\mathcal{L}(\hat{\mathbf{w}}_d,\mathbf{h}_d,\hat{\zeta}_d) = \mathcal{E}_d(\hat{\mathbf{w}}_d,\mathbf{h}_d) + \mu\left\|\hat{\mathbf{w}}_d - \sqrt{N}\mathbf{F}\mathbf{h}_d + \frac{1}{\mu}\hat{\zeta}_d\right\|_2^2 , \quad (13)$$

where $\hat{\zeta}_d \in \mathbb{C}^{N\times 1}$ is the Lagrangian multiplier in the d-th channel and μ denotes the penalty factor.

Then the ADMM technique [21] is applied to alternatively solve the following subproblems. The subproblems for solving $\hat{\mathbf{w}}_d$ and $\mathbf{h}_d$ both have closed-form solutions.

1) Subproblem $\hat{\mathbf{w}}_d$*:* If $\mathbf{h}_d$ and $\hat{\zeta}_d$ are fixed in Eq. (13), the optimal $\hat{\mathbf{w}}_d^{(i+1)}$ can be obtained by solving Eq. (14).

$$\begin{aligned}\hat{\mathbf{w}}_d^{(i+1)} = \arg\min_{\hat{\mathbf{w}}_d} \Bigg\{ & \|\hat{\mathbf{y}} - \hat{\mathbf{w}}_d^* \odot \hat{\mathbf{x}}_d\|_2^2 \\ & + \gamma \left\|(\hat{\mathbf{w}}_d^* - \hat{\mathbf{w}}_d^{p*}) \odot (\hat{\mathbf{x}}_d + \hat{\mathbf{x}}_d^p)\right\|_2^2 \\ & + \mu \left\|\hat{\mathbf{w}}_d - \sqrt{N}\mathbf{F}\mathbf{h}_d + \frac{1}{\mu}\hat{\zeta}_d\right\|_2^2 \Bigg\} .\end{aligned} \tag{14}$$

By taking the derivative with respect to $\hat{\mathbf{w}}_d^*$ to zero, we can get the solution for $\hat{\mathbf{w}}_d^{(i+1)}$:

$$\hat{\mathbf{w}}_d^{(i+1)} = \frac{\hat{\mathbf{x}}_d \odot \hat{\mathbf{y}}^* + \gamma(\hat{\mathbf{x}}_d + \hat{\mathbf{x}}_d^p) \odot (\hat{\mathbf{x}}_d^* + \hat{\mathbf{x}}_d^{p*}) \odot \hat{\mathbf{w}}_d^p + \mu\hat{\mathbf{h}}_d - \hat{\zeta}_d}{\hat{\mathbf{x}}_d \odot \hat{\mathbf{x}}_d^* + \gamma(\hat{\mathbf{x}}_d + \hat{\mathbf{x}}_d^p) \odot (\hat{\mathbf{x}}_d^* + \hat{\mathbf{x}}_d^{p*}) + \mu}, \tag{15}$$

where the fraction operator denotes element-wise division.

2) Subproblem $\mathbf{h}_d$*:* If $\hat{\mathbf{w}}_d$ and $\hat{\zeta}_d$ are given in Eq. (13), the optimal $\mathbf{h}_d^{(i+1)}$ can be solved by Eq. (16).

$$\mathbf{h}_d^{(i+1)} = \arg\min_{\mathbf{h}_d} \left\{ \lambda \|\mathbf{S}\mathbf{h}_d\|_2^2 + \mu \left\|\hat{\mathbf{w}}_d - \sqrt{N}\mathbf{F}\mathbf{h}_d + \frac{1}{\mu}\hat{\zeta}_d\right\|_2^2 \right\} . \tag{16}$$

The solution of $\mathbf{h}_d^{(i+1)}$ can be easily achieved by setting the derivation with respect to $\mathbf{h}_d$ to zero:

$$\mathbf{h}_d^{(i+1)} = \frac{\mathcal{F}^{-1}(\mu\hat{\mathbf{w}}_d + \hat{\zeta}_d)}{\frac{\lambda}{N}(\mathbf{s} \odot \mathbf{s}^*) + \mu} . \tag{17}$$

3) Updating Lagrangian multiplier $\hat{\zeta}_d$*:* The Lagrangian multiplier $\hat{\zeta}_d$ is updated by:

$$\hat{\zeta}_d^{(i+1)} = \hat{\zeta}_d^{(i)} + \mu\left(\hat{\mathbf{w}}_d^{(i+1)} - \hat{\mathbf{h}}_d^{(i+1)}\right) , \tag{18}$$

where $\hat{\mathbf{h}}_d^{(i+1)} = \sqrt{N}\mathbf{F}\mathbf{h}_d^{(i+1)}$. Within the i-th ADMM iteration, the factor μ is commonly updated as follows [21]:

$$\mu^{(i+1)} = \min(\mu_{\max}, \beta\mu^{(i)}) . \tag{19}$$

Algorithm 1: BiCF tracker

Input: Image: I_k.
The previous filter: $\mathbf{w}_d^p$.
The feature in frame #$k-1$: $\mathbf{x}_d^p$.
The spatial regularizer weights: $\mathbf{s}$.
Output: The current filter $\mathbf{w}_d$ in the k-th frame.

Extract features $\mathbf{x}_d$ from I_k.
Introduce the auxiliary variable $\mathbf{h}_d$ and build the equality constraint form Eq. (11).
Transform Eq. (11) to Eq. (12) by Parseval's theorem.
Initialize variables $\hat{\mathbf{w}}_d^{(0)}$, $\mathbf{h}_d^{(0)}$, and $\hat{\zeta}_d^{(0)}$.
for ADMM iteration $i = 1$ to *end* **do**
  Solve subproblem $\hat{\mathbf{w}}_d^{(i+1)}$ by Eq. (15).
  Solve subproblem $\mathbf{h}_d^{(i+1)}$ by Eq. (17).
  Update Lagrangian multiplier $\hat{\zeta}_d^{(i+1)}$ by Eq. (18).
  Update the penalty factor $\mu^{(i+1)}$ by Eq. (19).
end
Use Eq. (20) to update the appearance model.

Moreover, an online adaptive scheme is utilized to improve the filter's robustness, which can be formulated as:

$$\hat{\mathbf{x}}_{d,\text{model}} = (1-\eta)\hat{\mathbf{x}}_{d,\text{model}}^p + \eta\hat{\mathbf{x}}_d , \tag{20}$$

where $\hat{\mathbf{x}}_{d,\text{model}}$ and $\hat{\mathbf{x}}_{d,\text{model}}^p$ denotes the appearance model at the current frame and the previous frame, respectively. η is the online adaptation rate. The BiCF learning in the d-th channel in frame #k can be summarized in Algorithm 1.

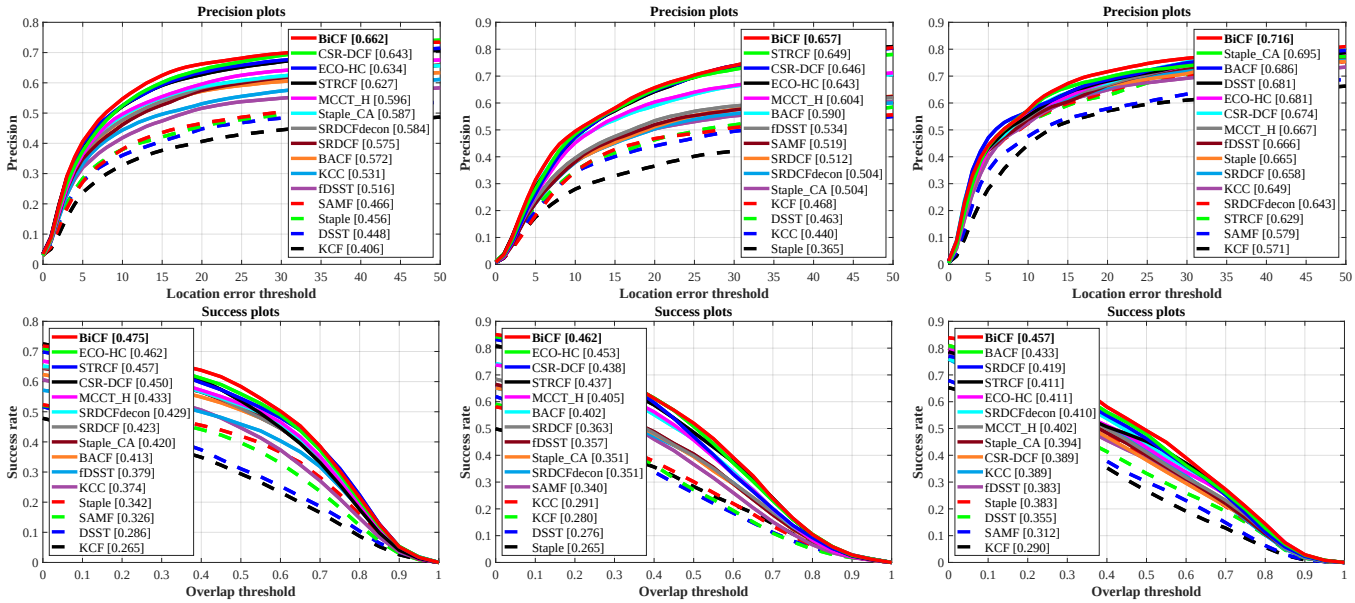
IV. Experiments

In this section, the proposed BiCF tracker is evaluated by considerable experiments on 243 challenging UAV image sequences from three datasets, which are widely used in UAV tracking, including UAV123@10fps [4], DTB70 [22], and UAVDT [23] datasets. The results are compared with 25 state-of-the-art tracking methods, such as KCF [10], DSST [12], SAMF [11], CF2 [24], C-COT [17], SRDCFdecon [25], Staple [9], BACF [14], CoKCF [26], CSR-DCF [27], ECO-HC, ECO [18], fDSST [28], IBCCF [13], MCPF [29], SRDCF [15], Staple_CA [30], KCC [31], MCCT-H [32], MCCT [32], STRCF, DeepSTRCF [19], TADT [33], UDT, and UDT+ [34].

A. Experimental setups

1) Evaluation metrics: The experiments are based on the one-pass evaluation, where two metrics including center location error (CLE) and success rate (SR) are used to evaluate all trackers on the UAV123@10fps, DTB70, and UAVDT datasets. CLE is used to measure the Euclidean distance between the estimated object location and the ground truth bounding box center. The precision plot shows the percentage of bounding boxes whose CLE is less than the given threshold. SR is applied to measure the intersection over union (IoU) of the estimated and the ground truth bounding box. The success plot indicates the ratio of the number of frames whose IoU is greater than the given threshold to the total number of frames. According to the common ranking metrics [4], CLE's threshold is set to 20 pixels to rank the precision of trackers, and the area under the curve (AUC) is applied to rank the success rate of trackers.

2) Implementation details: The proposed BiCF tracker is implemented in MATLAB R2018a on a computer with an i7-8700K CPU (3.7GHz) and an NVIDIA GeForce RTX 2080 GPU. The BiCF tracker uses a combination of hand-crafted features, including HOG [5], CN [6], and gray-scale features, for object representation. The regularization parameter λ and γ is set to 0.01 and 0.03, respectively. As for ADMM optimization parameters, the initial penalty factor μ, scale step β, and the maximum value $\mu_{\max}$ are set to 100, 10, and 10^5, respectively. The number of ADMM iterations is set to 4. All hyper-parameters remain fixed for all image sequences on three datasets. One MATLAB implementation and UAV tracking videos are available here: https://github.com/vision4robotics/BiCF-Tracker and https://youtu.be/fS12kosv37s.



(a) UAV123@10fps dataset (b) DTB70 dataset (c) UAVDT dataset

Fig. 3. Precision and success plots of BiCF and other 14 hand-crafted feature-based trackers on three datasets.

TABLE I

THE AVERAGE TRACKING SPEED OF BICF VERSUS OTHER HAND-CRAFTED FEATURE-BASED TRACKERS ON THREE DATASETS. THE TOP 3 TRACKING SPEED IS SHOWN IN RED, GREEN, AND BLUE FONTS. ALL RESULTS ARE GENERATED IN CPU MODE.

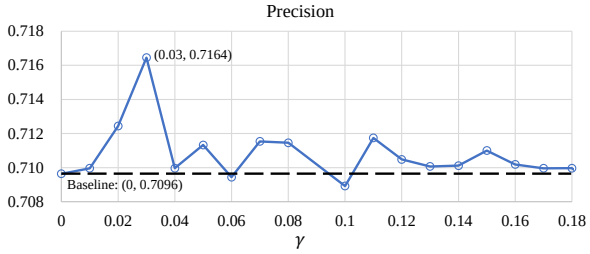
	KCF	DSST	BACF	SAMF	Staple	Staple_CA	SRDCF	SRDCFdecon	MCCT_H	CSR-DCF	STRCF	ECO-HC	fDSST	KCC	**BiCF**
Avg. FPS	651.1	106.5	56.0	12.8	65.4	58.9	14.0	7.5	59.7	12.1	28.5	69.3	168.1	46.1	**45.4**

B. Validity analysis of regularization factor γ

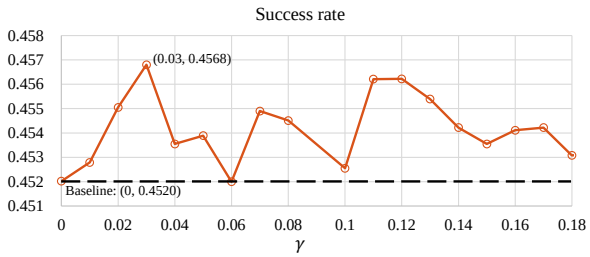
To shed light the effect of the bidirectional incongruity penalty factor γ on the overall performance, we test different numerical values of γ on the UAVDT dataset. γ values are set from 0 to 0.18 empirically for the trial, with a step size of 0.01. The results for precision and success rate are reported in Fig. 4. The precision corresponding to $\gamma = 0$ is used as the baseline, which is the black dash line in Fig. 4(a). The performance gradually increases with the increase of γ, and reaches the highest point (0.7164) at $\gamma = 0.03$. After that, the precision score decreases slightly and fluctuates around the baseline. For the success rate in Fig. 4(b), the performance sees the trend similar to precision and achieves the best score (0.4568) at $\gamma = 0.03$. Compared to the performance at $\gamma = 0$, the precision and success rate obtain a gain of 0.96% and 1.06% respectively when $\gamma = 0.03$. The results show that when γ is set in a certain range, the bidirectional incongruity regularization term can effectively improve the overall performance of the tracker.

C. Comparison with hand-crafted feature-based trackers

1) Overall performance comparison: The proposed BiCF tracker is compared with other trackers using hand-crafted features on the UAV123@10fps, DTB70, and UAVDT datasets. Fig. 3 shows that the BiCF tracker performs significantly better than other hand-crafted feature-based trackers on all three datasets. More specifically, the BiCF tracker achieves the best precision score (0.662) on UAV123@10fps, exceeding CSR-DCF (0.643) and ECO-HC (0.634) by 2.9% and 4.4%, respectively. BiCF also achieves the best AUC score (0.475) on UAV123@10fps, which leads to a significant gain of 2.9% and 4.0% compared to ECO-HC

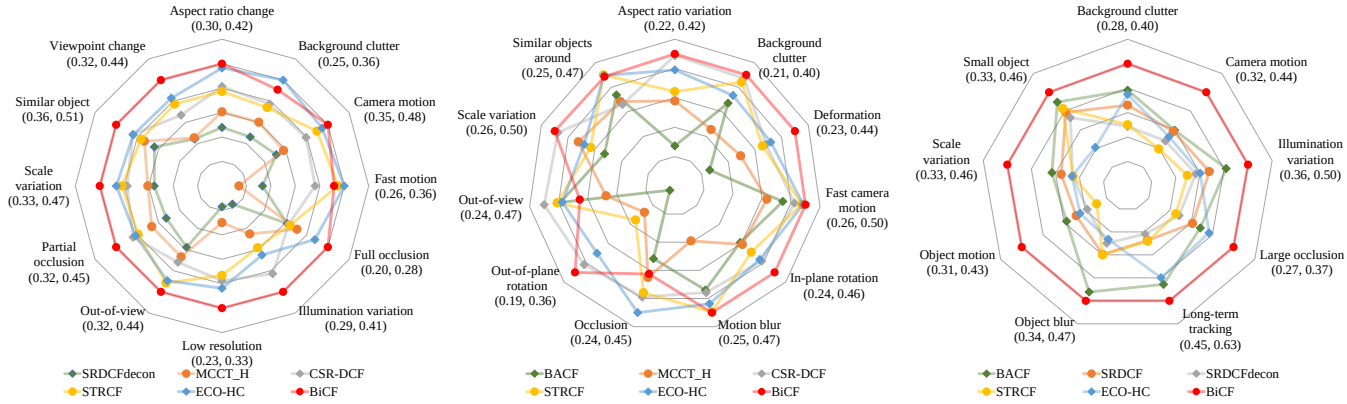


(a) Precision (at CLE = 20 pixels) under different values of γ.



(b) Success rate (at AUC score) under different values of γ.

Fig. 4. Different values of bidirectional incongruity factor γ are tested on UAVDT dataset. At $\gamma = 0.03$, both the precision and success rate reach the highest scores.



(a) UAV123@10fps dataset (b) DTB70 dataset (c) UAVDT dataset

Fig. 5. Attribute-based evaluation between BiCF and other hand-crafted feature-based trackers. The numerical interval of the attribute axis is displayed below the attribute name.

(0.462) and STRCF (0.457). On DTB70, BiCF provides best precision score (0.657). Moreover, the best AUC score (0.462) is also obtained by BiCF, followed by ECO-HC (0.453) and CSR-DCF (0.438). On the UAVDT dataset, BiCF provides the best precision score (0.716) compared to Staple_CA (0.695) and BACF (0.686). BiCF also achieves the best AUC score (0.457), followed by BACF (0.433) and SRDCF (0.419). In addition to excellent tracking performance, the speed of the proposed BiCF tracker (45.4 FPS) is sufficient for UAV real-time tracking, as shown in Table I. Although KCF obtains the best tracking speed (651.1 FPS), followed by fDSST (168.1 FPS) and DSST (106.5 FPS), their precision and AUC scores are much lower than BiCF.

2) Attribute-based evaluation: The image sequences on the UAV123@10fps, DTB70, and UAVDT datasets are annotated with 12, 11, and 9 different attributes, respectively. We compare the proposed BiCF tracker with the other 14 hand-crafted feature-based trackers in all attributes. Fig. 5 illustrates the attribute-based comparisons based on the AUC score on each dataset. Note that only the top 6 trackers in the AUC score on each dataset are compared with the BiCF tracker. The experimental results demonstrate that BiCF outperforms favorably than other competing trackers in most attributes, such as aspect ratio changes, camera/object motion, viewpoint changes, scale variations, to name a few. The results empirically demonstrate the bidirectional incongruity can be applied to learn a robust filter efficiently to counteract the object or background appearance variations during tracking.

D. Comparison with deep-based trackers

Table II presents the precision (% at CLE = 20 pixels) and success rate (% at AUC score) of the proposed BiCF tracker and other deep-based trackers on the UAVDT dataset. These deep-based trackers represent trackers that rely on deep features or pre-trained deep networks. Note that only the top 7 trackers on UAVDT performance are used for comparison. The results demonstrate that BiCF achieves the best precision (0.716) and obtains 2.4% and 2.8% gain respectively than ECO (0.700) and UDT+ (0.697). In terms of success rate, BiCF obtains the best AUC score of 0.457,

TABLE II

TRACKING PERFORMANCE AND SPEED COMPARISONS OF TOP 7 TRACKERS ON UAVDT. **RED** FONT INDICATES THE BEST RESULTS. ALL RESULTS ARE GENERATED BY THE SAME COMPUTER.

	ECO	UDT+	TADT	UDT	MCCT	DeepSTRCF	**BiCF**
Prec.	70.0	69.7	67.7	67.4	67.1	66.7	**71.6**
Succ.	45.4	41.6	43.1	44.1	43.7	43.7	**45.7**
FPS	16.4	**60.4**	32.5	76.4	8.6	6.6	50.2
GPU	✓	✓	✓	✓	✓	✓	✗

which outperforms ECO (0.454) and UDT(0.441) by 0.5% and 3.4% respectively. Table II also gives the tracking speed of competing trackers on the UAVDT dataset. UDT obtains the best tracking speed (76.4 FPS), followed by UDT+ (60.4 FPS). The higher speed of these trackers benefits from the GPU, while the tracking performance is lower than BiCF both in precision and success rate, which runs at a single CPU with a real-time speed of 50.2 FPS.

V. CONCLUSIONS

In this work, a bidirectional incongruity-aware correlation filter, i.e., BiCF, is proposed to perform various types of UAV object tracking tasks. By incorporating bidirectional response-based error into the CF learning, the novel tracker can sense the inter-frame information on the appearance changes. Considerable experiments are conducted to verify the proposed approach on 243 challenging aerial sequences from three datasets. Extensive experimental results show that the presented BiCF tracker outperforms favorably against 25 state-of-the-art tracking methods in accuracy, robustness, and efficiency. Moreover, BiCF has the advantage of a small computational burden and is suitable for performing real-time UAV tracking missions. The results of BiCF will further extend the development of bidirectional incongruity suppression strategy in UAV visual tracking applications.

ACKNOWLEDGMENT

The work was supported by the National Natural Science Foundation of China under Grant 61806148 and the State Key Laboratory of Mechanical Transmissions (Chongqing University) under Grant SKLMT-KFKT-201802.

REFERENCES

[1] M. Monajjemi, J. Bruce, S. A. Sadat, J. Wawerla, and R. Vaughan, "Uav, do you see me? establishing mutual attention between an uninstrumented human and an outdoor uav in flight," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2015, pp. 3614–3620.

[2] S. Lin, M. A. Garratt, and A. J. Lambert, "Monocular vision-based real-time target recognition and tracking for autonomously landing an uav in a cluttered shipboard environment," *Autonomous Robots*, vol. 41, no. 4, pp. 881–901, 2017.

[3] M. Mueller, G. Sharma, N. Smith, and B. Ghanem, "Persistent aerial tracking system for uavs," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2016, pp. 1562–1569.

[4] M. Mueller, N. Smith, and B. Ghanem, "A benchmark and simulator for uav tracking," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016, pp. 445–461.

[5] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 1, 2005, pp. 886–893.

[6] J. V. de Weijer, C. Schmid, J. J. Verbeek, and D. Larlus, "Learning color names for real-world applications," *IEEE Transactions on Image Processing*, vol. 18, pp. 1512–1523, 2009.

[7] D. S. Bolme, J. R. Beveridge, B. A. Draper, and Y. M. Lui, "Visual object tracking using adaptive correlation filters," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010, pp. 2544–2550.

[8] M. Danelljan, F. Shahbaz Khan, M. Felsberg, and J. Van de Weijer, "Adaptive color attributes for real-time visual tracking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 1090–1097.

[9] L. Bertinetto, J. Valmadre, S. Golodetz, O. Miksik, and P. H. Torr, "Staple: Complementary learners for real-time tracking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 1401–1409.

[10] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista, "High-Speed tracking with kernelized correlation filters," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 3, pp. 583–596, 2015.

[11] Y. Li and J. Zhu, "A scale adaptive kernel correlation filter tracker with feature integration," in *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, 2014, pp. 254–265.

[12] M. Danelljan, G. Häger, F. Khan, and M. Felsberg, "Accurate scale estimation for robust visual tracking," in *Proceedings of the British Machine Vision Conference (BMVC)*, 2014.

[13] 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 *Proceedings of the IEEE International Conference on Computer Vision (ICCV) Workshops*, 2017.

[14] H. Kiani Galoogahi, A. Fagg, and S. Lucey, "Learning background-aware correlation filters for visual tracking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1135–1143.

[15] M. Danelljan, G. Häger, F. Shahbaz Khan, and M. Felsberg, "Learning spatially regularized correlation filters for visual tracking," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 4310–4318.

[16] C. Fu, Z. Huang, Y. Li, R. Duan, and P. Lu, "Boundary effect-aware visual tracking for uav with online enhanced background learning and multi-frame consensus verification," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2019.

[17] M. Danelljan, A. Robinson, F. S. Khan, and M. Felsberg, "Beyond correlation filters: Learning continuous convolution operators for visual tracking," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016, pp. 472–488.

[18] M. Danelljan, G. Bhat, F. Shahbaz Khan, and M. Felsberg, "Eco: efficient convolution operators for tracking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 6638–6646.

[19] F. Li, C. Tian, W. Zuo, L. Zhang, and M. H. Yang, "Learning spatial-temporal regularized correlation filters for visual tracking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[20] Z. Huang, C. Fu, Y. Li, F. Lin, and P. Lu, "Learning aberrance repressed correlation filters for real-time uav tracking," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019.

[21] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," in *Foundations and Trends in Machine Learning*, vol. 3, 2010, pp. 1–122.

[22] S. Li and D.-Y. Yeung, "Visual object tracking for unmanned aerial vehicles: A benchmark and new motion models," in *AAAI*, 2017.

[23] D. Du, Y. Qi, H. Yu, Y. Yang, K. Duan, G. Li, W. Zhang, Q. Huang, and Q. Tian, "The unmanned aerial vehicle benchmark: Object detection and tracking," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 370–386.

[24] C. Ma, J.-B. Huang, X. Yang, and M.-H. Yang, "Hierarchical convolutional features for visual tracking," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 3074–3082.

[25] M. Danelljan, G. Häger, F. Shahbaz Khan, and M. Felsberg, "Adaptive decontamination of the training set: A unified formulation for discriminative visual tracking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 1430–1438.

[26] L. Zhang and P. N. Suganthan, "Robust visual tracking via co-trained kernelized correlation filters," *Pattern Recognition*, vol. 69, pp. 82–93, 2017.

[27] A. Lukežič, T. Voj'iř, L. Čehovin Zajc, J. Matas, and M. Kristan, "Discriminative correlation filter with channel and spatial reliability," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[28] M. Danelljan, G. Häger, F. S. Khan, and M. Felsberg, "Discriminative scale space tracking," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 8, pp. 1561–1575, 2016.

[29] T. Zhang, C. Xu, and M.-H. Yang, "Multi-task correlation particle filter for robust object tracking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 4335–4343.

[30] M. Mueller, N. Smith, and B. Ghanem, "Context-aware correlation filter tracking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1396–1404.

[31] C. Wang, L. Zhang, L. Xie, and J. Yuan, "Kernel cross-correlator," in *AAAI Conference on Artificial Intelligence*, 2018, pp. 4179–4186.

[32] N. Wang, W. Zhou, Q. Tian, R. Hong, M. Wang, and H. Li, "Multi-cue correlation filters for robust visual tracking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 4844–4853.

[33] X. Li, C. Ma, B. Wu, Z. He, and M.-H. Yang, "Target-aware deep tracking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 1369–1378.

[34] N. Wang, Y. Song, C. Ma, W. Zhou, W. Liu, and H. Li, "Unsupervised deep tracking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.