Learning Temporary Block-Based Bidirectional Incongruity-Aware Correlation Filters for Efficient UAV Object Tracking

Fuling Lin[1], Changhong Fu[1,*], Yujie He[2], Fuyu Guo[3], and Qian Tang[3]

***Abstract*—In the field of UAV object tracking, correlation filter based approaches have received lots of attention due to their computational efficiency. The methods learn filters by the ridge regression and generate response maps to distinguish the specified target from the background. An ideal filter can predict the object's position in a new frame, and in turn, can backtrack the object in the past frames. However, the neglect of tracking reversibility in most methods limits the potential of using inter-frame information to improve performance. In this work, a novel bidirectional incongruity-aware correlation filter is presented based on the nature of tracking reversibility. The proposed method incorporates the response-based bidirectional incongruity, which represents the gap between the filters' discriminative difference in the forward and backward tracking perspective caused by object appearance changes. It enables the filter not only to inherit the discriminability from previous filters but also to enhance the generalization capability to unpredictable appearance variations in upcoming frames. Moreover, a temporary block-based strategy is introduced to empower the filter accommodate more drastic object appearance changes and make more effective use of inter-frame information. Comprehensive experiments are conducted on three challenging UAV tracking benchmarks, including UAV123@10fps, DTB70, and UAVDT. Experimental results indicate that the proposed method has superior performance compared with the other 34 state-of-the-art trackers. Our approach permits real-time performance at ∼46.8 FPS on a single CPU and is suitable for UAV online tracking applications.**

***Index Terms*—Aerial video analysis, unmanned aerial vehicle, visual object tracking, discriminative correlation filter, temporary block-based bidirectional incongruity**

I. Introduction

UNMANNED aerial vehicles (UAVs) with visual perception systems have been widely used in various scenarios, such as autonomous landing [2], object following [3], infrastructure inspection [4], and human-computer interaction [5]. In UAV object tracking, the objective is to predict the specified target's location and size over the entire sequence captured by an onboard camera. The problem is particularly challenging because only the initial state including the target's location and size is known. Despite the considerable progress made in object tracking in recent years, there still leave many challenges in UAV tracking scenarios, such as limited computing capability, restricted power source, large viewpoint changes, as well as fast object and UAV motion.

Recently, the discriminative correlation filter (CF) based tracking approach has received widespread attention due to its high computational efficiency and excellent performance. As a tracking-by-detection method, the CF-based approach learns a discriminative correlation filter by minimizing the squared error between the expected and actual correlation responses. The learned filter can be used to estimate the object's location in each new frame. The computational efficiency of CFs originates from the correlation computation in the Fourier domain at both training and detection stages. Recent works further improve CFs' performance by applying kernel trick [6], [7], multiple hand-crafted features [8]–[12], boundary effects suppression [13]–[17], joint reliability learning [18]–[20], deep features from convolutional neural networks [11], [21]–[26], and deep learning strategies [27]–[30]. Most CF-based trackers are learned using only intra-frame information and ignore the inter-frame information that can maintain tracking reversibility. The reversibility represents that an ideal filter can not only predict the object's location in a new frame but also backtrack the location in the previous frames. In this work, we investigate the problem of incorporating the nature of inter-frame based tracking reversibility into CF learning for robust UAV tracking. In many UAV tracking scenes, the object appearance frequently changes due to the fast motion of the object and the UAV. It leads to the suboptimal discriminative power of CFs that only use intra-frame information for training. To exploit the limited information in the tracking process, a straightforward method is to incorporate all historical samples into CF learning [13]. However, this *sample-level* scheme has high computational requirements and is not fit for real-time UAV tracking applications. Another strategy is based on the *filter-level* and utilizes the temporal regularization term [25] to approximately replace multiple historical training samples. Such an approximation, which aims to make the filter more similar to the previous one, results in suboptimal generalization ability of the learned model and limits the performance when encountering drastic appearance changes. Since the inference of the object's location relies on the filter's response to the new frame, the *response-level*

Preliminary results of this work were introduced in [1].

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

[2]Yujie He is with the School of Engineering, École polytechnique fédérale de Lausanne, 1015, Lausanne, Switzerland.

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

*Corresponding author

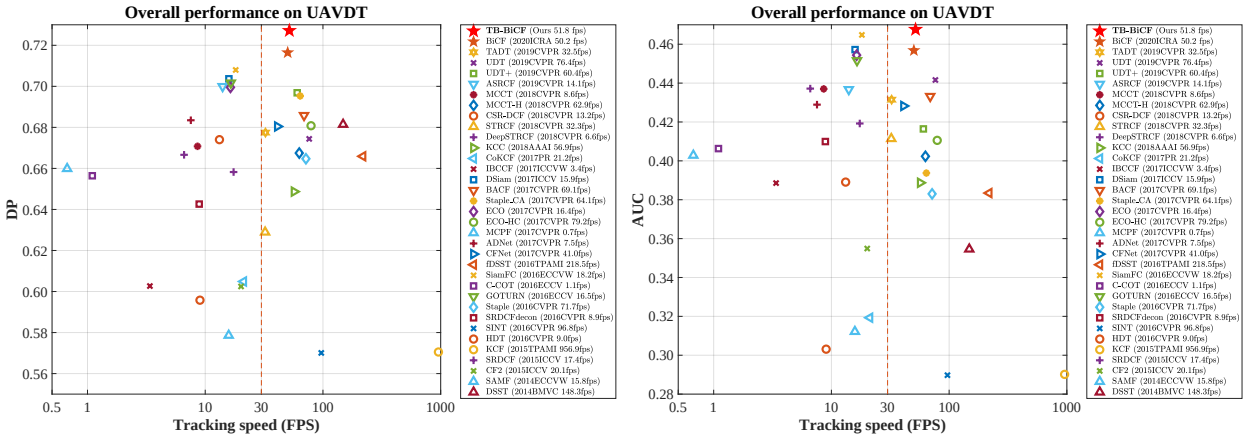


Fig. 1. Plots of DP-based precision versus tracking speed and AUC-based success rate versus tracking speed of the proposed TB-BiCF tracker and other 34 state-of-the-art trackers on the challenging UAVDT [31] benchmark. Results show that TB-BiCF achieves the impressive performance in the precision and success rate, and permits a real-time tracking speed over 30 FPS, *i.e.*, the red dash line in the figure. Note that the FPS axis is in the log10 scale.

inter-frame information is considered in this work, which contributes to building the bidirectional restraint component from the perspective of tracking reversibility.

In this paper, a novel temporary block-based bidirectional incongruity-aware correlation filter (TB-BiCF) is proposed for UAV object tracking. We systematically dissect the tracking reversibility and present the *response-level* bidirectional incongruity that contains inter-frame information about changes in both the object's appearance and the filter's discriminability. The bidirectional incongruity represents the gap of filters' discriminative difference between the forward detection and backward relocation. By incorporating the bidirectional incongruity into CF learning, a novel incongruity-aware optimization problem is formulated. It enables the filter to develop the discriminant ability inheritance from previous filters as well as generalization ability enhancement against significant appearance changes. Thus the learned filter can be more robust and accurate in the challenging UAV tracking scenarios. Moreover, the introduction of the temporary block-based strategy allows the filter to consider the incongruity between frames at a greater interval instead of only previous and current frames. The inter-frame information within the temporary block enables the filter to resist severe appearance changes better, and further improves the robustness without sacrificing efficiency.

Considering the nature of UAV online tracking, a computationally efficient learning strategy is of vital importance. Therefore, we decompose the novel incongruity-aware optimization problem into several subproblems and apply the alternating direction method of multipliers (ADMM) [32] to solve them iteratively. The resultant closed-form solutions containing element-wise operations allow the TB-BiCF to perform efficiently at real-time speed. Additionally, we propose to apply multiple hand-crafted features to achieve comprehensive representations of the object and background appearance, including the histogram of oriented gradients (HOG) [33], pixel intensity, and color names (CN) [34].

To validate the performance of the proposed TB-BiCF, we perform both quantitative and qualitative experiments on three challenging UAV benchmark datasets: UAV123@10fps [35] with 123 videos, DTB70 [36] with 70 videos, and UAVDT [31] with 50 videos. The total frame number from all benchmarks exceeds 90K. The TB-BiCF tracker demonstrates the outstanding results compared with other advanced methods using the hand-crafted features on the three benchmarks. We further show that the TB-BiCF tracker outperforms the other trackers based on deep features or end-to-end learning with a high tracking speed on a single CPU. Figure 1 shows the overall results of our approach and other trackers on the UAVDT benchmark and demonstrates that the TB-BiCF tracker is suitable for real-time UAV tracking tasks due to its superior performance in precision, success rate, and efficiency.

In this work, we make the following main contributions:

- A novel *response-level* bidirectional incongruity-aware CF is proposed to achieve the trade-off between inheriting the discriminant ability from previous filters and enhancing the generalization ability against appearance changes.
- A temporary block-based inter-frame information is presented to make the filter accommodate more significant appearance changes and further improves the filter's generalization ability.
- A new filter learning problem with the temporary block-based bidirectional incongruity is formulated and optimized by the ADMM technique where each subproblem has closed-form solutions with element-wise operations, resulting in computational efficiency.
- The proposed approach outperforms favorably other state-of-the-art trackers on three challenging UAV benchmarks over 90K images and reaches an average speed of ∼46.8 frames per second (FPS) using a single CPU, which is suitable for real-time UAV applications.

The remaining parts of this paper are organized as follows. Section II gives related works. Section III presents the proposed temporary block-based bidirectional incongruity-aware correlation filter (TB-BiCF). Experimental results are shown in Section IV and conclusions are given in Section V.

II. Related works

In this section, we discuss tracking methods that are most relevant to our work, *i.e*., tracking with CFs, tracking using historical information including sample-level, filter-level, and response-level information, and methods based on the tracking reversibility.

A. Tracking with CFs

As a discriminative tracking method, CF-based methods aim at differentiating the object from the background. Starting with the MOSSE tracker proposed by D. S. Bolme *et al*. [37], the CF-based method has achieved widespread attention with a high tracking speed and gained popularity in the tracking community. The high computational efficiency of CF-based methods originates from the use of the fast Fourier transformation (FFT). It transforms operations in the spatial domain into operations in the Fourier domain, *i.e*., from complex circular correlation to element-wise operations. J. F. Henriques *et al*. investigated the circulant structure of the dense sampled training patches [6], and further showed that the ridge regression on all training patches can be used equivalently to learn a CF [7], which can be incorporated with the kernel trick for more powerful regression. Several works [9], [38]–[40] were proposed to empower the CFs to estimate the scale variations and significantly improved the tracking performance in robustness. In [15], a background-aware CF (BACF) was proposed to handle the problem of lack of real negative samples in filter learning and alleviate the influence induced by boundary effects. Y. Sun *et al*. introduced the ROI-based pooling operation to the CF formula, which can help construct a more robust filter against the target deformation [41]. The use of multi-channel features [7], [8], [15] and the combination of multiple features [9]–[11] enrich the representations of the object appearance and enhance the discriminative power of CFs. In [42], [43], deep features extracted from the convolutional networks are applied for filter learning, but at the expense of tracking speed. Some tracking methods encourage the CF to focus on more reliable areas by using spatial information around the object [19], [44], and some learn the channel-wise weighting coefficients [18], [20] to emphasize the impact of important channels and to mitigate the effects of contaminated channels. However, the methods mentioned above focus more on how to make full use of the contents of the current frame, while ignoring to make an efficient use of the historical information that can help handle the inherent problem of limited samples.

B. Tracking using historical information

During the tracking process, the tracker knows only the initial state of the object in the first frame and needs to estimate the state in each new frame. In the situation with limited data, integrating the historical information into the current filter learning can increase the robustness and discriminability of the learned filter. In the CF-based approaches, the inter-frame information can be classified into the following categories, *i.e*., sample-level, filter-level, and response-level information.

1) Sample-level information: In [13], [45], the historical training samples are incorporated into the CF learning phase to improve the filter's discrimination. As time goes by, the number of training samples increases and brings a sizable computational burden, which is not fit for UAV online tracking applications. To mitigate the influence of samples corrupted by background clutters, occlusion, and other factors on the filter learning, M. Danelljan *et al*. [14] proposed a unified formulation to manage the training set dynamically. The method can simultaneously learn the filter and estimate the quality of training samples to alleviate the degradation of the discriminative power induced by corrupted samples. In [24], a Gaussian mixture model was employed to generate a compact model of the training set. Although the strategy can reduce the redundant samples in the training set and memory consumption, the large number of Gaussian components will also increase the computational load.

2) Filter-level information: F. Li *et al*. [25] introduced the temporal regularization to make the current filter more similar to the one in the previous frame, thus serving as an approximation to the effect of using multiple training samples in the CF learning. The approximation can confine any sudden change of the filter and thus improve tracking robustness. However, this approximation also constrains the changes in the filter caused by the object appearance changes and limits the generalization ability of the learned filter, which makes the tracker struggle in the case of significant appearance changes, especially in the complex UAV tracking scenarios with fast UAV/object motion.

3) Response-level information: Z. Huang *et al*. [46] presented the ARCF tracker using the response-level information to repress the response aberrance. Since the decision to locate the object is based on the correlation response containing the information of both samples and filters, the ARCF tracker used the detection and training correlation outputs to build the response-based regularization, which improves the credibility of the detection result in a new frame and reduces the risk of tracking drift. However, ARCF only utilizes the filter's detection power and ignores its ability to relocate in historical samples, which makes ARCF vulnerable to interference caused by contaminated detection response.

Different from the methods mentioned above, the proposed method makes better use of inter-frame information at the response level based on the nature of tracking reversibility. By dissecting the tracking process, we investigate the response-based bidirectional incongruity between the forward detection and backward relocation. Compared with the forward-backward error in [47], [48], the novel response-based bidirectional incongruity is based on the characteristics of correlation filters and is incorporated into the CF learning problem. Furthermore, the temporary block are applied to construct the bidirectional regularization to help the filter accommodate more significant appearance changes. The proposed improvements can well enhance the generalization power and inherit the discriminability from previous filters without sacrificing the tracking speed, which is suitable for UAV real-time and robust tracking.

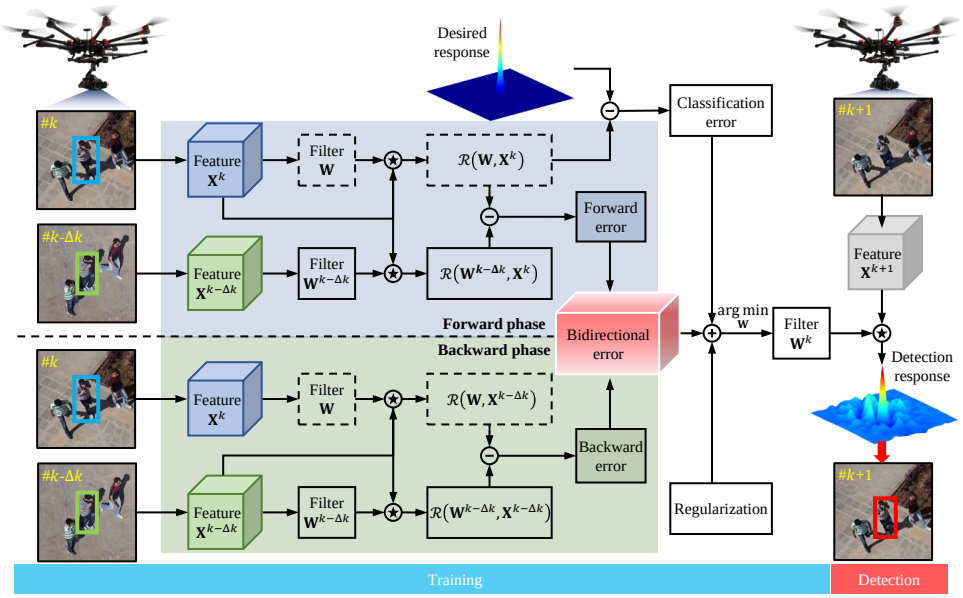


Fig. 2. A flowchart of the proposed TB-BiCF tracker. In the forward phase, the existence of the forward tracking error is resulted from the difference between the responses, $\mathcal{R}_d(\mathbf{W},\mathbf{X}^k)$ and $\mathcal{R}_d(\mathbf{W}^{k-\Delta k},\mathbf{X}^k)$. In the backward phase, the tracking error is also caused by the inconsistency of the responses between $\mathcal{R}_d(\mathbf{W}^{k-\Delta k},\mathbf{X}^{k-\Delta k})$ and $\mathcal{R}_d(\mathbf{W},\mathbf{X}^{k-\Delta k})$. Both forward and backward errors constitute the bidirectional incongruity, which reflects the gap of filters' discriminative difference. The figure is modified from our previous work [1].

III. Proposed Method

A. Correlation filter for visual tracking

1) Training stage: UAV object tracking aims to sequentially estimate the state of the specified object given the initial position and size. Based on the correlation filter based method [37], [38], the learned filter can be used to differentiate the object from the background. The filter learning problem in frame #k is to minimize the sum of squared error $\mathcal{E}$ between the desired response label $\mathbf{y} \in \mathbb{R}^N$ and the correlation output:

$$\mathcal{E}(\mathbf{W};\mathbf{X}^k) = \sum_{d=1}^{D} \|\mathbf{y} - \mathbf{w}_d \star \mathbf{x}_d^k\|_2^2 + \lambda \sum_{d=1}^{D} \|\mathbf{w}_d\|_2^2 \ , \quad (1)$$

where $\star$ is the circular correlation operator. $\mathbf{w}_d \in \mathbb{R}^N$ and $\mathbf{x}_d^k \in \mathbb{R}^N$ denote the filter and the vectorized feature map of the training sample in the d-th channel. λ is a regularization factor. Filters and feature maps across all channels are concatenated and grouped as $\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_D]$ and $\mathbf{X}^k = [\mathbf{x}_1^k, \mathbf{x}_2^k, \ldots, \mathbf{x}_D^k]$ for clarity. Thus the optimization problem is $\mathbf{W}^k = \arg\min_{\mathbf{W}} \mathcal{E}(\mathbf{W};\mathbf{X}^k)$. By transferring Eq. (1) from the spatial domain to the Fourier domain, an analytical solution to $\mathbf{W}^k$ can be obtained efficiently.

Remark 1: The feature maps $\mathbf{X}^k$ of the training sample are obtained by extracting the patch's feature in the search area. The size and location of the search area are determined by the size and location of the object in frame #k.

2) Detection stage: When the new frame #k+1 arrives, the learned filter is applied to build the response map for detection:

$$\mathcal{R}(\mathbf{W}^k,\mathbf{Z}^{k+1}) = \sum_{d=1}^{D} \mathcal{R}_d(\mathbf{W}^k,\mathbf{Z}^{k+1}) = \sum_{d=1}^{D} \mathbf{w}_d^k \star \mathbf{z}_d^{k+1} \ , \quad (2)$$

where $\mathbf{Z}^{k+1} = [\mathbf{z}_1^{k+1}, \mathbf{z}_2^{k+1}, \ldots, \mathbf{z}_D^{k+1}]$ is the feature maps of the detection sample, which is cropped from frame #k+1 on the basis of the object's location and size in frame #k. Thus the object's location can be derived by exploring the apogee of the detection response map $\mathcal{R}(\mathbf{W}^k,\mathbf{Z}^{k+1})$.

B. Temporary block-based bidirectional incongruity modeling

1) Bidirectional incongruity: Different from standard CFs, we introduce the *response-level* information on the basis of tracking reversibility and propose the bidirectional incongruity regularization term:

$$\|\Delta_F \mathcal{R}_d - \Delta_B \mathcal{R}_d\|_2^2 \ , \quad (3)$$

where $\Delta_F \mathcal{R}_d$ and $\Delta_B \mathcal{R}_d$ measures the filter's discriminate difference from the forward and backward aspects in the tracking process, respectively. $\Delta_F \mathcal{R}_d$ denotes the forward tracking error and is computed by:

$$\Delta_F \mathcal{R}_d = \mathcal{R}_d(\mathbf{W},\mathbf{X}^k) - \mathcal{R}_d(\mathbf{W}^{k-\Delta k},\mathbf{X}^k) \ . \quad (4)$$

It estimates the discriminate difference of the filter by using the feature maps $\mathbf{X}^k$ of the current training sample, as shown in the forward phase in Fig. 2. As for the historical backtrace error, it can be obtained by:

$$\Delta_B \mathcal{R}_d = \mathcal{R}_d(\mathbf{W}^{k-\Delta k},\mathbf{X}^{k-\Delta k}) - \mathcal{R}_d(\mathbf{W},\mathbf{X}^{k-\Delta k}) \ . \quad (5)$$

The backward error weights the filter's discriminability using the feature maps $\mathbf{X}^{k-\Delta k}$ of the historical sample, as shown in the backward phase in Fig. 2. In this work, if the tracking process begins with frame #$k-\Delta k$ and ends with #k, it is considered as a forward tracking process; if the tracking process is mirrored, that is, starting with frame #k and ending with #$k-\Delta k$, it is regarded as a backward tracking process. The tracking errors reflected by these two viewing aspects are forward and backward tracking errors respectively, *i.e.*, $\Delta_F \mathcal{R}_d$

and $\Delta_B\mathcal{R}_d$, which show the filter's discriminate difference between the first and the last frames in a tracking process.

Remark 2: The proposed method centers on the problem of inconsistent response in the CF framework, and tries to narrow the performance gap between the forward and backward tracking. In other words, the discriminate differences observed from two aspects are expected to be close. Therefore, the forward and backward tracking errors are correlated and a bidirectional incongruity regularization term $\|\Delta_F\mathcal{R}_d - \Delta_B\mathcal{R}_d\|_2^2$ is designed to strengthen the learned filter's robustness to the appearance changes.

2) Temporary block: Figure 3 shows the temporary block-based learning strategy. Each temporary block represents a sequence in which the object changes continuously over a period of time. By introducing the first and last frames in the temporary block into the bidirectional incongruity-aware learning, the influence of appearance changes on the filters' discriminate difference can be effectively weighted. Therefore, the suppression of the temporary block-based bidirectional incongruity can alleviate the degradation of the filter's discriminative ability caused by severe appearance changes.

Remark 3: In this work, the $k{-}\Delta k$-th and k-th frames denote the first and last frames within the temporary block at a Δk interval. Compared with applying all frames in the block, the proposed temporary block-based strategy can reduce the risk of overfitting and maintain efficient operational speed.

By analyzing the tracking process from different angles, it can be found that there exist error in the reversibility of the tracking process. Both $\Delta_F\mathcal{R}_d$ and $\Delta_B\mathcal{R}_d$ within the temporary block constitute the bidirectional incongruity term Eq. (3), which reflects the degree of variations in the information gap. Intuitively, measurements from different perspectives tend to be close since they both represents the same difference between the first and last frames within the temporary block.

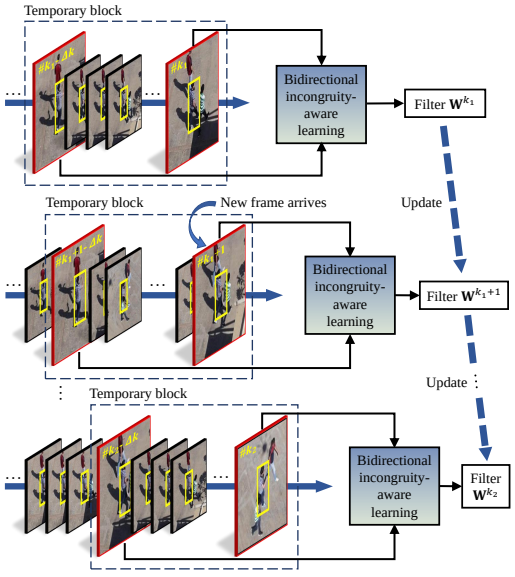


Fig. 3. A flowchart of the proposed temporary block-based filter learning strategy. Each temporary block represents a short sequence where the object changes continuously over a period of time. The first and last frames within the temporary block are applied to build the bidirectional incongruity regularization.

But drastic changes in appearance usually occur in UAV tracking scenarios, *e.g.*, fast object/UAV motion, and thus there exists the bidirectional incongruity between the two measurements. Therefore, suppressing the bidirectional inconsistency can reduce the influence of the significant appearance changes on the measurement results. In other words, the filter can achieve the trade-off between inheriting discriminant power from the previous filter and enhancing generalization power against the appearance changes.

Remark 4: The temporary block-based bidirectional regularization is different from the temporal regularization, *i.e.*, $\|\mathbf{W} - \mathbf{W}^{k-1}\|_2^2$, in STRCF [25]. Although the temporal regularization making the filter close to the previous one reduces the risk of tracking drift, the generalization ability of the filter is limited when the appearance variation is drastic. In contrast, the proposed regularization allows the filter to retain the characteristic of the previous one and have the generalization ability against large appearance changes.

C. Objective function of TB-BiCF

In this work, the bidirectional incongruity regularization term based on the temporary block is incorporated into the CF learning. The overall optimization problem is expressed as follows:

$$\begin{aligned}\mathbf{W}^k &= \arg\min_{\mathbf{W}} \mathcal{E}_{\mathrm{s}}(\mathbf{W}; \mathbf{X}^k, \mathbf{W}^{k-\Delta k}, \mathbf{X}^{k-\Delta k}) \\ &= \arg\min_{\mathbf{W}} \Big\{ \mathcal{E}_{\mathrm{c}}(\mathbf{W}; \mathbf{X}^k) \\ &\quad + \mathcal{E}_{\mathrm{t}}(\mathbf{W}; \mathbf{X}^k, \mathbf{W}^{k-\Delta k}, \mathbf{X}^{k-\Delta k}) + \mathcal{E}_{\mathrm{r}}(\mathbf{W}) \Big\} .\end{aligned} \tag{6}$$

1) Classification error term: The first term $\mathcal{E}_{\mathrm{c}}$ denotes the classification error between the correlation output and the predefined Gaussian label $\mathbf{y}$:

$$\mathcal{E}_{\mathrm{c}}(\mathbf{W}; \mathbf{X}^k) = \sum_{d=1}^{D} \left\| \mathbf{y} - \mathbf{w}_d \star \mathbf{x}_d^k \right\|_2^2 . \tag{7}$$

2) Temporary block-based bidirectional incongruity regularization term: The second term $\mathcal{E}_{\mathrm{t}}$ represents the proposed bidirectional incongruity regularization term:

$$\begin{aligned}\mathcal{E}_{\mathrm{t}}(\mathbf{W}; \mathbf{X}^k, \mathbf{W}^{k-\Delta k}, \mathbf{X}^{k-\Delta k}) &= \gamma \sum_{d=1}^{D} \|\Delta_F \mathbf{r}_d - \Delta_B \mathbf{r}_d\|_2^2 \\ &= \gamma \sum_{d=1}^{D} \|(\mathbf{w}_d \star \mathbf{x}_d^k - \mathbf{w}_d^{k-\Delta k} \star \mathbf{x}_d^k) \\ &\quad - (\mathbf{w}_d^{k-\Delta k} \star \mathbf{x}_d^{k-\Delta k} - \mathbf{w}_d \star \mathbf{x}_d^{k-\Delta k})\|_2^2 \\ &= \gamma \sum_{d=1}^{D} \left\| (\mathbf{w}_d - \mathbf{w}_d^{k-\Delta k}) \star (\mathbf{x}_d^k + \mathbf{x}_d^{k-\Delta k}) \right\|_2^2 ,\end{aligned} \tag{8}$$

where γ is a regularization factor.

3) Regularization term: The third term $\mathcal{E}_{\mathrm{r}}$ is used to constrains the complexity of the filter $\mathbf{W}$:

$$\mathcal{E}_{\mathrm{r}}(\mathbf{W}) = \lambda \sum_{d=1}^{D} \|\mathbf{s} \odot \mathbf{w}_d\|_2^2 , \tag{9}$$

where $\mathbf{s}$ is the spatial regularizer following [13] and λ denotes the regularization parameter.

Remark 5: Note that $\mathcal{E}_{\mathrm{s}}(\mathbf{W};\mathbf{X}^k,\mathbf{W}^{k-\Delta k},\mathbf{X}^{k-\Delta k})$ 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.

D. TB-BiCF learning

By introducing an auxiliary variable $\mathbf{h}_d \in \mathbb{R}^N$ and requiring $\mathbf{w}_d = \mathbf{h}_d$, $\mathcal{E}_d$ can be equivalently written as the equality constraint form:

$$\begin{aligned}\mathcal{E}_d(\mathbf{w}_d,\mathbf{h}_d) &= \left\|\mathbf{y}-\mathbf{w}_d\star\mathbf{x}_d^k\right\|_2^2+\lambda\left\|\mathbf{s}\odot\mathbf{h}_d\right\|_2^2 \\ &\quad+\gamma\left\|(\mathbf{w}_d-\mathbf{w}_d^{k-\Delta k})\star(\mathbf{x}_d+\mathbf{x}_d^{k-\Delta k})\right\|_2^2 , \\ &\textit{s.t.}\quad \mathbf{w}_d=\mathbf{h}_d,\ d=1,\ldots,D\ .\end{aligned} \tag{10}$$

The complicated correlation operation can be converted to the element-wise operation by transforming Eq. (10) from the spatial domain into the Fourier domain, thus improving computational efficiency:

$$\begin{aligned}\mathcal{E}_d(\hat{\mathbf{w}}_d,\mathbf{h}_d) &= \left\|\hat{\mathbf{y}}-\hat{\mathbf{w}}_d^*\odot\hat{\mathbf{x}}_d^k\right\|_2^2+\lambda\left\|\mathbf{s}\odot\mathbf{h}_d\right\|_2^2 \\ &\quad+\gamma\left\|(\hat{\mathbf{w}}_d^*-\hat{\mathbf{w}}_d^{k-\Delta k*})\odot(\hat{\mathbf{x}}_d^k+\hat{\mathbf{x}}_d^{k-\Delta k})\right\|_2^2 , \\ &\textit{s.t.}\quad \hat{\mathbf{w}}_d=\sqrt{N}\mathbf{F}\mathbf{h}_d,\ d=1,\ldots,D\ ,\end{aligned} \tag{11}$$

where $\odot$ stands for the Hadamard product. The superscript $\hat{}$ and * are 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 $\mathbf{v}\in\mathbb{R}^N$ into the frequency domain, such that $\hat{\mathbf{v}}=\sqrt{N}\mathbf{F}\mathbf{v}$. Eq. (11) can be formulated as the augmented Lagrangian form:

$$\mathcal{L}(\hat{\mathbf{w}}_d,\mathbf{h}_d,\hat{\zeta}_d)=\mathcal{E}_d(\hat{\mathbf{w}}_d,\mathbf{h}_d)+\mu\left\|\hat{\mathbf{w}}_d-\sqrt{N}\mathbf{F}\mathbf{h}_d+\frac{1}{\mu}\hat{\zeta}_d\right\|_2^2, \tag{12}$$

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

Then the ADMM technique [32] is applied to alternatively solve the following subproblems.

Remark 6: 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. (12), the optimal $\hat{\mathbf{w}}_d^{(i+1)}$ can be obtained by solving Eq. (13):

$$\begin{aligned}\hat{\mathbf{w}}_d^{(i+1)} = \arg\min_{\hat{\mathbf{w}}_d}\Bigg\{ &\left\|\hat{\mathbf{y}}-\hat{\mathbf{w}}_d^*\odot\hat{\mathbf{x}}_d^k\right\|_2^2 \\ &+\gamma\left\|(\hat{\mathbf{w}}_d^*-\hat{\mathbf{w}}_d^{k-\Delta k*})\odot(\hat{\mathbf{x}}_d^k+\hat{\mathbf{x}}_d^{k-\Delta k})\right\|_2^2 \\ &+\mu\left\|\hat{\mathbf{w}}_d-\sqrt{N}\mathbf{F}\mathbf{h}_d+\frac{1}{\mu}\hat{\zeta}_d\right\|_2^2\Bigg\}\ .\end{aligned} \tag{13}$$

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

$$\hat{\mathbf{w}}_d^{(i+1)}=\frac{\hat{\mathbf{x}}_d^k\odot\hat{\mathbf{y}}^*+\gamma\hat{\mathbf{m}}_d^{\mathbf{xx}}\odot\hat{\mathbf{w}}_d^{k-\Delta k}+\mu\hat{\mathbf{h}}_d-\hat{\zeta}_d}{\hat{\mathbf{x}}_d^k\odot\hat{\mathbf{x}}_d^{k*}+\gamma\hat{\mathbf{m}}_d^{\mathbf{xx}}+\mu}, \tag{14}$$

where $\hat{\mathbf{m}}_d^{\mathbf{xx}}=(\hat{\mathbf{x}}_d^k+\hat{\mathbf{x}}_d^{k-\Delta k})\odot(\hat{\mathbf{x}}_d^{k*}+\hat{\mathbf{x}}_d^{k-\Delta k*})$, and the fraction operator represents element-wise division.

Remark 7: Detailed derivation can be found in Appendix A.

2) Subproblem $\mathbf{h}_d$: If $\hat{\mathbf{w}}_d$ and $\hat{\zeta}_d$ are given in Eq. (12), the optimal $\mathbf{h}_d^{(i+1)}$ can be solved by Eq. (15):

$$\mathbf{h}_d^{(i+1)}=\arg\min_{\mathbf{h}_d}\left\{\lambda\left\|\mathbf{s}\odot\mathbf{h}_d\right\|_2^2+\mu\left\|\hat{\mathbf{w}}_d-\sqrt{N}\mathbf{F}\mathbf{h}_d+\frac{1}{\mu}\hat{\zeta}_d\right\|_2^2\right\}. \tag{15}$$

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

$$\mathbf{h}_d^{(i+1)}=\frac{\mathcal{F}^{-1}(\mu\hat{\mathbf{w}}_d+\hat{\zeta}_d)}{\frac{\lambda}{N}(\mathbf{s}\odot\mathbf{s}^*)+\mu}\ , \tag{16}$$

where $\mathcal{F}^{-1}$ is the inverse discrete Fourier transform.

Remark 8: Detailed derivation can be found in Appendix B.

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

$$\hat{\zeta}_d^{(i+1)}=\hat{\zeta}_d^{(i)}+\mu\left(\hat{\mathbf{w}}_d^{(i+1)}-\hat{\mathbf{h}}_d^{(i+1)}\right)\ , \tag{17}$$

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 [32]:

$$\mu^{(i+1)}=\min(\mu_{\max},\beta\mu^{(i)})\ . \tag{18}$$

Furthermore, the filter's robustness is improved by the following online adaptive scheme:

$$\hat{\mathbf{x}}_{d,\mathrm{model}}^k=(1-\eta)\hat{\mathbf{x}}_{d,\mathrm{model}}^{k-1}+\eta\hat{\mathbf{x}}_d^k\ , \tag{19}$$

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

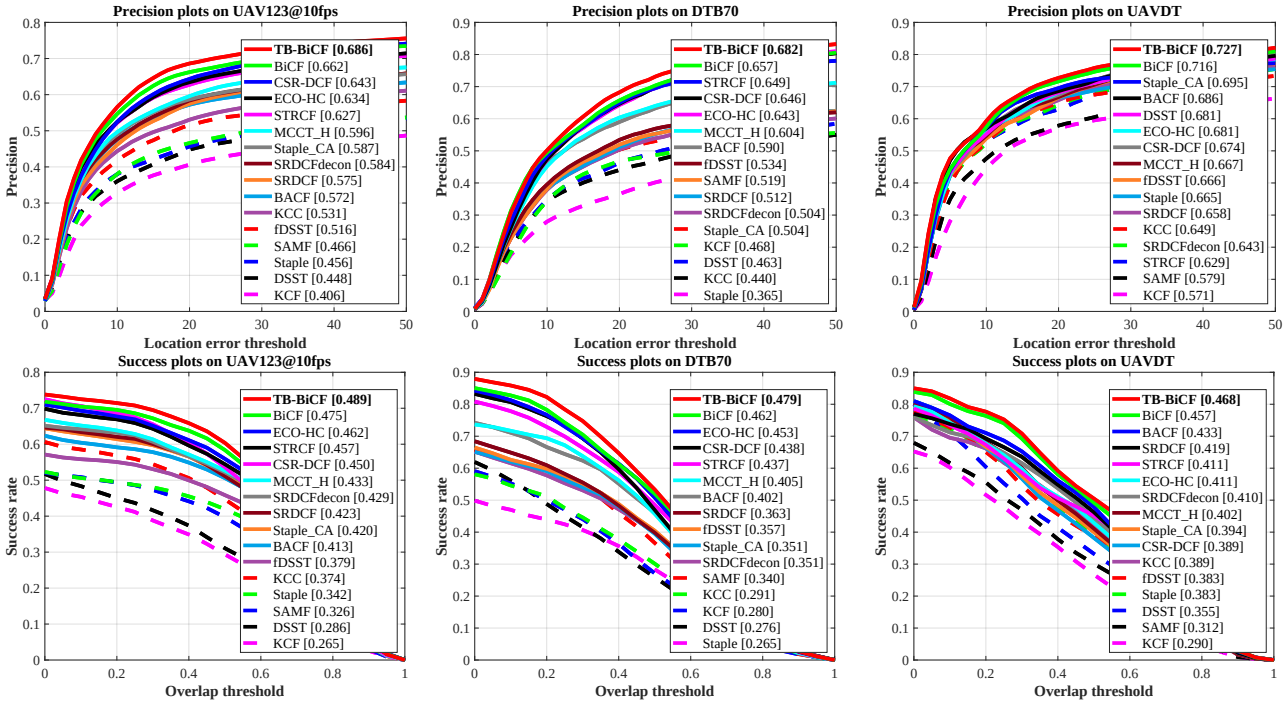
IV. Experiments

Quantitative and qualitative experiments are conducted to evaluate the proposed TB-BiCF tracker on 243 challenging image sequences from three well-known UAV tracking benchmarks, including UAV123@10fps [35], DTB70 [36], and UAVDT [31]. We carry out a comprehensive analysis of TB-BiCF and state-of-the-art tracking methods in this section, consisting of 15 trackers using hand-crafted features and 19 trackers based on deep learning.

Algorithm 1: TB-BiCF learning

Input: Image: I_k.
Maximum frame interval within temporary block: Δk.
Filters from the $k-\Delta k$-th frame: $\mathbf{W}^{k-\Delta k}$.
Feature maps from the $k-\Delta k$-th frame: $\mathbf{X}^{k-\Delta k}$.
Spatial regularizer weights: $\mathbf{s}$.
Output: The current filter $\mathbf{W}^k$ in the k-th frame.

1 Extract features $\mathbf{X}^k$ from I_k.
2 Introduce the auxiliary variable $\mathbf{h}_d$ and build the equality constraint form Eq. (10).
3 Transform Eq. (10) to Eq. (11) by Parseval's theorem.
4 Initialize variables $\hat{\mathbf{w}}_d^{(0)}$, $\mathbf{h}_d^{(0)}$, and $\hat{\zeta}_d^{(0)}$.
5 **for** ADMM iteration $i=1$ to *end* **do**
6 Solve subproblem $\hat{\mathbf{w}}_d^{(i+1)}$ by Eq. (14).
7 Solve subproblem $\mathbf{h}_d^{(i+1)}$ by Eq. (16).
8 Update Lagrangian multiplier $\hat{\zeta}_d^{(i+1)}$ by Eq. (17).
9 Update the penalty factor $\mu^{(i+1)}$ by Eq. (18).
10 **end**
11 Use Eq. (19) to update the appearance model.



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

Fig. 4. Precision and success plots of TB-BiCF and the other state-of-the-art 15 trackers using hand-crafted features on three UAV benchmarks. The scores of DP and AUC are respectively given in square brackets in the precision and success plots. Experimental results demonstrate that the proposed TB-BiCF tracker has the outstanding performance on all three challenging benchmarks.

A. Experimental setups

1) Evaluation metrics: The experiments employ two metrics, including center location error (CLE) and overlap, to evaluate all methods on the UAV123@10fps, DTB70, and UAVDT benchmarks based on the one-pass evaluation. The CLE is used to measure the Euclidean distance between the estimated object location and the center of the ground truth bounding box. The precision plot represents the curve formed by the percentage of bounding boxes whose CLE is less than a given threshold. According to the standard ranking metrics [35], the percentage of the CLE at 20 pixels is used as the distance precision (DP) to rank trackers in term of precision. The overlap metric is applied to measure the intersection over union (IoU) between the estimated bounding box and the ground truth. The curve on the success plot consists of the ratio of the number of frames with IoU greater than a given threshold to the total number of frames. In this work, the area under the curve (AUC) [35] in the success plot is employed to rank trackers in terms of success rate. Additionally, both frames per second and milliseconds per frame (MSPF) are applied to evaluate the tracking speed.

2) Implementation details: The TB-BiCF tracker applies a combination of hand-crafted features to meet the real-time requirements of tracking speed, including a fast version of HOG [49] with 31 channels, CN [34] with 10 channels, and gray-scale features with 1 channel, for object representation. Thus the total number of feature channels D is set to 42. We set the regularization parameter $\lambda = 1$ and $\gamma = 0.1$, and the length of the temporary block $\Delta k = 8$. Following the setting of large values of μ and β and few iterations of ADMM optimization in [15], the initial penalty factor $\mu = 100$, the maximum value $\mu_{\max} = 10^5$, and scale step $\beta = 50$ are employed in this work, and the number of ADMM iterations is set to 3 for a better trade-off between efficiency and accuracy. The online adaptation rate η is set to 0.039. All hyper-parameters remain fixed for all image sequences on all benchmarks. The proposed TB-BiCF runs on MATLAB R2018a on a computer with an i7-8700K CPU and an RTX 2080 GPU. Our MATLAB implementation is publicly released at https://github.com/vision4robotics/TB-BiCF-Tracker.

Remark 9: Objective evaluations of trackers used for comparison are performed on the same computer, by utilizing the open-source codes and default settings provided by the authors.

B. Comparison with trackers using hand-crafted features

1) Overall performance comparison: The proposed TB-BiCF tracker is first compared with 15 state-of-the-art trackers using hand-crafted features, *i.e.*, KCF [7], DSST [38], SAMF [9], SRDCFdecon [14], Staple [10], BACF [15], CSR-DCF [18], ECO-HC [24], fDSST [39], SRDCF [13], Staple_CA [44], KCC [50], MCCT-H [11], STRCF [25], and BiCF [1], on three UAV benchmarks in terms of overall performance. Figure 4 presents that the TB-BiCF tracker obtains the outstanding performance compared with the other 15 hand-crafted feature-based trackers on all three challenging UAV benchmarks .

UAV123@10fps. The UAV123@10fps benchmark [35] is one of the most widely used benchmarks captured from low-altitude UAVs and contains 123 image sequences with various

TABLE I
THE OVERALL TRACKING PERFORMANCE AND SPEED OF TB-BICF VERSUS OTHER HAND-CRAFTED FEATURE-BASED TRACKERS ON THREE CHALLENGING UAV BENCHMARKS. THE TRACKING PERFORMANCE OF THE TOP 3 TRACKERS IS SHOWN IN **RED**, **GREEN**, AND **BLUE** FONTS. ALL RESULTS ARE GENERATED IN CPU MODE. THE PROPOSED TB-BICF ACHIEVES THE BEST DP AND AUC RESULTS OVER ALL CHALLENGING UAV BENCHMARKS WITH A REAL-TIME SPEED.

Tracker	Venus	UAV123@10fps			DTB70			UAVDT			Average		
		DP	AUC	Speed	DP	AUC	Speed	DP	AUC	Speed	DP	AUC	Speed
KCF	15'TPAMI	0.406	0.265	**618.7**	0.468	0.280	**377.5**	0.571	0.290	**956.9**	0.481	0.278	**651.1**
DSST	14'BMVC	0.448	0.286	**98.5**	0.463	0.276	**72.7**	0.681	0.355	**148.3**	0.531	0.305	**106.5**
BACF	17'CVPR	0.572	0.413	52.5	0.590	0.402	46.5	0.686	**0.433**	69.1	0.616	0.416	56.0
SAMF	14'ECCVW	0.466	0.326	12.5	0.519	0.340	10.0	0.579	0.312	15.8	0.521	0.326	12.8
Staple_CA	17'CVPR	0.587	0.420	56.0	0.504	0.351	56.5	**0.695**	0.394	64.1	0.595	0.388	58.9
SRDCF	15'ICCV	0.575	0.423	13.9	0.512	0.363	10.7	0.658	0.419	17.4	0.582	0.402	14.0
SRDCFdecon	16'CVPR	0.584	0.429	7.6	0.504	0.351	6.0	0.643	0.410	8.9	0.577	0.397	7.5
MCCT_H	18'CVPR	0.596	0.433	57.2	0.604	0.405	59.0	0.667	0.402	62.9	0.622	0.414	59.7
CSR-DCF	17'CVPR	**0.643**	0.450	11.3	0.646	0.438	11.8	0.674	0.389	13.2	**0.655**	0.426	12.1
STRCF	18'CVPR	0.627	0.457	26.9	**0.649**	0.437	26.3	0.629	0.411	32.3	0.635	0.435	28.5
ECO-HC	17'CVPR	0.634	**0.462**	66.6	0.643	**0.453**	62.2	0.681	0.411	79.2	0.653	**0.442**	69.3
fDSST	17'TPAMI	0.516	0.379	**153.7**	0.534	0.357	**132.0**	0.666	0.383	**218.5**	0.572	0.373	**168.1**
KCC	18'AAAI	0.531	0.374	40.8	0.440	0.291	40.7	0.649	0.389	56.9	0.540	0.351	46.1
Staple	16'CVPR	0.456	0.342	62.1	0.365	0.265	62.5	0.665	0.383	71.7	0.495	0.330	65.4
BiCF	20'ICRA	**0.662**	**0.475**	44.0	**0.657**	**0.462**	42.1	**0.716**	**0.457**	50.2	**0.679**	**0.465**	45.4
TB-BiCF	Ours	**0.686**	**0.489**	45.4	**0.682**	**0.479**	43.1	**0.727**	**0.468**	51.8	**0.698**	**0.479**	46.8

challenging factors. Since the UAV123@10fps dataset is generated by temporally downsampling UAV123 [35] (captured at 30 FPS) to 10 FPS, the absence of intermediate frames means the object has more significant displacement and appearance changes between frames, and thus makes the tracking more difficult. Therefore, the UAV123@10fps benchmark is chosen for the evaluation in this work. In Fig. 4(a), the TB-BiCF tracker has the leading DP score (0.686) exceeding the second highest tracker BiCF (0.662) and the third highest tracker CSR-DCF (0.643) by 2.4% and 4.3%, respectively. TB-BiCF also leads in the AUC score (0.489), which is 1.4% higher than BiCF (0.475) in the second place and 2.7% higher than ECO-HC (0.462) in the third place.

DTB70. The DTB70 benchmark [36] contains 15777 pictures, making up 70 image sequences captured by drone cameras. As shown in Fig. 4(b), TB-BiCF provides the best DP score (0.682), which exceeds 2.5% and 3.3% of the second best tracker BiCF (0.657) and the third best tracker STRCF (0.649). Moreover, the best AUC score is also obtained by TB-BiCF (0.479), followed by BiCF and ECO-HC.

UAVDT. The UAVDT benchmark [31] focusing on complex scenarios with ∼37K representative frames for UAV object tracking task. In Fig. 4(c), TB-BiCF achieves the best DP score (0.727) compared to BiCF (0.716) and Staple_CA (0.695). TB-BiCF also gets the best AUC score (0.468), followed by BiCF (0.457) and BACF (0.433).

Remark 10: In addition to impressive tracking performance, the speed of the proposed TB-BiCF tracker (46.8 FPS) is sufficient for UAV real-time tracking (Table I). Although KCF obtains the best tracking speed (651.1 FPS), followed by fDSST (168.1 FPS) and DSST (106.5 FPS), their DP and AUC scores are much lower than TB-BiCF. The outstanding performance proves the effectiveness of the proposed temporary block-based bidirectional incongruity learning strategy.

2) Attribute-based evaluation: The attribute-based evaluation is performed to manifest the performance of the presented tracker in different challenging UAV tracking scenarios.

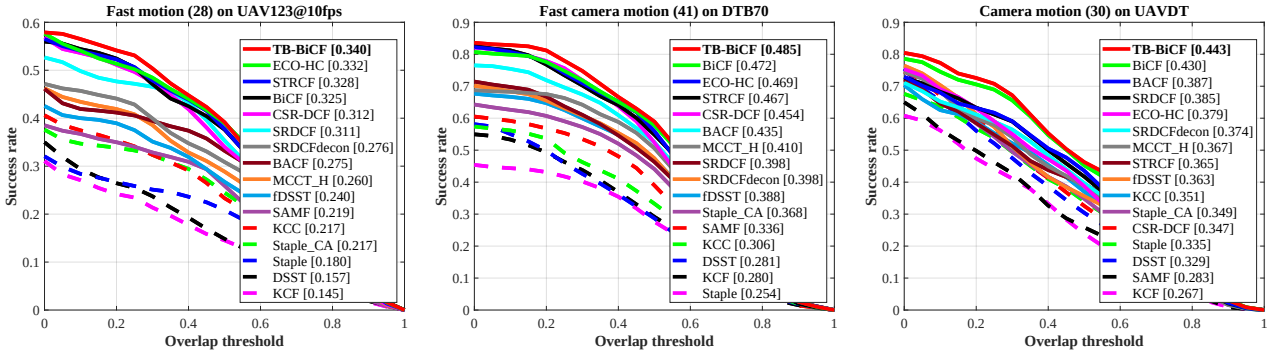
UAV123@10fps. The video sequences on the UAV123@10fps benchmark is annotated with 12 different attributes, including camera motion (CM), fast motion (FM), illumination variation (IV), low resolution (LR), scale variation (SV), viewpoint change (VC), partial occlusion (POC), aspect ratio change (ARC), out-of-view (OV), similar object (SOB), background clutter (BC), and full occlusion (FOC).

DTB70. The 11 different attributes from the DTB70 benchmark are fast camera motion (FCM), deformation (DEF), aspect ratio variation (ARV), in-plane rotation (IPR), out-of-plane rotation (OPR), out-of-view (OV), occlusion (OCC), and similar objects around (SOA).

UAVDT. The UAVDT benchmark includes the following 9 attributes: object blur (OB), camera motion (CM), object motion (OM), illumination variations (IV), small object (SO), background clutter (BC), scale variations (SV), long-term tracking (LTT), and large occlusion (LO).

To further corroborate the validity of TB-BiCF in the face of fast motion, some success plots of the related attributes from three challenging benchmarks are presented in Fig. 5. More specifically, in Fig. 5(a), the TB-BiCF tracker obtains the best AUC score (0.340) in the attribute of fast motion, exceeding the second best tracker ECO-HC (0.332) by 0.8% and the third tracker STRCF (0.328) by 1.2%. In Fig. 5(b), TB-BiCF also achieves the best AUC score (0.485) in the attribute of fast camera motion and exceeds BiCF (0.472) and ECO-HC (0.469) by 1.3% and 1.6%, respectively. In the camera motion from the UAVDT benchmark, the highest AUC score belongs to TB-BiCF (0.443) improving the second best tracker BiCF (0.430) by 1.2% and the third tracker BACF (0.387) by 5.6%. The impressive performance mainly contributes to the proposed temporary block-based bidirectional incongruity-aware learning. When sudden object appearance variation occurs, the proposed tracker can efficiently repress the bidirectional incongruity between the frames within the temporary block, and thus the tracking robustness is generally enhanced.

Remark 11: Figure 6 provides the complete attribute-based AUC scores from the three challenging benchmarks. The top 5 trackers with average performance are used to present the



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

Fig. 5. Success plots of the proposed TB-BiCF and the other 15 handcrafted-based trackers in the attributes related to fast motion from the three challenging UAV benchmarks. The AUC scores are given in square brackets. Experimental results show that TB-BiCF achieves the outstanding attribute-based performance.

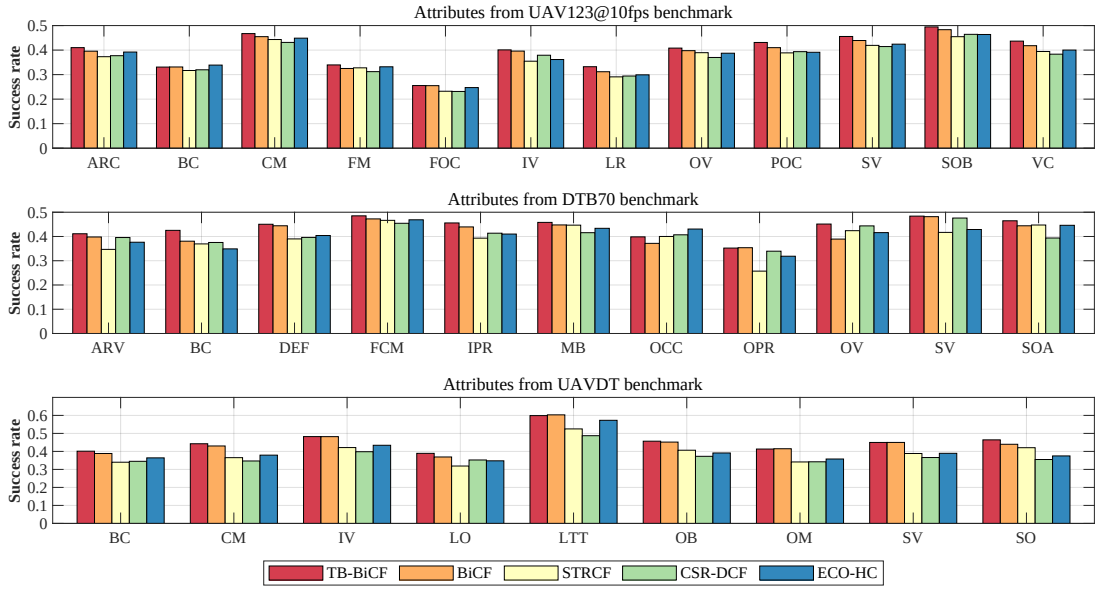


Fig. 6. AUC-based scores in different attributes on the three challenging UAV benchmarks. The top 5 trackers with average performance are used in the attributed-based evaluation. Experimental results show that the TB-BiCF tracker achieves the superior attribute-based performance in most attributes.

attribute-based results. The experimental results demonstrate that TB-BiCF outperforms favorably than other competing trackers in most attributes.

3) Qualitative evaluation: Extensive evaluations are performed to present the performance of the proposed TB-BiCF tracker in comparison with the other top 4 trackers with average performance, *i.e.*, BiCF, STRCF, CSR-DCF, and ECO-HC. In Fig. 7, three challenging sequences, *i.e.*, *Car16_1* from UAV123@10fps, *MountainBike5* from DTB70, and *S0601* from UAVDT, are applied to provide the tracking results, the CLE and overlap curves. The CLE curve denotes the Euclidean distance between the estimated object location and the ground truth bounding box center in each frame. The overlap rate represents the IoU between the predicted bounding boxes and the ground truth in each frame. Fast motion and motion blur are great challenges in these tracking scenes. Most trackers have tracking drift when the object appearance changes rapidly, that is, the CLE gradually increases while the overlap decreases. In contrast, the proposed TB-BiCF tracker can well accommodate the sudden appearance variation and is robust against the large appearance changes because of the bidirectional incongruity learning strategy. In other words, the TB-BiCF can achieve the trade-off between discriminate power inheritance and generalization power enhancement.

Remark 12: More representative tracking results are visualized in Fig. 8 to demonstrate the discriminative capability against motion blur, fast UAV/object motion, low resolutions, and small objects.

4) Failure cases: Figure 9 provides three challenging sequences from different benchmarks that the proposed TB-BiCF tracker fails to track the object. In the sequence *person19_2* and *RcCar3*, the object is out of the onboard camera's view, and the TB-BiCF tracker fails to perceive the object reappearing from the scene. Thus the inter-frame information without the existence of the object is used in filter learning, which greatly affects the discriminative power. In the sequence *S1606*, the car is mostly blocked by the lamp post, and the available effective object information is scarce. In this case,

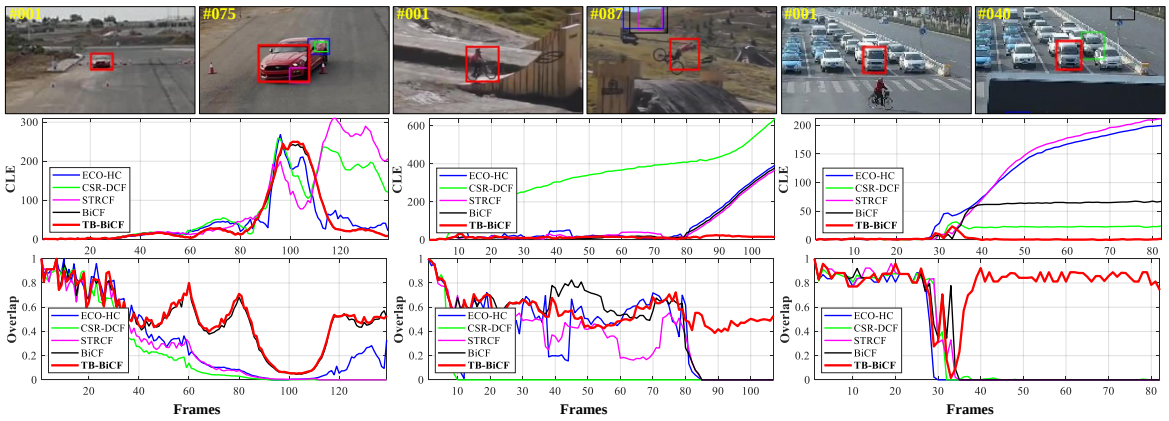


Fig. 7. CLE and overlap curves in three challenging sequences, including *car16_1* from UAV123@10fps, *MountainBike5* from DTB70, and *S0601* from UAVDT. Fast motion and motion blur are great challenges in these tracking scenes. Most trackers have tracking drift when the object appearance changes rapidly while the proposed TB-BiCF tracker can maintain the discriminative power.

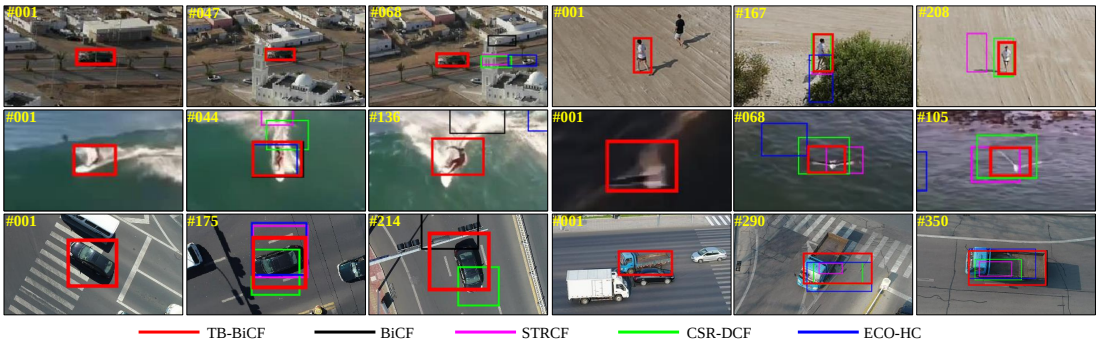


Fig. 8. Qualitative comparisons of the proposed TB-BiCF tracker with other advanced trackers. From left to right and top to bottom, these sequences are *truck2* from UAV123@10fps, *group2_2* from UAV123@10fps, *Surfing04* from DTB70, *Gull1* from DTB70, *S0309* from UAVDT, and *S1607* from UAVDT, respectively. The proposed TB-BiCF tracker shows the superior performance in complex environment. The UAV object tracking videos are available at https://youtu.be/PFR-AB79iWY.

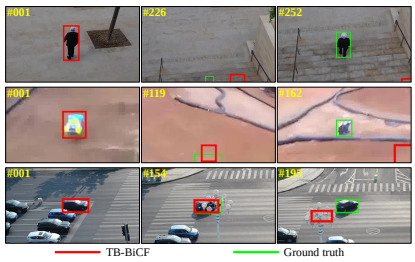


Fig. 9. Failure cases of the presented TB-BiCF tracker. The first, second, and third row show the *person19_2* from UAV123@10fps, *RcCar3* from DTB70, and *S1606* from UAVDT.

the features of the appearance are contaminated by irrelevant information, which also interferes the filter training.

C. Comparison with deep-based trackers

This work further evaluates the tracking performance and efficiency of the TB-BiCF tracker in comparison with other deep-based state-of-the-art trackers on the UAVDT benchmark. These 19 deep-based advanced trackers can be categorized as follows: CF-based trackers that rely on convolutional features, *i.e.*, ECO [24], DeepSTRCF [25], CF2 [51], C-COT [23], MCPF [52], CoKCF [53], IBCCF [40], MCCT [11], and ASRCF [54]; Trackers based on deep learning architecture, *i.e.*, SINT [55], HDT [56], SiamFC [27], GOTURN [57], ADNet [58], CFNet [28], DSiam [59], UDT, UDT+ [30], and TADT [29].

TABLE II
TRACKING PERFORMANCE AND SPEED COMPARISONS WITH THE OTHER 19 DEEP-BASED TRACKERS ON UAVDT BENCHMARK. THE BEST, SECOND, AND THIRD PERFORMANCE ARE REPRESENTED BY THE **RED**, **GREEN**, AND **BLUE** FONTS. NOTE THT THE DEEPSTRCF TRACKER IS DENOTED BY D-STRCF IN THE TABLE FOR A CONCISE REPRESENTATION.

Trackers	Venue	DP	AUC	FPS	MSPF	GPU
CF2	15'ICCV	0.602	0.355	20.1	49.75	✓
C-COT	16'ECCV	0.656	0.406	1.1	909.09	✓
ECO	17'CVPR	0.700	0.454	16.4	60.98	✓
MCPF	17'CVPR	0.660	0.403	0.7	1428.57	✓
CoKCF	17'PR	0.605	0.319	21.2	47.14	✓
IBCCF	17'ICCVW	0.603	0.389	3.4	294.12	✓
MCCT	18'CVPR	0.671	0.437	8.6	116.28	✓
D-STRCF	18'CVPR	0.667	0.437	6.6	151.52	✓
ASRCF	19'CVPR	0.700	0.437	14.1	70.92	✓
SINT	16'CVPR	0.570	0.290	**96.8**	**10.33**	✓
HDT	16'CVPR	0.596	0.303	9.0	111.11	✓
SiamFC	16'ECCVW	**0.708**	**0.465**	18.2	54.95	✓
GOTURN	16'ECCV	0.702	0.452	16.5	60.61	✓
ADNet	17'CVPR	0.605	0.319	7.5	133.33	✓
CFNet	17'CVPR	0.680	0.428	41.0	24.39	✓
DSiam	17'ICCV	**0.704**	**0.457**	15.9	64.89	✓
UDT	19'CVPR	0.674	0.441	**76.4**	**13.09**	✓
UDT+	19'CVPR	0.697	0.416	**60.4**	**16.56**	✓
TADT	19'CVPR	0.677	0.431	32.5	30.77	✓
TB-BiCF	Ours	**0.727**	**0.468**	51.8	19.31	✗

Table II reports the DP and AUC scores of TB-BiCF and trackers that are used for comparison. The results demonstrate that TB-BiCF achieves the best DP (0.727) and obtains 1.9% and 2.3% gain respectively than SiamFC (0.708) and DSiam (0.704). In terms of success rate, TB-BiCF obtains the highest AUC score of 0.468 outperforming SiamFC (0.465), DSiam (0.457), and ECO (0.454) by 0.3%, 1.1%, and 1.4%, respectively. In terms of operational efficiency, Table II also gives the tracking speed of all trackers on the UAVDT benchmark. SINT obtains the best tracking speed with 96.8 FPS, followed by UDT (76.4 FPS) and UDT+ (60.4 FPS). The higher speed of these trackers benefits from the employment of GPUs, while the tracking performance of these trackers is lower than TB-BiCF both in the DP and AUC score, which runs on a single CPU with a real-time speed of 51.8 FPS. Besides, TB-BiCF outperforms all of 9 CF-based trackers with satisfactory speed and achieves better tracking performance as well.

Remark 13: The in-depth evaluation between 19 deep-based trackers demonstrates higher tracking performance and high-efficiency of the proposed TB-BiCF tracker, which can provide accurate and efficient performance for future tracking applications on the real-time vision-based aerial platform.

D. Validity analysis of key parameters

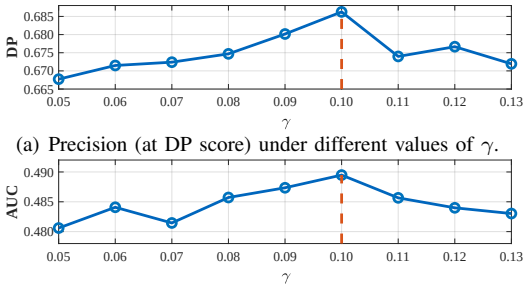
To shed light the effect of key parameters on the performance, an ablation study on the regularization factor λ and the parameters of the temporary block-based bidirectional incongruity-aware model, *i.e.*, γ and Δk, are conducted on the UAV123@10fps benchmark. Then the sensitivity analysis of the bidirectional incongruity penalty factor γ and maximum frame interval Δk within the temporary block are carried out.

1) Ablation study: By only applying the classification error term Eq. (7) to learn the filter, the TB-BiCF-NBR is obtained without the proposed bidirectional incongruity and the regularization term, that is, $\gamma = 0$ and $\lambda = 0$. Then the TB-BiCF-NB whose $\lambda = 1$ is learned by adding the regularization term to the TB-BiCF-NBR. On the basis of the TB-BiCF-NB, the TB-BiCF-1 whose $\gamma = 0.1$ and $\Delta k = 1$ is obtained by adding the proposed bidirectional incongruity term. The TB-BiCF is the final version with $\gamma = 0.1$ and $\Delta k = 8$. Table III presents the results of each model for the DP and AUC scores. Specifically, the results indicate that the tracking performance is gradually improved with the improvement of the model and the final version TB-BiCF obtains the best performance compared with other models.

2) Regularization factor γ: γ values are set from 0.05 to 0.13 empirically for the trial, with a step size of 0.01. Figure 10 reports the results of the DP and AUC scores. The performance gradually increases before reaching the highest point (0.686) at $\gamma = 0.1$ in Fig. 10(a). After that, the DP score decreases slightly. For the AUC score in Fig. 10(b), the performance sees the trend similar to the DP plots and achieves the best score (0.489) at $\gamma = 0.1$. Compared to the performance at $\gamma = 0.05$, the precision and success rate obtain a gain of 1.8% and 0.8% respectively when $\gamma = 0.1$. The results show that when γ is set in a certain range, the proposed bidirectional incongruity-aware learning strategy can effectively improve the overall performance. Therefore, γ is set to 0.1 in this work.

TABLE III
ABLATION STUDY OF THE KEY PARAMETERS. THE FINAL VERSION TB-BICF ACHIEVES THE BEST DP AND AUC SCORES.

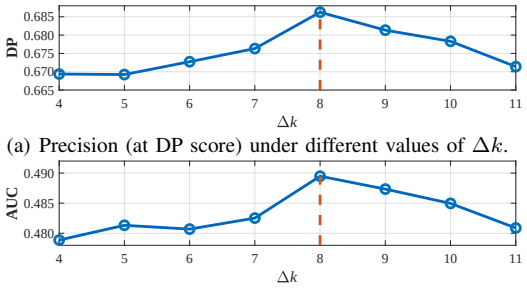
	TB-BiCF-NBR	TB-BiCF-NB	TB-BiCF-1	TB-BiCF
(λ,γ,Δk)	(0,0,-)	(1,0,-)	(1,0.1,1)	(1,0.1,8)
DP	0.534	0.667	0.668	**0.686**
AUC	0.380	0.477	0.483	**0.489**



(a) Precision (at DP score) under different values of γ.

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

Fig. 10. Different values of bidirectional incongruity factor γ are tested on UAV123@10fps dataset. At $\gamma = 0.1$, both the DP and AUC scores reach the highest scores.



(a) Precision (at DP score) under different values of Δk.

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

Fig. 11. Different values of frame interval Δk are tested on UAV123@10fps dataset. At $\Delta k = 8$, both the DP and AUC scores reach the highest scores.

3) Frame interval Δk: Δk values are set from 4 to 11 empirically for the trial, with a step size of 1. The results of DP and AUC scores are reported in Fig. 11. The performance also gradually increases before reaching the highest point (0.686) at $\Delta k = 8$ in Fig. 11(a). Then the DP score decreases slightly until $\Delta k = 11$. In Fig. 11(b), the AUC score has the trend similar to the DP and obtains the best score (0.489) at $\Delta k = 8$. The DP and AUC score at $\Delta k = 8$ obtains a gain of 1.7% and 1.0% respectively compared with the performance at $\Delta k = 4$. The results indicate that the temporary block-based strategy can effectively improve the overall performance when Δk is set in a certain range. Therefore, Δk is set to 8 in this work.

E. Temporary block setting

The proposed temporary block-based strategy can be further extended to the introduction of multiple samples, and thus the extension form of the temporary block-based bidirectional incongruity regularization term Eq. (8) is expressed as follows:

$$\mathcal{E}_{\mathrm{t,ext}}(\mathbf{W}) = \gamma \sum_{d=1}^{D} \sum_{p=1}^{\Delta k} \left\| (\mathbf{w}_d - \mathbf{w}_d^{k-p}) \star (\mathbf{x}_d^k + \mathbf{x}_d^p) \right\|_2^2. \quad (20)$$

The extension form uses the pairwise combination of the previous frames (from $\#k-\Delta k$ to $\#k-1$) and the current frame

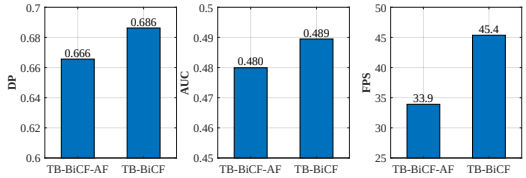


Fig. 12. Tracking performance and speed comparisons with TB-BiCF-AF and TB-BiCF. The proposed TB-BiCF obtains the outstanding performance in DP, AUC, and FPS.

$\#k$ to construct the bidirectional incongruity, that is, it applies all frames in the temporary block to build the bidirectional incongruity. By substituting $\mathcal{E}_{\mathrm{t}}(\mathbf{W})$ in Eq. (6) with the extension form $\mathcal{E}_{\mathrm{t,ext}}(\mathbf{W})$, a new overall optimization problem can be obtained. The solving process of the optimization problem is the same as the original one. Thus the solution to subproblem $\mathbf{h}_d$ is the same as Eq. (16), and the solution to $\hat{\mathbf{w}}_d$ is:

$$\hat{\mathbf{w}}_d^{(i+1)} = \frac{\hat{\mathbf{x}}_d^k \odot \hat{\mathbf{y}}^* + \gamma \sum_{p=1}^{\Delta k} \hat{\mathbf{m}}_d^{k-p,\mathbf{xx}} \odot \hat{\mathbf{w}}_d^{k-p} + \mu \hat{\mathbf{h}}_d - \hat{\zeta}_d}{\hat{\mathbf{x}}_d^k \odot \hat{\mathbf{x}}_d^{k*} + \gamma \sum_{p=1}^{\Delta k} \hat{\mathbf{m}}_d^{k-p,\mathbf{xx}} + \mu}, \tag{21}$$

where $\hat{\mathbf{m}}_d^{k-p,\mathbf{xx}} = (\hat{\mathbf{x}}_d^k + \hat{\mathbf{x}}_d^{k-p}) \odot (\hat{\mathbf{x}}_d^{k*} + \hat{\mathbf{x}}_d^{k-p*})$.

Figure 12 presents the tracking performance of TB-BiCF and the extension model using all frames in the temporary block, *i.e.*, TB-BiCF-AF. The results show that the TB-BiCF has the best DP and AUC scores compared to TB-BiCF-AF. The suboptimal performance of TB-BiCF-AF can be attributed to the fact that the extension model introduces all frames in the temporary block, which can lead to over-fitting of the model and introduce the risk of model degradation. Moreover, the speed of TB-BiCF is 1.3× faster than TB-BiCF-AF, since TB-BiCF-AF needs to process all frames in the temporary block in each filter learning stage. Therefore, the temporary block-based strategy considering the first and last frames in the block can better improve the discrimination ability and generalization ability of the filter and can be efficiently applied to the UAV tracking applications.

V. Conclusions

In this work, a temporary block-based bidirectional incongruity-aware correlation filter, *i.e.*, TB-BiCF, is proposed to perform real-time UAV object tracking. By considering the temporary block-frame bidirectional incongruity in the correlation filter learning, the novel tracker can make full use of the inter-frame information to obtain the balance between discriminate power inheritance and generalization power enhancement. Considerable experiments are performed to verify the proposed approach on three challenging UAV tracking benchmarks. Comprehensive experimental results manifest that the presented TB-BiCF tracker has the outstanding performance against 34 state-of-the-art tracking methods in accuracy, robustness, and efficiency. Additionally, the proposed method with less computation is suitable for real-time UAV tracking tasks. The results of TB-BiCF will further expand the development of the temporary block-based bidirectional incongruity repression strategy in UAV object tracking applications. We strongly believe that the development of future work can further improve the proposed method and promote the development of UAV aerial video analysis.

Appendix A
Derivation of the Optimization Problem Eq. (13)

The original objective function in Eq. (13) can be equivalently expressed as follows:

$$\begin{aligned} &\left\|\hat{\mathbf{y}} - \hat{\mathbf{X}}_d^k \hat{\mathbf{w}}_d^*\right\|_2^2 + \gamma \left\| \left(\hat{\mathbf{X}}_d^k + \hat{\mathbf{X}}_d^{k-\Delta k}\right) \left(\hat{\mathbf{w}}_d^* - \hat{\mathbf{w}}_d^{k-\Delta k*}\right) \right\|_2^2 \\ &+ \mu \left\| \hat{\mathbf{w}}_d - \sqrt{N}\mathbf{F}\mathbf{h}_d + \frac{1}{\mu}\hat{\zeta}_d \right\|_2^2 , \end{aligned} \tag{22}$$

where $\hat{\mathbf{X}}_d^k = \mathrm{diag}\left(\hat{\mathbf{x}}_d^k\right)$ and $\hat{\mathbf{X}}_d^{k-\Delta k} = \mathrm{diag}\left(\hat{\mathbf{x}}_d^{k-\Delta k}\right)$ denote the diagonal matrices.

By differentiating the objective function Eq. (22) with respect to the filter $\hat{\mathbf{w}}_d^*$ in the d-th channel and setting the outcome to zero, we can obtain:

$$\begin{aligned} &\hat{\mathbf{X}}_d^{k\top}\left(\hat{\mathbf{X}}_d^{k*}\hat{\mathbf{w}}_d - \hat{\mathbf{y}}^*\right) + \mu\hat{\mathbf{w}}_d - \mu\sqrt{N}\mathbf{F}\mathbf{h}_d + \hat{\zeta}_d + \\ &\gamma\left(\hat{\mathbf{X}}_d^k + \hat{\mathbf{X}}_d^{k-\Delta k}\right)^\top \left(\left(\hat{\mathbf{X}}_d^k + \hat{\mathbf{X}}_d^{k-\Delta k}\right)^* \left(\hat{\mathbf{w}}_d - \hat{\mathbf{w}}_d^{k-\Delta k}\right)\right) = 0, \end{aligned} \tag{23}$$

where the superscript * and $^\top$ denotes the conjugate and the transpose of a vector or matrix, respectively. By the transposition of terms, the following equation can be achieved:

$$\begin{aligned} &\left(\hat{\mathbf{X}}_d^{k\top}\hat{\mathbf{X}}_d^{k*} + \gamma\hat{\mathbf{M}}_d^{\mathbf{xx}} + \mu\mathbf{I}_N\right)\hat{\mathbf{w}}_d \\ &= \hat{\mathbf{X}}_d^{k\top}\hat{\mathbf{y}}^* + \mu\sqrt{N}\mathbf{F}\mathbf{h}_d - \hat{\zeta}_d + \gamma\hat{\mathbf{M}}_d^{\mathbf{xx}}\hat{\mathbf{w}}_d^{k-\Delta k} , \end{aligned} \tag{24}$$

where $\hat{\mathbf{M}}_d^{\mathbf{xx}} = \left(\hat{\mathbf{X}}_d^k + \hat{\mathbf{X}}_d^{k-\Delta k}\right)^\top \left(\hat{\mathbf{X}}_d^k + \hat{\mathbf{X}}_d^{k-\Delta k}\right)^*$ and $\mathbf{I}_N$ is an identity matrix of size $N \times N$. Note that the matrix on the left side of the equation, *i.e.*, $\left(\hat{\mathbf{X}}_d^{k\top}\hat{\mathbf{X}}_d^{k*} + \gamma\hat{\mathbf{M}}_d^{\mathbf{xx}} + \mu\mathbf{I}_N\right)$, is diagonal, so the inverse of the matrix can be easily obtained by replacing each element in the diagonal with its reciprocal. The solution to $\hat{\mathbf{w}}_d$ can be achieved by the element-wise operation:

$$\begin{aligned} \hat{\mathbf{w}}_d &= \left(\hat{\mathbf{X}}_d^{k\top}\hat{\mathbf{X}}_d^{k*} + \gamma\hat{\mathbf{M}}_d^{\mathbf{xx}} + \mu\mathbf{I}_N\right)^{-1} \left(\hat{\mathbf{X}}_d^{k\top}\hat{\mathbf{y}}^* \right. \\ &\left. + \mu\sqrt{N}\mathbf{F}\mathbf{h}_d - \hat{\zeta}_d + \gamma\hat{\mathbf{M}}_d^{\mathbf{xx}}\hat{\mathbf{w}}_d^{k-\Delta k}\right) \\ &= \frac{\hat{\mathbf{x}}_d^k \odot \hat{\mathbf{y}}^* + \mu\hat{\mathbf{h}}_d - \hat{\zeta}_d + \gamma\hat{\mathbf{m}}_d^{\mathbf{xx}} \odot \hat{\mathbf{w}}_d^{k-\Delta k}}{\hat{\mathbf{x}}_d^k \odot \hat{\mathbf{x}}_d^{k*} + \gamma\hat{\mathbf{m}}_d^{\mathbf{xx}} + \mu} , \end{aligned} \tag{25}$$

where the vector $\hat{\mathbf{m}}_d^{\mathbf{xx}}$ is composed of the diagonal elements of the matrix $\hat{\mathbf{M}}_d^{\mathbf{xx}}$. This solution is equivalent to Eq. (14). ■

Appendix B
Derivation of the Optimization Problem Eq. (15)

By differentiating the objective function Eq. (15) with respect to the auxiliary variable $\hat{\mathbf{h}}_d^*$ in the d-th channel and setting the outcome to zero, we can achieve:

$$\lambda\mathbf{S}^{\mathrm{H}}\mathbf{S}\mathbf{h}_d - \mu\sqrt{N}\mathbf{F}^{\mathrm{H}}\hat{\mathbf{w}}_d + \mu N\mathbf{h}_d - \sqrt{N}\mathbf{F}^{\mathrm{H}}\hat{\zeta}_d = 0 , \tag{26}$$

where the superscript $^{\mathrm{H}}$ denotes the conjugate transpose of a matrix and $\mathbf{S} = \mathrm{diag}(\mathbf{s})$ is the diagonal matrix. By the transposition of terms, we can easily obtain the solution to the subproblem $\hat{\mathbf{h}}_d$:

$$\mathbf{h}_d = \left(\lambda\mathbf{S}^{\mathrm{H}}\mathbf{S} + \mu\mathbf{I}_N\right)^{-1}\left(\mu\sqrt{N}\mathbf{F}^{\mathrm{H}}\hat{\mathbf{w}}_d + \sqrt{N}\mathbf{F}^{\mathrm{H}}\hat{\zeta}_d\right) . \tag{27}$$

Note that the matrix $\left(\lambda \mathbf{S}^{\mathrm{H}}\mathbf{S} + \mu \mathbf{I}_N\right)$ is the diagonal matrix, so the inverse of the matrix can be achieved by replacing each element in the diagonal with its reciprocal. Therefore, the solution to $\hat{\mathbf{h}}_d$ can be expressed by the element-wise operation as follows:

$$\begin{aligned}\mathbf{h}_d &= \frac{\mu\sqrt{N}\mathbf{F}^{\mathrm{H}}\hat{\mathbf{w}}_d + \sqrt{N}\mathbf{F}^{\mathrm{H}}\hat{\zeta}_d}{\lambda(\mathbf{s}\odot\mathbf{s}^*)+\mu} = \frac{\frac{1}{\sqrt{N}}\mathbf{F}^{\mathrm{H}}\left(\mu\hat{\mathbf{w}}_d + \hat{\zeta}_d\right)}{\frac{\lambda}{N}(\mathbf{s}\odot\mathbf{s}^*)+\mu} \\ &= \frac{\mathcal{F}^{-1}(\mu\hat{\mathbf{w}}_d + \hat{\zeta}_d)}{\frac{\lambda}{N}(\mathbf{s}\odot\mathbf{s}^*)+\mu}\ .\end{aligned} \tag{28}$$

The resolution is equivalent to Eq. (16). ■

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] F. Lin, C. Fu, Y. He, F. Guo, and Q. Tang, "BiCF: Learning Bidirectional Incongruity-Aware Correlation Filter for Efficient UAV Object Tracking," in *Proceedings of the International Conference on Robotics and Automation (ICRA)*, 2020, pp. 1–7.

[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)*, 2016, pp. 1562–1569.

[4] C. Martinez, C. Sampedro, A. Chauhan, and P. Campoy, "Towards Autonomous Detection and Tracking of Electric Towers for Aerial Power Line Inspection," in *Proceedings of the International Conference on Unmanned Aircraft Systems (ICUAS)*, 2014, pp. 284–295.

[5] 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.

[6] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista, "Exploiting the Circulant Structure of Tracking-by-Detection with Kernels," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2012, pp. 702–715.

[7] 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.

[8] M. Danelljan, F. S. Khan, M. Felsberg, and J. v. d. 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] 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*, 2015, pp. 254–265.

[10] L. Bertinetto, J. Valmadre, S. Golodetz, O. Miksik, and P. H. S. 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.

[11] 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.

[12] C. Fu, F. Lin, Y. Li, and G. Chen, "Correlation Filter-Based Visual Tracking for UAV with Online Multi-Feature Learning," *Remote Sensing*, vol. 11, no. 5, pp. 1–23, 2019.

[13] M. Danelljan, G. Häger, F. S. 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.

[14] M. Danelljan, G. Häger, F. S. 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.

[15] H. K. Galoogahi, A. Fagg, and S. Lucey, "Learning Background-Aware Correlation Filters for Visual Tracking," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 1144–1152.

[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, pp. 4415–4422.

[17] C. Fu, J. Xu, F. Lin, F. Guo, T. Liu, and Z. Zhang, "Object Saliency-Aware Dual Regularized Correlation Filter for Real-Time Aerial Tracking," *IEEE Transactions on Geoscience and Remote Sensing*, pp. 1–12, 2020.

[18] A. Lukežic, T. Vojír, L. C. 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, pp. 4847–4856.

[19] C. Sun, D. Wang, H. Lu, and M. Yang, "Correlation Tracking via Joint Discrimination and Reliability Learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 489–497.

[20] F. Du, P. Liu, W. Zhao, and X. Tang, "Joint Channel Reliability and Correlation Filters Learning for Visual Tracking," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 6, pp. 1625–1638, 2020.

[21] M. Danelljan, G. Häger, F. S. Khan, and M. Felsberg, "Convolutional Features for Correlation Filter Based Visual Tracking," in *Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCVW)*, 2015, pp. 621–629.

[22] L. Wang, W. Ouyang, X. Wang, and H. Lu, "Visual Tracking with Fully Convolutional Networks," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 3119–3127.

[23] 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.

[24] M. Danelljan, G. Bhat, F. S. 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. 6931–6939.

[25] F. Li, C. Tian, W. Zuo, L. Zhang, and M. Yang, "Learning Spatial-Temporal Regularized Correlation Filters for Visual Tracking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 4904–4913.

[26] Y. Li, C. Fu, Z. Huang, Y. Zhang, and J. Pan, "Intermittent Contextual Learning for Keyfilter-Aware UAV Object Tracking Using Deep Convolutional Feature," *IEEE Transactions on Multimedia*, pp. 1–13, 2020.

[27] L. Bertinetto, J. Valmadre, J. F. Henriques, A. Vedaldi, and P. H. S. Torr, "Fully-Convolutional Siamese Networks for Object Tracking," in *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, 2016, pp. 850–865.

[28] J. Valmadre, L. Bertinetto, J. Henriques, A. Vedaldi, and P. H. S. Torr, "End-to-End Representation Learning for Correlation Filter Based Tracking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 5000–5008.

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

[30] 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, pp. 1308–1317.

[31] 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. 375–391.

[32] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, *Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers*. Now Publishers, Inc., 2011, vol. 3, no. 1.

[33] 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)*, 2005, pp. 886–893.

[34] J. van de Weijer, C. Schmid, J. Verbeek, and D. Larlus, "Learning Color Names for Real-World Applications," *IEEE Transactions on Image Processing*, vol. 18, no. 7, pp. 1512–1523, 2009.
[35] 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.
[36] S. Li and D.-Y. Yeung, "Visual Object Tracking for Unmanned Aerial Vehicles: A Benchmark and New Motion Models," in *Proceedings of the Association for the Advancement of Artificial Intelligence (AAAI)*, 2017, p. 4140–4146.
[37] 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.
[38] M. Danelljan, G. Häger, F. Shahbaz Khan, and M. Felsberg, "Accurate Scale Estimation for Robust Visual Tracking," in *Proceedings of the British Machine Vision Conference (BMVC)*, 2014, pp. 1–11.
[39] 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, 2017.
[40] F. Li, Y. Yao, P. Li, D. Zhang, W. Zuo, and M. Yang, "Integrating Boundary and Center Correlation Filters for Visual Tracking with Aspect Ratio Variation," in *Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCVW)*, 2017, pp. 2001–2009.
[41] Y. Sun, C. Sun, D. Wang, Y. He, and H. Lu, "ROI Pooled Correlation Filters for Visual Tracking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 5776–5784.
[42] Z. Han, P. Wang, and Q. Ye, "Adaptive Discriminative Deep Correlation Filter for Visual Object Tracking," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 1, pp. 155–166, 2020.
[43] X. Lu, C. Ma, B. Ni, and X. Yang, "Adaptive Region Proposal with Channel Regularization for Robust Object Tracking," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–14, 2019.
[44] 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. 1387–1395.
[45] Y. Sui, G. Wang, and L. Zhang, "Joint Correlation Filtering for Visual Tracking," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 1, pp. 167–178, 2020.
[46] 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, pp. 2891–2900.
[47] Z. Kalal, K. Mikolajczyk, and J. Matas, "Forward-Backward Error: Automatic Detection of Tracking Failures," in *Proceedings of the International Conference on Pattern Recognition (ICPR)*, 2010, pp. 2756–2759.
[48] Z. Kalal, K. Mikolajczyk, and J. Matas, "Tracking-Learning-Detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 7, pp. 1409–1422, 2012.
[49] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object Detection with Discriminatively Trained Part-Based Models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1627–1645, 2010.
[50] C. Wang, L. Zhang, L. Xie, and J. Yuan, "Kernel cross-correlator," in *Proceedings of the Association for the Advancement of Artificial Intelligence (AAAI)*, 2018, pp. 4179–4186.
[51] C. Ma, J. Huang, X. Yang, and M. Yang, "Hierarchical Convolutional Features for Visual Tracking," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 3074–3082.
[52] T. Zhang, C. Xu, and M. 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. 4819–4827.
[53] L. Zhang and P. N. Suganthan, "Robust Visual Tracking via Co-trained Kernelized Correlation Filters," *Pattern Recognition*, vol. 69, pp. 82–93, 2017.
[54] K. Dai, D. Wang, H. Lu, C. Sun, and J. Li, "Visual Tracking via Adaptive Spatially-Regularized Correlation Filters," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 4665–4674.
[55] R. Tao, E. Gavves, and A. W. M. Smeulders, "Siamese Instance Search for Tracking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 1420–1429.
[56] Y. Qi, S. Zhang, L. Qin, H. Yao, Q. Huang, J. Lim, and M. Yang, "Hedged Deep Tracking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 4303–4311.
[57] D. Held, S. Thrun, and S. Savarese, "Learning to Track at 100 FPS with Deep Regression Networks," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016, pp. 749–765.
[58] S. Yun, J. Choi, Y. Yoo, K. Yun, and J. Y. Choi, "Action-Decision Networks for Visual Tracking with Deep Reinforcement Learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1349–1358.
[59] Q. Guo, W. Feng, C. Zhou, R. Huang, L. Wan, and S. Wang, "Learning Dynamic Siamese Network for Visual Object Tracking," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 1781–1789.

Fuling Lin received his B.Eng. degree in Mechanical Engineering from Tongji University, Shanghai, China. He is currently pursuing M.Sc. degree in Mechanical Engineering in Tongji University, Shanghai, China. His research interests include robotics, visual object tracking and computer vision.

Changhong Fu received his Ph.D. degree in Robotics and Automation from Computer Vision and Aerial Robotics (CVAR) Lab, Technical University of Madrid, Spain. During his Ph.D., he held two research positions at Arizona State University, USA and Nanyang Technological University (NTU), Singapore. After received his Ph.D., he worked at the NTU as Post-doc Research Fellow. He has worked on 2 international, 2 national and 4 industrial projects related to the vision for UAV. In addition, he has pubilshed more than 50 journal and conference papers (including IEEE TMM, IEEE TGRS, IEEE TMECH, IEEE TIE, CVPR, ICCV, ICRA and IROS) related to the intelligent vision and control for UAV. Currently, he is an Assistant Professor at School of Mechanical Engineering, Tongji University, China, and leading 6 projects related to the vision for Unmanned Systems (US). His research areas are Intelligent Vision and Control for US in Complex Environment.

Yujie He received his B.Eng. degree in Mechanical Engineering from Tongji University, Shanghai, China. He is currently pursuing M.Sc. degree in Robotics, École polytechnique fédérale de Lausanne (EPFL), Switzerland. His research interests include robotics, visual object tracking, and place recognition.

Fuyu Guo is currently a Ph.D student in the School of Mechanical Engineering, Chongqing University. He holds a B.Eng. in Mechanical Engineering from Northeastern University, China. In 2018 he was a visiting student for 6 months at Assoc. Prof. Pham Quang Cuong's CRI group in the School of Mechanical and Aerospace Engineering, Nanyang Technological University, Singapore. His research interests include robotic manipulation, intelligent perception.

Qian Tang received her Bachelor's, Master's and Ph.D. degrees in Chongqing University, China, respectively in 1991, 1994, and 1997. From 2003 to 2004, she was a visiting scholar at the Department of Mechanical and Industrial Engineering, the University of Toronto, Canada. She is a senior fellow of the Chinese Mechanical Engineering Society. Currently, she is a professor at College of Mechanical Engineering, Chongqing University, China. Her research interests include intelligent manufacturing and additive manufacturing.