ReCF: Exploiting Response Reasoning for Correlation Filters in Real-Time UAV Tracking

Fuling Lin, *Graduate Student Member, IEEE*, Changhong Fu, *Member, IEEE*, Yujie He, Weijiang Xiong, and Fan Li

***Abstract*— Object tracking is a fundamental task for the visual perception system on the intelligent unmanned aerial vehicle (UAV). The high efficiency of correlation filter (CF) based trackers has advanced the widespread development of online UAV object tracking. This kind of method can effectively train a filter to discriminate the target from the background. However, most CF-based methods require a fixed label function over all the previous samples, leading to over-fitting and filter degradation, especially in complex drone scenarios. To address this problem, a novel adaptive response reasoning approach is proposed for CF learning. It can leverage temporal information in filter training and significantly promote the robustness of the tracker. Specifically, the proposed response reasoning method goes beyond the standard response consistency requirement and constructs an auxiliary label of the current sample. Besides, it helps learn a generic relationship between the previous and current filters, thereby realizing self-regulated filter updating and enhancing the discriminability of the filter. Extensive experiments on four well-known challenging UAV tracking benchmarks with 278 videos sequences show that the presented method yields superior results to 40 state-of-the-art trackers with real-time performance on a single CPU, which is suitable for UAV online tracking missions.**

***Index Terms*— Real-time UAV tracking, discriminative correlation filter, adaptive response reasoning.**

I. Introduction

RECENT years have witnessed rapid progress in computer vision and artificial intelligence, and the unmanned aerial vehicle (UAV) equipped with intelligent vision-based perception technologies can help humans perform advanced analytical applications. Currently, object tracking for UAVs has become a research hotspot in the intelligent transportation field. It has made the execution of many practical tasks more accessible and efficient, such as path following [1], [2], traffic surveillance [3], [4], and object monitoring [5]. Though many object tracking approaches have made considerable progress recently, there are still many challenges in real-time UAV tracking scenarios.

Discriminative correlation filter (CF) based methods have attracted considerable attention in visual tracking. These approaches learn a CF by minimizing the least-squares loss for a set of shift samples generated by performing a cyclic shift operator on the training sample. Since its optimization problem can be solved by point-wise operations in the Fourier domain, CFs show the advantage of high computational efficiency and enable fast online learning and detection. Besides, online learning allows CFs to track arbitrary objects by constructing and maintaining the target appearance model.

Most CF-based trackers have achieved excellent performance [6]–[10]. The high efficiency owned by these methods makes them suitable for various scenarios that require real-time performance. Besides, they also have competitive CPU-based results on numerous generic benchmarks [11]–[13]. Due to the unique characteristics of the UAV tracking scenarios, benchmarks for UAV object tracking have gradually emerged [14]–[17]. In the aerial tracking perspective, the relative motion between the airborne camera and the target is often accompanied by lots of distinctive and challenging ingredients. Recently, several works [18]–[20] have carried out the research and design of the CF-based trackers for the UAV tracking scenes with great improvements. The CPU-based real-time tracking methods still necessitate investigations given the UAV's duration and the onboard processor's performance, and the CF is an effective and feasible research entry point.

In complex UAV tracking scenarios, the object appearance model maintained by CFs is frequently changing due to various challenging factors such as fast motion and viewpoint changes caused by the relative motion between the UAV and the object. Though the introduction of previous samples allows the filter to remember the historical details and makes the filter more discriminative, the substantial training samples stored continuously in the tracking process can put pressure on the computing resources. Moreover, the training labels for historical samples are identical and fixed, which means that the current filter's responses to all previous samples are close to the same label. The response consistency requirement introduces the risk of over-fitting the model and omits the impact of historical filters. On the other hand, the sole utilization of the desirable label for the current sample is difficult to resist fluctuations in the detection response. To this end, this work investigates the CPU-based real-time UAV tracking problem, emphasizing the use of historical information and the resistance of turbulent response.

In this paper, the response reasoning is exploited for correlation filters in UAV real-time tracking, which includes

Manuscript received 7 June 2020; revised 15 April 2021 and 14 June 2021; accepted 29 June 2021. Date of publication 14 July 2021; date of current version 9 August 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 61806148 and in part by the Natural Science Foundation of Shanghai under Grant 20ZR1460100. The Associate Editor for this article was L. M. Bergasa. *(Corresponding author: Changhong Fu.)*

The authors are with the School of Mechanical Engineering, Tongji University, Shanghai 201804, China (e-mail: changhongfu@tongji.edu.cn).

Digital Object Identifier 10.1109/TITS.2021.3094654

two components. Firstly, the design of historical response regularization exploits the previous sample to establish the adaptive training label for the historical sample, which realizes the approximation of using excessive historical samples. Secondly, the introduction of inferred response regularization can effectively suppress response fluctuations and enable the filter to have self-regulated update ability in collaboration with the proposed historical response regularization. Thus no additional learning rate is needed to maintain the apparent model. The primary contributions of the proposed approach are summarized as follows:

- A historical response regularization is proposed to eliminate the response consistency requirement by substituting the historical sample's fixed training labels, which requires only the last frame sample instead of numerous historical samples.
- An inferred response regularization is presented to assist the filter in suppressing turbulent response fluctuations enabling the self-regulated update capability with the historical response regularization and making the filter more robust to intricate UAV tracking scenarios.
- The proposed filter learning optimization problem can be decomposed into several subproblems and solved efficiently by the alternating direction method of multipliers (ADMM) [21] for closed-form solutions.
- Comprehensive experiments are performed on four UAV benchmarks to verify the proposed method, *i.e.*, UAV123@10fps [14], DTB70 [15], VisDrone2019-test-dev [16], and UAVDT [17]. Experimental results show that the proposed response reasoning-based correlation filter (ReCF) can achieve competitive performance against the other 40 state-of-the-art trackers, including those using handcrafted or convolutional features or trained by deep models. What is more, ReCF can operate with real-time efficiency on a single CPU.

Remark 1: As shown in Fig. 1, the proposed ReCF has the best precision and success rate with a high tracking speed of ∼70.8 frames per second (FPS).

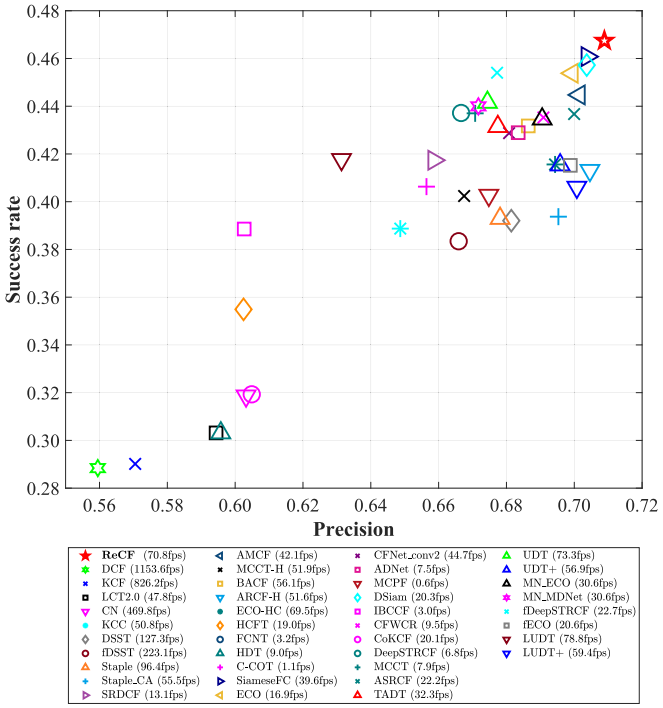


Fig. 1. Overall performance of the proposed ReCF and the other 40 advanced trackers on the UAVDT benchmark. The plot shows that the ReCF tracker has achieved the superiority in the precision, success rate, and operating efficiency.

II. Related Works

A. Correlation Filters for Visual Tracking

The objective of UAV tracking is to estimate the object's location and size in the video sequence captured by onboard cameras. Recently, the correlation filter based trackers have aroused great attention since their computational efficiency, which is suitable for UAV real-time applications. In the seminal work [22], correlation filters were learned by the minimum output sum of squared error. Upon the influential approach, J. F. Henriques *et al*. utilized the circulant structure of training samples [23] and applied the kernel trick to CFs [6] without additional computational expense comparing with linear CFs. Several works [6], [24]–[27] investigated the multi-channel visual cues in the CF learning, which enables the filter to apply more discriminative features. In [8], [28], a binary matrix was introduced to crop the central patch of each shifted samples to alleviate the boundary effects resulting from the periodic assumption [29]. H. Zhu *et al.* presented a bilateral weighted regression ranking method to make the filter more robust, avoiding the problems of model degradation and sample imbalance [30]. In [10], the cross-correlator breaks the limitation of circulant structure on training samples and allows the application of any kernel function. Several work [31]–[34] incorporated the accurate scale estimation in the CF to improve the tracking performance. In [35], the channel reliability was used as feature weights in localization. Since the CF works by constructing the appearance model, the excellent performance of convolutional neural networks has encouraged some works to extract deep features containing semantic and structural information to improve the filter's discriminative ability [9], [26], [36], [37]. In [7], a factorized convolution operator was used to build a compact set of filters and to reduce the computational complexity of the model.

However, the computationally expensive deep feature extraction leads to poor real-time tracking performance. In this work, only the combination of handcrafted features, *i.e.*, pixel intensity, histograms of oriented gradients (HOG) [38], and color names [39], is used to ensure the real-time UAV tracking.

B. Temporal Information for Visual Tracking

The limited set of training samples affects the model's generalization capacity because of the online nature of the tracking problem. Several works [9], [19], [29], [40]–[42] investigate the temporal information in the tracking process to alleviate the model degradation in the successive frames formed by continuous dynamic scenes. In [29], [42], the tracker utilized historical training samples in the filter learning to achieve a well-generalized and robust appearance model. The method

of SRDCFdecon [43] realized the dynamic management of the training set to alleviate the adverse impact of corrupted samples. Though these methods have achieved promising results, they remain some limitations.

Firstly, the tracking speed is limited by the size of the training set and all the historical samples in the training set have identical label. The resulting response consistency requirement introduces the over-fitting risk. Different label functions on different features have significantly impacted filter learning [44] though the weights of label functions are highly targeted and need to be carefully fine-tuned. On the contrary, in [45], the tracker utilizes the noise model to build a reference label for the training label, showing marked improvement. Therefore, a novel training label for the historical sample is introduced to reduce the number of historical samples and overcome the response consistency requirement in this work. The proposed label can be achieved by the historical response, which only needs the historical sample of the previous frame. The resulting historical response regularization can approximate the employment of multiple historical samples through the continuity of inter-frame information, which can greatly reduce computational consumption.

Secondly, the turbulent response fluctuations have a remarkable influence on the tracking performance, especially in the cluttered UAV scenes. Z. Huang *et al.* proposed a strategy of abnormal response suppression to enhance the filter's discriminative power [19]. The aberrance repressed strategy introduces an extra training label for the current sample in filter learning. The new label is built on the response in the detection stage, *i.e.,* the response of the historical filter to the detection sample. However, to ensure that the target position is aligned with the maximum value of the detection response in the filter learning, an additional shift operation is required for this detection response to obtain the required training label. Hence, to reduce the additional shift operation, this work proposes a novel training label based on the response of the previous filter to the previous training sample instead of the detection sample. Due to the prediction ability of the previous filter, the constraint with the proposed training label is called the inferred response regularization in this work.

III. Proposed Method

A. Revisit of Spatially Regularized Correlation Filters

To help better understand the proposed approach, the SRDCF [29] used as the baseline in this work is first briefly revisited. The SRDCF uses a training set $\{(\mathbf{X}^{(f)}, \mathbf{Y})\}_{f=1:k}$ to learn the filter $\mathbf{W}^{(k)}$ at frame k. $\mathbf{X}^{(f)} = [\mathbf{x}_1^{(f)}, \mathbf{x}_2^{(f)}, \ldots, \mathbf{x}_D^{(f)}]$ represents the features of all D channels extracted from the corresponding frame f and each channel is a vectorized feature, *i.e.*, $\mathbf{x}_d^{(f)} \in \mathbb{R}^{N\times 1}$. $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_D]$ is the Gaussian label and has the same size as $\mathbf{X}^{(f)}$. $\mathbf{W}^{(f)} = [\mathbf{w}_1^{(f)}, \mathbf{w}_2^{(f)}, \ldots, \mathbf{w}_D^{(f)}]$ denotes the filter with D channels. The optimization problem of SRDCF is to minimize the regression error:

$$\varepsilon = \sum_{f=1}^{k} \alpha^{(f)} \left\| \mathbf{y}_d - \sum_{d=1}^{D} \mathbf{w}_d^{(k)} \star \mathbf{x}_d^{(f)} \right\|_2^2 + \sum_{d=1}^{D} \left\| \mathbf{s} \odot \mathbf{w}_d^{(k)} \right\|_2^2, \quad (1)$$

where $\star$ and $\odot$ denotes the circular correlation and Hadamard product, respectively. $\alpha^{(f)}$ is a weight coefficient to emphasize the most recent samples, and $\mathbf{s}$ is the spatial regularization weights proposed in SRDCF. The spatial regularizer $\mathbf{s}$ helps the filter to focus on the reliable region close to the target center. It suppresses the filter coefficients close to the target edge, which makes the filter more robust in the complex environment, *e.g.*, viewpoint changes, and fast motion. Though SRDCF can efficiently alleviate the boundary effects, the use of most historical samples results in a high computational burden. The optimization of Eq. (1) tries to make the historical samples' responses close to the same label $\mathbf{Y}$.

Remark 2: The response consistency requirement introduces the risk of over-fitting and can lead to tracking drift. Moreover, SRDCF discards the impact of the historical filters, which can be used to construct a prior response incorporating the target spatial information.

B. Response Reasoning

1) Historical Response Regularization: To circumvent the response consistency requirement of the previous samples, the information from the previous frame is used to make the response template contain more spatial information than the ideal label. In other words, the label to the previous training sample goes beyond a fixed ideal response $\mathbf{Y}$ and becomes the response between the previous filter $\mathbf{W}^{(k-1)}$ and training features $\mathbf{X}^{(k-1)}$ at the $(k-1)$-th frame. Therefore, the historical response regularization is expressed as follows:

$$e_H = \gamma_H \sum_{d=1}^{D} \left\| \tilde{\mathbf{y}}_d^{(k-1)} - \mathbf{w}_d \star \mathbf{x}_d^{(k-1)} \right\|_2^2, \quad (2)$$

where γ_H is the historical response regularization weight and $\tilde{\mathbf{y}}_d^{(k-1)} = \mathbf{w}_d^{(k-1)} \star \mathbf{x}_d^{(k-1)}$. Compared with the first term in Eq. (1) regarding the samples $\{\mathbf{X}^{(f)}\}_{f=1:k-1}$, the proposed regularization changes the training label of the historical samples from the fixed ideal label $\mathbf{Y}$ to the label $\tilde{\mathbf{Y}} = [\tilde{\mathbf{y}}_1, \tilde{\mathbf{y}}_2, \ldots, \tilde{\mathbf{y}}_D]$, which is built by the filter and training sample from the previous frame. Since the filter $\mathbf{W}^{(k-1)}$ is constructed based on the appearance model, its response $\tilde{\mathbf{Y}}$ to the corresponding sample $\mathbf{X}^{(k-1)}$ can be more realistic than the fixed single centered Gaussian $\mathbf{Y}$ and contain more spatial information about the target, such as the historical response $\tilde{\mathbf{Y}}^{(k-1)}$ shown Fig. 2. Since the introduction of the historical filter $\mathbf{W}^{(k-1)}$ preserves the continuity of inter-frame information, only the previous sample is used in the filter learning.

2) Inferred Response Regularization: In addition, the prediction ability of the previous filter $\mathbf{W}^{(k-1)}$ is utilized to construct the inferred response regularization, which can efficiently repress the turbulent response fluctuations. In classical correlation filters, the previous filter is used to compute the response map of the detection sample $\mathbf{Z}^{(k)} = [\mathbf{z}_1^{(k)}, \mathbf{z}_2^{(k)}, \ldots, \mathbf{z}_D^{(k)}]$, which is cropped through the target center and size at frame $k-1$. The detection response $\mathbf{R}_s^{(k)}$ is obtained by:

$$\mathbf{R}_s^{(k)} = \sum_{d=1}^{D} \mathbf{r}_d^{(k)} = \sum_{d=1}^{D} \mathbf{w}_d^{(k-1)} \star \mathbf{z}_d^{(k)}. \quad (3)$$

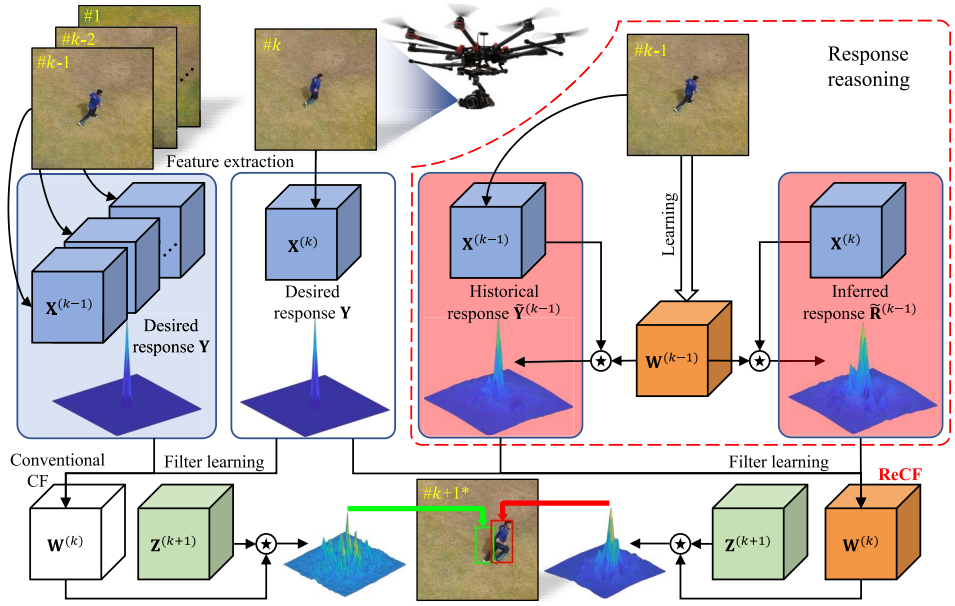


Fig. 2. Main difference between the proposed ReCF and SRDCF. SRDCF uses the previous samples $\{(\mathbf{X}^{(f)}, \mathbf{Y})\}_{f=1:k-1}$ to learn filters based on the historical response consistency requirement. In the proposed response reasoning, our method utilizes the historical filter $\mathbf{W}^{(k-1)}$ to build a training label $\tilde{\mathbf{Y}}$ corresponding to the past feature $\mathbf{X}^{(k-1)}$. Furthermore, the ReCF constructs the auxiliary inferred response $\tilde{\mathbf{R}}^{(k-1)}$ with the prediction ability of $\mathbf{W}^{(k-1)}$. $\#k+1^*$ means the detection patch ($\mathbf{Z}^{(k+1)}$) cropped from the $(k+1)$-th frame.

By searching the highest value of $\mathbf{R}_s^{(k)}$, the relative displacement of the object can be inferred. In this work, the inferred response $\tilde{\mathbf{R}}^{(k)} = [\tilde{\mathbf{r}}_1^{(k)}, \tilde{\mathbf{r}}_2^{(k)}, \ldots, \tilde{\mathbf{r}}_D^{(k)}]$ is built by the filter $\mathbf{W}^{(k-1)}$ and the current training sample $\mathbf{X}^{(k)}$, *i.e.*,

$$e_I = \gamma_I \sum_{d=1}^{D} \left\| \tilde{\mathbf{r}}_d^{(k)} - \mathbf{w}_d \star \mathbf{x}_d^{(k)} \right\|_2^2, \tag{4}$$

where γ_I is the inferred response regularization weight and $\tilde{\mathbf{r}}_d^{(k)} = \mathbf{w}_d^{(k-1)} \star \mathbf{x}_d^{(k)}$. Compared with the standard filter learning using only the single centered Gaussian $\mathbf{Y}$ as the training label of $\mathbf{X}^{(k)}$, the proposed regularization introduces the target prior spatial information, *i.e.,* the inferred response $\tilde{\mathbf{R}}^{(k)}$ depicted in Fig. 2, in the filter learning.

Remark 3: The additional training label in [19] for response regularization is constructed by shifting the detection response, that is, the correlation response between the previous filter $\mathbf{W}^{(k-1)}$ and the detection sample $\mathbf{Z}^{(k)}$. In contrast, the proposed inferred response regularization in this work employs the previous filter $\mathbf{W}^{(k-1)}$ and current training sample $\mathbf{X}^{(k)}$ to build the response regularization without the extra shift operation.

Furthermore, the combination of the historical and inferred regularization enables the self-regulated filter updating, making it more robust against the fast motion, illumination variations, and viewpoint changes. Overall, only the previous filter $\mathbf{W}^{(k-1)}$ and features $\mathbf{X}^{(k-1)}$ are applied to decrease onboard computational burden and memory consumption significantly.

C. ReCF Learning

Following [35], the overall optimization problem is to minimize the training error ε in channel independent form:

$$\varepsilon = \sum_{d=1}^{D} \left\| \mathbf{y} - \mathbf{w}_d \star \mathbf{x}_d^{(k)} \right\|_2^2 + \sum_{d=1}^{D} \|\mathbf{s} \odot \mathbf{w}_d\|_2^2 + e_H + e_I. \tag{5}$$

Therefore, Eq. (5) can be decomposed into D subproblems since they are independent across all channels and the d-th channel is chosen to derive. For simplified presentation, the subscript $(\cdot)_d$ is omitted in the following derivation. The d-th subproblems is expressed as follows:

$$\begin{aligned} \min_{\mathbf{w}} \ & \left\| \mathbf{y} - \mathbf{w} \star \mathbf{x}^{(k)} \right\|_2^2 + \|\mathbf{s} \odot \mathbf{w}\|_2^2 \\ & + \gamma_H \left\| \mathbf{w}^{(k-1)} \star \mathbf{x}^{(k-1)} - \mathbf{w} \star \mathbf{x}^{(k-1)} \right\|_2^2 \\ & + \gamma_I \left\| \mathbf{w}^{(k-1)} \star \mathbf{x}^{(k)} - \mathbf{w} \star \mathbf{x}^{(k)} \right\|_2^2. \end{aligned} \tag{6}$$

Here an auxiliary variable $\mathbf{h}$ is introduced by requiring $\mathbf{h} = \mathbf{w}$ to make Eq. (6) become an equality constrained optimization problem, *i.e.*,

$$\begin{aligned} \min_{\mathbf{w}} \ & \left\| \mathbf{y} - \mathbf{w} \star \mathbf{x}^{(k)} \right\|_2^2 + \|\mathbf{s} \odot \mathbf{h}\|_2^2 \\ & + \gamma_H \left\| \mathbf{w}^{(k-1)} \star \mathbf{x}^{(k-1)} - \mathbf{w} \star \mathbf{x}^{(k-1)} \right\|_2^2 \\ & + \gamma_I \left\| \mathbf{w}^{(k-1)} \star \mathbf{x}^{(k)} - \mathbf{w} \star \mathbf{x}^{(k)} \right\|_2^2, \\ s.t. \ & \mathbf{w} = \mathbf{h}. \end{aligned} \tag{7}$$

In Eq. (7), the circular correlation is computationally demanding while it can be computed efficiently using element-wise operations in the Fourier domain. By applying the Parseval's theorem, Eq. (7) can be transferred to the Fourier domain:

$$\begin{aligned} \min_{\hat{\mathbf{w}}} \ & \left\| \hat{\mathbf{y}} - \bar{\hat{\mathbf{w}}} \odot \hat{\mathbf{x}}^{(k)} \right\|_2^2 + \|\mathbf{s} \odot \mathbf{h}\|_2^2 \\ & + \gamma_H \left\| \bar{\hat{\mathbf{w}}}^{(k-1)} \odot \hat{\mathbf{x}}^{(k-1)} - \bar{\hat{\mathbf{w}}} \odot \hat{\mathbf{x}}^{(k-1)} \right\|_2^2 \\ & + \gamma_I \left\| \bar{\hat{\mathbf{w}}}^{(k-1)} \odot \hat{\mathbf{x}}^{(k)} - \bar{\hat{\mathbf{w}}} \odot \hat{\mathbf{x}}^{(k)} \right\|_2^2, \\ s.t. \ & \hat{\mathbf{w}} - \sqrt{N}\mathbf{F}\mathbf{h} = 0. \end{aligned} \tag{8}$$

The $\hat{}$ denotes the discrete Fourier transformation and $\bar{}$ represents the complex conjugation. By introducing the Lagrange multiplier $\hat{\zeta}$, the objective function in Eq. (8) can be formulated as the augmented Lagrangian form:

$$\begin{aligned}\mathcal{L}(\hat{\mathbf{w}}, \hat{\zeta}) = &\left\|\hat{\mathbf{y}} - \bar{\hat{\mathbf{w}}} \odot \hat{\mathbf{x}}^{(k)}\right\|_2^2 + \|\mathbf{s} \odot \mathbf{h}\|_2^2 \\ &+ \mu \left\|\hat{\mathbf{w}} - \sqrt{N}\mathbf{F}\mathbf{h} + \frac{1}{\mu}\hat{\zeta}\right\|_2^2 \\ &+ \gamma_H \left\|\bar{\hat{\mathbf{w}}}^{(k-1)} \odot \hat{\mathbf{x}}^{(k-1)} - \bar{\hat{\mathbf{w}}} \odot \hat{\mathbf{x}}^{(k-1)}\right\|_2^2 \\ &+ \gamma_I \left\|\bar{\hat{\mathbf{w}}}^{(k-1)} \odot \hat{\mathbf{x}}^{(k)} - \bar{\hat{\mathbf{w}}} \odot \hat{\mathbf{x}}^{(k)}\right\|_2^2, \end{aligned} \tag{9}$$

where μ is the penalty factor.

Remark 4: The alternating direction method of multipliers (ADMM) [21] is used to solve the minimization of Eq. (9), which decomposes the optimization problem into two subproblems and optimizes them at each iteration.

1) Subproblem $\hat{\mathbf{w}}$*:* The solution to $\hat{\mathbf{w}}^{i+1}$ at the $(i+1)$-th ADMM iteration can be solved by fixing $\mathbf{h}$ and $\hat{\zeta}$ in Eq. (9). Thus the optimization of $\hat{\mathbf{w}}^{i+1}$ is formulated as:

$$\begin{aligned}\hat{\mathbf{w}}^{i+1} = \arg\min_{\hat{\mathbf{w}}} \Bigg\{ &\left\|\hat{\mathbf{y}} - \bar{\hat{\mathbf{w}}} \odot \hat{\mathbf{x}}^{(k)}\right\|_2^2 + \mu \left\|\hat{\mathbf{w}} - \sqrt{N}\mathbf{F}\mathbf{h} + \frac{1}{\mu}\hat{\zeta}\right\|_2^2 \\ &+ \gamma_H \left\|\bar{\hat{\mathbf{w}}}^{(k-1)} \odot \hat{\mathbf{x}}^{(k-1)} - \bar{\hat{\mathbf{w}}} \odot \hat{\mathbf{x}}^{(k-1)}\right\|_2^2 \\ &+ \gamma_I \left\|\bar{\hat{\mathbf{w}}}^{(k-1)} \odot \hat{\mathbf{x}}^{(k)} - \bar{\hat{\mathbf{w}}} \odot \hat{\mathbf{x}}^{(k)}\right\|_2^2 \Bigg\}. \end{aligned} \tag{10}$$

By minimizing the objective function in Eq. (10), a closed-form solution to $\hat{\mathbf{w}}^{i+1}$ can be achieved:

$$\hat{\mathbf{w}}^{i+1} = \frac{\hat{S}_{\mathbf{xy}}^{(k)} + \left(\gamma_I \hat{S}_{\mathbf{xx}}^{(k)} + \gamma_H \hat{S}_{\mathbf{xx}}^{(k-1)}\right) \odot \hat{\mathbf{w}}^{(k-1)} + \mu\sqrt{N}\mathbf{F}\mathbf{h}^i - \hat{\zeta}^i}{(1+\gamma_I)\hat{S}_{\mathbf{xx}}^{(k)} + \gamma_H \hat{S}_{\mathbf{xx}}^{(k-1)} + \mu}, \tag{11}$$

where $\hat{S}_{\mathbf{xy}}^{(k)} = \hat{\mathbf{x}}^{(k)} \odot \bar{\hat{\mathbf{y}}}$ and $\hat{S}_{\mathbf{xx}}^{(k)} = \hat{\mathbf{x}}^{(k)} \odot \bar{\hat{\mathbf{x}}}^{(k)}$. The complexity of Eq. (11) is $\mathcal{O}(N)$, where N is the signal length.

2) Subproblem $\mathbf{h}$*:* If $\hat{\mathbf{w}}$ and $\hat{\zeta}$ are fixed in Eq. (9), $\mathbf{h}^{i+1}$ can be solved by the following optimization problem:

$$\mathbf{h}^{i+1} = \arg\min_{\mathbf{h}} \|\mathbf{s} \odot \mathbf{h}\|_2^2 + \mu \left\|\hat{\mathbf{w}} - \sqrt{N}\mathbf{F}\mathbf{h} + \frac{1}{\mu}\hat{\zeta}\right\|_2^2. \tag{12}$$

By taking the derivative of the objective function in Eq. (12) regarding $\mathbf{h}$ to zero, a closed-form solution to $\mathbf{h}^{i+1}$ can be obtained:

$$\mathbf{h}^{i+1} = \frac{\mathcal{F}^{-1}\left(\mu\hat{\mathbf{w}}^{i+1} + \hat{\zeta}^i\right)}{\frac{1}{N}(\mathbf{s} \odot \mathbf{s}) + \mu}, \tag{13}$$

where $\mathcal{F}^{-1}$ denotes the inverse discrete Fourier transformation. The cost of computing Eq. (13) is bounded by $\mathcal{O}(N\log(N))$, *i.e.,* the complexity of the inverse fast Fourier transformation for a given signal of length N.

In the last step of the $i+1$-th iteration, the Lagrangian multiplier $\hat{\zeta}^{i+1}$ is updated by

$$\hat{\zeta}^{i+1} = \mu\left(\hat{\mathbf{w}}^{i+1} - \hat{\mathbf{h}}^{i+1}\right). \tag{14}$$

The penalty factor μ at the $i+1$-th iteration is updated by $\mu^{i+1} = \min(\mu_{\max}, \beta\mu^i)$, where β is a pre-set step size factor. As a result, the complexity of our ReCF tracker is bounded by $\mathcal{O}(N_{\text{ADMM}} D N \log(N))$, where D and N_{ADMM} denote the number of filter channels and ADMM iterations, respectively.

In the detection step of frame $k+1$, the CF obtains the features $\mathbf{z}^{(k+1)}$ of the searching region according to the location and size of the target at frame k. Then the CF is applied to compute the spatial response $\mathbf{r}^{(k+1)}$ by:

$$\mathbf{r}^{(k+1)} = \mathcal{F}^{-1}\left(\bar{\hat{\mathbf{w}}}^{(k)} \odot \hat{\mathbf{z}}^{(k+1)}\right). \tag{15}$$

Remark 5: By searching the maximum value in the spatial response map, the CF can estimate the relative displacement between the frame k and $k+1$.

IV. Experiments

A. Experimental Setup

1) Evaluation Methodology: In this work, two metrics of precision and success rate are employed to evaluation all trackers by the one-pass evaluation protocol [14]. The precision metric is used to measure the center location error (CLE) between the estimated bounding box and the ground-truth bounding box. For the success rate, it measures the overlap of the estimated bounding box and the ground-truth bounding box. For better results visualization, the precision plot is applied to show the percentage of frames where the CLE is within a given threshold. The success plot is also used to visualize the percentage of frames where the overlap exceeds a given threshold.

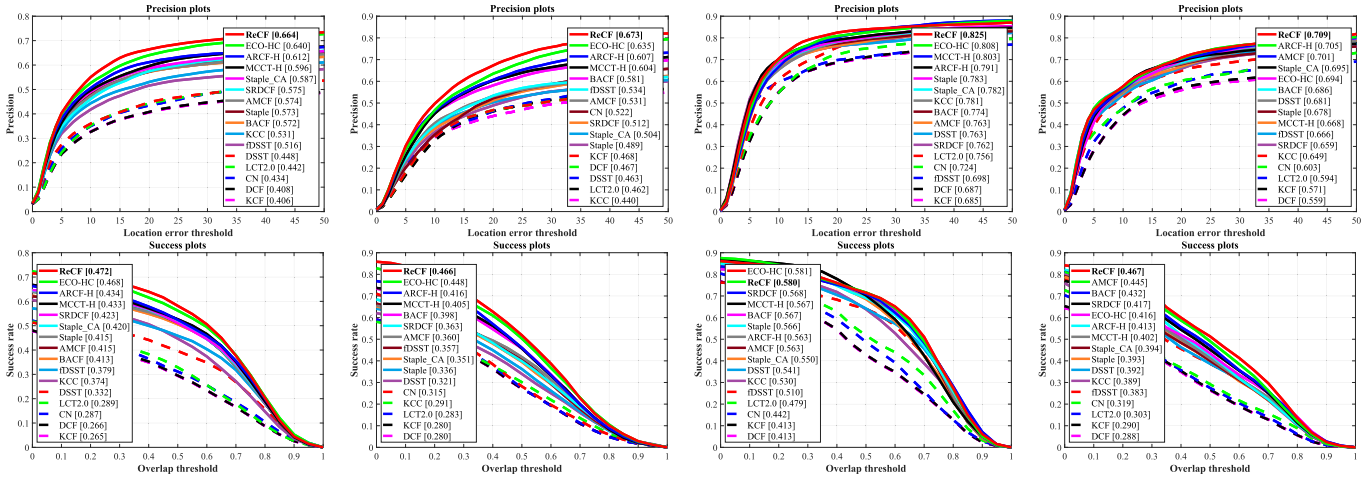
Remark 6: In the precision plot, trackers are ranked by the distance precision (DP) at a conventional CLE threshold of 20 pixels. Moreover, trackers in the success plot are ranked by the area under the curve (AUC). For a complete introduction to these metrics, please refer to [14].

2) Implementation Details: The ideal response map is built by a 2D Gaussian function with a bandwidth of $\sqrt{wh}/16$ where (w, h) denotes the object size. The handcrafted features, *i.e.*, grayscale, HOG, and CN, are used to represent the object. The regularization factors, γ_I and γ_H in Eq. (11), are set to 102.2 and 28, respectively. The iteration number N_{ADMM} is 3 for the ADMM optimization. The initial and maximum penalty factors, μ^0 and $\mu_{\max}$, are set to 10^2 and 10^5, and the scale step β is set to 500. Following [7], an additional correlation filter is utilized for scale estimation. All hyper-parameters remain fixed for all experiments.

Remark 7: The proposed method is implemented in MATLAB R2019a on a computer with an i7-8700K (3.70GHz) CPU. The source code of ReCF is available at https://github.com/vision4robotics/ReCF-Tracker. Other trackers used for comparisons are evaluated on the same computer using the open-source code provided by the original authors to ensure fair evaluations.

B. Comparison With Handcrafted-Based Trackers

The proposed method is evaluated extensively on four challenging UAV benchmarks, *i.e.*, UAV123@10fps [14] with 123 videos, DTB70 [15] with 70 videos, VisDrone2019-test-dev [16] with 35 videos, and UAVDT [17] with 50 videos, with



(a) UAV123@10fps (b) DTB70 (c) VisDrone2019-test-dev (d) UAVDT

Fig. 3. Precision and success plots of ReCF, the baseline SRDCF, and the other 14 real-time handcrafted based trackers on (a) UAV123@10fps, (b) DTB70, (c) VisDrone2019-test-dev, and (d) UAVDT. The precision (@20 pixels) and AUC scores of success plots are given in brackets. The proposed ReCF has obtained the competitive precision and success rate in all four challenging benchmarks.

TABLE I

AVERAGE PRECISION AT 20 PIXELS (DP), SUCCESS RATE (AUC), AND TRACKING SPEED (FPS) OF TOP-12 HANDCRAFTED BASED TRACKERS ARE REPORTED BY AVERAGING THE RESULTS OF ALL FOUR BENCHMARKS. THE **RED**, **GREEN**, AND **BLUE** FONTS INDICATE THE BEST THREE RESULTS. ALL EXPERIMENTAL RESULTS ARE GENERATED BY THE SAME COMPUTER FOR FAIR COMPARISONS

	DSST	KCC	fDSST	Staple	Staple_CA	AMCF	BACF	MCCT-H	ARCF-H	ECO-HC	SRDCF	**ReCF**
Avg. DP	0.589	0.600	0.604	0.631	0.642	0.643	0.653	0.667	**0.678**	**0.695**	0.627	**0.718**
Avg. AUC	0.397	0.396	0.407	0.428	0.429	0.446	0.452	0.452	**0.457**	**0.478**	0.443	**0.496**
Avg. FPS	**83.6**	38.2	**162.8**	**91.0**	51.0	38.9	41.8	48.3	39.9	60.1	10.2	60.2

the baseline tracker SRDCF [29] and the other 14 real-time trackers (with a speed of exceeding 30 FPS) using handcrafted features, including DSST [32], CN [24], LCT2.0 [46], DCF, KCF [6], KCC [10], fDSST [33], Staple [25], Staple_CA [47], BACF [8], MCCT-H [26], ECO-HC [7], ARCF-H [19], AMCF [20].

1) Overall Evaluation: Results in Fig. 3 demonstrate that the proposed method achieves the competitive precision and AUC scores over all four UAV benchmarks. On UAV123@10fps [14], the presented approach provides the highest precision and AUC score, outperforming the second best tracker by 2.4% and 0.4%, respectively. It is worth pointing out that the UAV123@10fps benchmark generated at 10 FPS makes tracking more difficult because of the larger inter-frame displacement of the object. On DTB70 [15], the ReCF achieves the best precision of 0.673 and the highest AUC score of 0.466. On VisDrone2019-test-dev [16], ReCF performs 1.7% better than the second best tracker in precision, and the AUC score of ReCF is rather close to the best score. ReCF also has the best precision and AUC score of the comparison on the UAVDT [17] benchmark. The evaluation demonstrates that compared with SRDCF the proposed method can significantly improve the UAV tracking performance. It also shows the importance of the response reasoning, which helps learn a more robust CF from the handcrafted features and historical information. Besides, the fixed learning rate in standard CFs does not guarantee optimal tracking for all sequences, especially for the complex UAV scenarios. By response reasoning, the proposed method can adaptively update the filter without a fixed learning rate and obtain better robustness.

Remark 8: Table I provides the average tracking speed of top-12 ranked trackers running on a single CPU over all four benchmarks. fDSST has the highest speed, followed by Staple and DSST. Our method has an average tracking speed of 60.2 FPS, meeting the real-time UAV tracking requirement. Some qualitative results are illustrated in Fig. 4.

2) Attribute-Based Evaluation: UAV123@10fps, DTB70, VisDrone2019-test-dev, and UAVDT are fully annotated by 12, 11, 12, and 9 challenging visual attributes, respectively, *e.g.*, fast motion (FM), illumination variation (IV), viewpoint change (VC), background clutter (BC), motion blur (MB), camera motion (CM), and object motion(OM). Thanks to the proposed response reasoning module, the ReCF tracker has achieved satisfactory results in most visual attributes compared with other trackers. Some examples of success plots are provided in Fig. 5. On UAV123@10fps, ReCF exceeds the second best tracker (ECO-HC) by 0.7% in FM and outperforms the second best tracker (ECO-HC) by 1.9% in IV. For MB, FM, and BC on the DTB70 benchmark, ReCF has the best performance and remarkable improvement against other top trackers. Furthermore, the experimental attribute-based results on VisDrone2019-test-dev (in CM, FM, and BC) and UAVDT (in CM, OM, and BC) demonstrate that ReCF can perform better than other trackers in UAV scenarios when encountering fast motion and background clutter attributes.

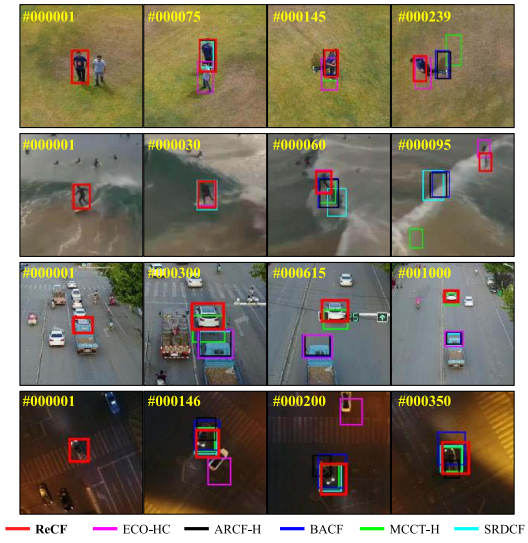


Fig. 4. Qualitative performance evaluation of ReCF, the baseline SRDCF, and the other top-4 ranked trackers (ECO-HC, ARCF-H, BACF, MCCT-H). The first, second, third, and fourth rows depict the tracking results on *person11* from UAV123@10fps, *Surfing06* from DTB70, *uav0000164_00000_s* from VisDrone2019-test-dev, and *S1301* from UAVDT. More UAV tracking videos are available at https://youtu.be/T31ueXiqzyk.

C. Comparison With Deep-Based Trackers

For further evaluation, the proposed method is compared with the other 25 state-of-the-art trackers on UAVDT, including HCFT [36], FCNT [48], HDT [49], C-COT [42], ECO [7], CFNet [50], ADNet [51], MCPF [52], IBCCF [34], CoKCF [53], DeepSTRCF [9], MCCT [26], ASRCF [54], TADT [55], UDT [56], UDT+ [56], SiameseFC [57] DSiam [58], CFWCR [59], MN_ECO [60], MN_MDNet [60], fDeepSTRCF [61], fECO [61], LUDT [62], and LUDT+ [62]. Table II shows that the proposed method has the best precision and AUC scores against the other trackers. More specifically, in the AUC score, the proposed method outperforms SiameseFC and DSiam by 0.6% and 1.0%, respectively. The evaluation highlights the competitive performance of these trackers in the UAV benchmarks and demonstrates the proposed method using only handcrafted features can perform well in the complex UAV scenarios. It is noted that the deep-based trackers using GPUs for training or extracting convolutional features show high performance while coming at the cost of tracking speed.

Remark 9: On UAVDT, the proposed method can reach a real-time speed of 70.8 FPS on a single CPU, which is 1.8× faster than SiameseFC using GPUs. The experimental results show that the proposed method can be applied efficiently in the online UAV tracking applications.

D. Ablation Study of Response Reasoning

In this section, the ablation study of the presented response reasoning is evaluated systematically over all four UAV benchmarks. The ReCF-N represents the tracker without response reasoning. By introducing the previous training sample in filter learning based on the historical response regularization (HRR), the ReCF-HRR can be obtained. As shown in Table III,

TABLE II

COMPARISONS OF THE PROPOSED RECF AND THE OTHER 25 DEEP-BASED TRACKERS ON UAVDT. THE **RED**, **GREEN**, AND **BLUE** FONTS SHOW THE BEST THREE RESULTS, RESPECTIVELY. RECF ACHIEVES THE HIGHEST SCORES AND HAS A REAL-TIME SPEED OF 70.8 FPS ON A SINGLE CPU

Tracker	Venue	DP	AUC	FPS	GPU
HCFT	15'ICCV	0.602	0.355	19.0	✓
FCNT	15'ICCV	0.656	0.245	3.2	✓
HDT	16'CVPR	0.596	0.303	9.0	✓
C-COT	16'ECCV	0.656	0.406	1.1	✓
SiameseFC	16'ECCVW	**0.704**	**0.461**	39.6	✓
ECO	17'CVPR	0.699	0.454	16.9	✓
CFNet_conv2	17'CVPR	0.681	0.429	44.7	✓
ADNet	17'CVPR	0.683	0.429	7.5	✓
MCPF	17'CVPR	0.675	0.403	0.6	✓
DSiam	17'ICCV	**0.704**	**0.457**	20.3	✓
IBCCF	17'ICCVW	0.603	0.389	3.0	✓
CFWCR	17'ICCVW	0.691	0.435	9.5	✓
CoKCF	17'PR	0.605	0.319	20.1	✓
DeepSTRCF	18'CVPR	0.667	0.437	6.8	✓
MCCT	18'CVPR	0.671	0.437	7.9	✓
ASRCF	19'CVPR	0.700	0.437	22.2	✓
TADT	19'CVPR	0.677	0.431	32.3	✓
UDT	19'CVPR	0.674	0.442	**73.3**	✓
UDT+	19'CVPR	0.696	0.415	56.9	✓
MN_ECO	20'ACM	0.691	0.435	30.6	✓
MN_MDNet	20'ACM	0.672	0.440	30.6	✓
fDeepSTRCF	20'TIP	0.677	0.454	22.7	✓
fECO	20'TIP	0.699	0.415	20.6	✓
LUDT	21'IJCV	0.631	0.418	**78.8**	✓
LUDT+	21'IJCV	0.701	0.406	59.4	✓
ReCF	Ours	**0.709**	**0.467**	**70.8**	✗

TABLE III

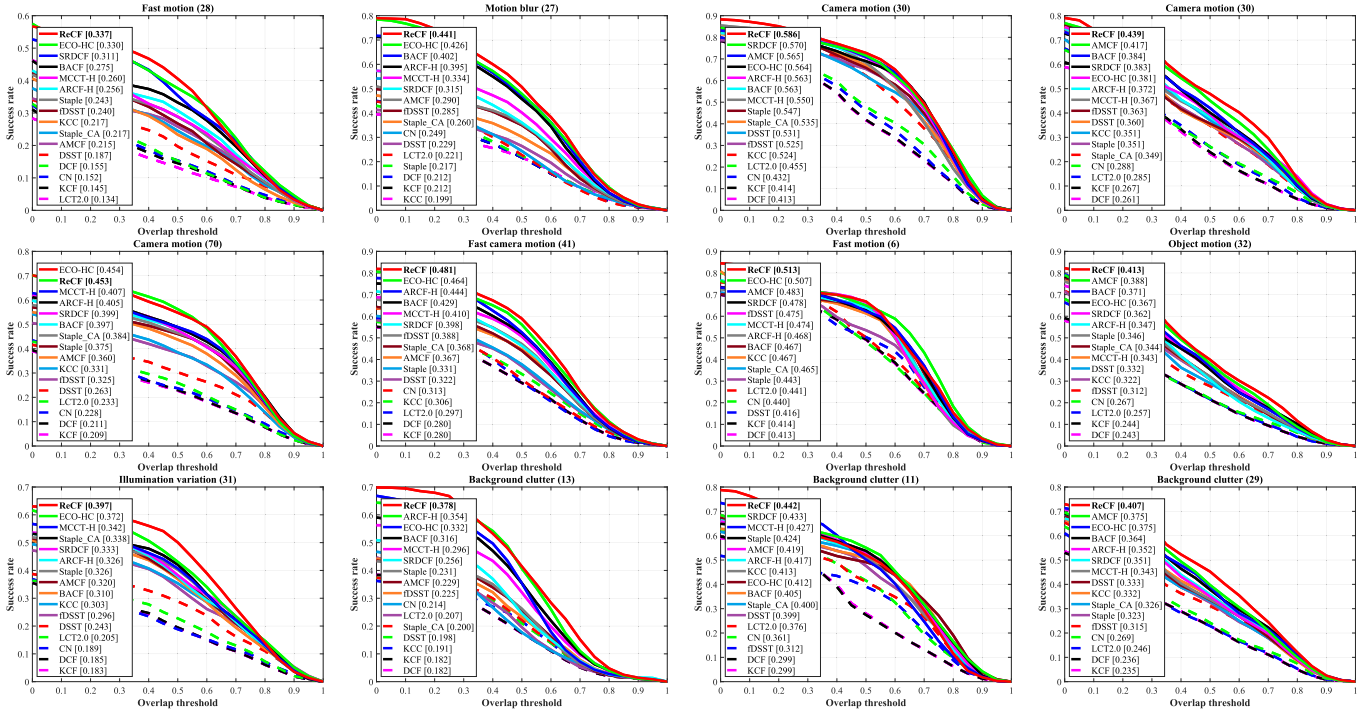
ABLATION STUDY OF THE PROPOSED RESPONSE REASONING. THE PRECISION (@20 PIXELS) AND SUCCESS RATE (AUC SCORES) ARE PRESENTED ON UAV123@10FPS, DTB70, AND UAVDT

Tracker	DP			
	UAV123@10fps	DTB70	VisDrone2019-test-dev	UAVDT
ReCF-N	0.364	0.462	0.302	0.245
ReCF-HRR	0.552	0.600	0.536	0.543
ReCF-IRR	0.650	0.672	0.788	0.691
ReCF (Final)	**0.664**	**0.673**	**0.825**	**0.709**

Tracker	AUC			
	UAV123@10fps	DTB70	VisDrone2019-test-dev	UAVDT
ReCF-N	0.242	0.322	0.225	0.134
ReCF-HRR	0.390	0.415	0.369	0.349
ReCF-IRR	0.465	0.463	0.563	0.451
ReCF (Final)	**0.472**	**0.466**	**0.581**	**0.467**

ReCF-HRR has a significant improvement compared to the ReCF-N tracker in precision and AUC scores on all benchmarks, showing the validity of HRR module. The ReCF-IRR tracker denotes the ReCF-N track with the inferred response regularization (IRR). Compared with ReCF-N, ReCF-IRR also obtains remarkable gain in Table III, indicating the effectiveness of introducing IRR module. ReCF represents the ReCF-N with the complete response reasoning module, *i.e.,* equipped with both HRR and IRR, which has the best precision and AUC scores.

Remark 10: The main reason for the improvements is originated from the introduction of HRR and IRR, which makes the filter break through the limitations brought by the response consistency requirement and have the capacity to resist turbulent response fluctuations. As a result, the integration of the historical and inferred response regularization



(a) UAV123@10fps (b) DTB70 (c) VisDrone2019-test-dev (d) UAVDT

Fig. 5. Success plots of the proposed ReCF, the baseline SRDCF, and the other 14 real-time handcrafted based trackers in different attributes on each benchmark. The AUC scores are given in brackets. The results show that the proposed ReCF achieves the best AUC scores in most attributes.

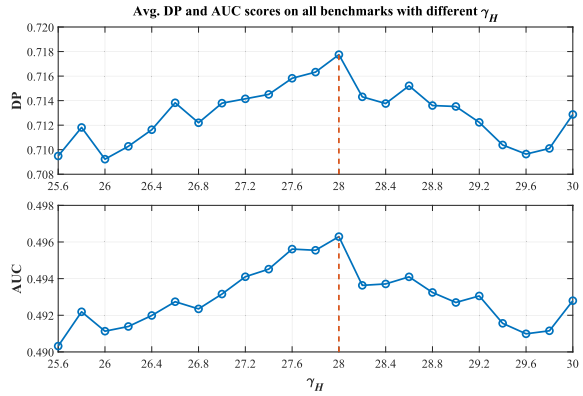


Fig. 6. Impact of historical response regularization term γ_H on all four benchmarks.

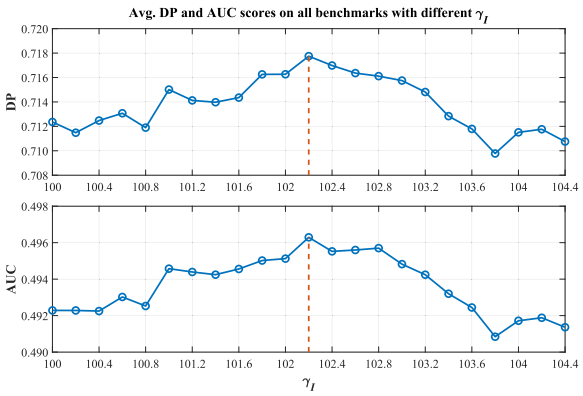


Fig. 7. Impact of inferred response regularization term γ_I on all four benchmarks.

enables the filter to have self-regulated update ability and can realize the state-of-the-art tracking performance.

E. Analysis of Key Parameters

1) γ_H in Historical Response Regularization: The first experiment is conducted to study the influence of the term γ_H in the historical response regularization. Different γ_H from 25.6 to 30 with a step size of 0.2 are analyzed on all the four benchmark. In this experiment, the factor γ_I in the inferred response regularization is set to 102.2 and remains fixed. Figure 6 shows that the performance of ReCF will be depressed when γ_H is small because of the insufficient impact of the HRR. The DP and AUC scores empirically reach the highest point of 0.7177 and 0.4962 respectively at $\gamma_H = 28$. Besides, ReCF has a tendency to obtain suboptimal results when γ_H is larger, which indicates the tracker focuses more on the historical sample while ignoring the magnitude of the current sample.

2) γ_I in Inferred Response Regularization: In the second experiment, the sensitiveness of the term γ_I in the inferred response regularization is studied from 100 to 104.4 with a step size of 0.2. Here γ_H is fixed to 28. Figure 7 shows that the tracker is less robust when γ_I is small since the tracker puts less attention on the inferred response label and thus drops the prior spatial information of the object. When $\gamma_I = 102.2$, the DP and AUC scores empirically reaches the turning point of 0.7177 and 0.4962 respectively. On the overall

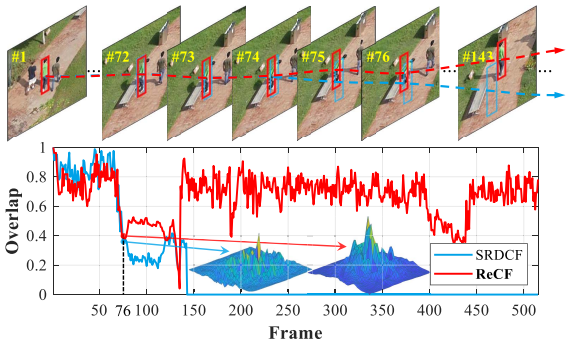


Fig. 8. Overlap of ReCF and SRDCF on the sequence *group3_3* from UAV123@10fps. ReCF can track the person successfully while SRDCF lose the target after frame #143.

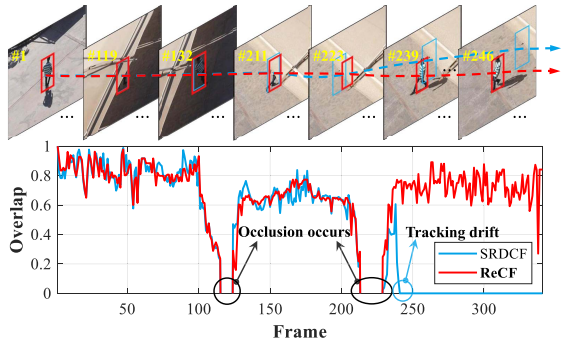


Fig. 9. Overlap of ReCF and SRDCF on the sequence *person12_2* from UAV123@10fps. ReCF can track the person successfully while SRDCF lose the target after frame #239.

trend, the AUC scores fluctuate around 0.495 when $\gamma_I \in [101, 103.2]$, which means that γ_I has a general improvement on the tracker robustness with low sensitiveness.

F. Tracking Process Analysis

1) Target Appearance Variation: Figure 8 presents the tracking precision of ReCF and SRDCF in overlap at each frame on the *group3_3* sequence from UAV123@10fps. In the sequence, when the person's posture undergoes significant change, SRDCF cannot adapt to this variation, although it does not get totally lost the target (before frame #75) as shown in Fig. 8. The detection response of SRDCF at frame #76 becomes less salient and contains more noise compared to ReCF, since SRDCF collects deviated samples and still assumes a fixed single center Gaussian label over them. Therefore, SRDCF begins to drift gradually and finally fails to track the person (frame #143). As a result, the overlap of SRDCF remains 0 after frame #143.

Remark 11: On the contrary, ReCF can give proper response labels for the previous and current samples by the proposed response reasoning strategy to estimate the motion of the person successfully.

2) Occlusion: Figure 9 presents the tracking precision of ReCF and SRDCF in overlap at each frame on the *person12_2* sequence from UAV123@10fps. Although the eaves obscure

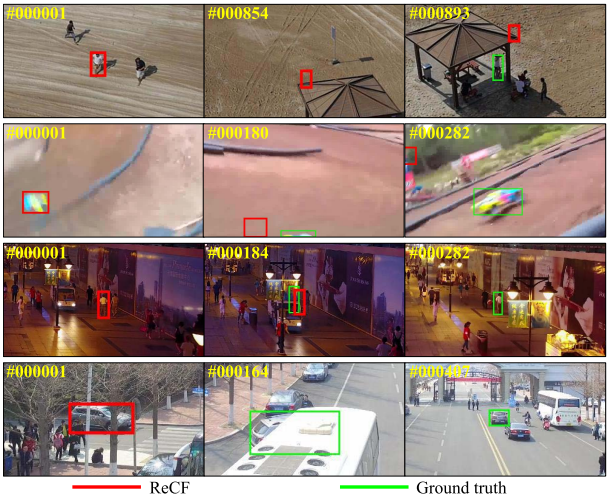


Fig. 10. Failure cases of the proposed method. The first, second, third, and fourth rows depict the tracking results on *group2_3* from UAV123@10fps, *RcCar4* from DTB70, *uav0000074_01656_s* from VisDrone2019-test-dev, and *S0801* from UAVDT.

parts of the human body from frame #100 and #128, ReCF and SRDCF can track the person when he reappears. When the person is entirely obscured by the eaves and appears again (frame #205 to #239), SRDCF arises tracking drift and regards the eave as the target. The main reason can be attributed to SRDCF using the contaminated samples in training which leads to the deviation of the model. Thanks to the proposed response reasoning modules, ReCF possesses the self-regulated update ability and does not require an additional learning rate to maintain the appearance model. Thus ReCF can successfully track the person after he is temporarily occluded by the eaves.

G. Failure Cases

Figure 10 shows some failure cases of the proposed method. In these sequences, when the target disappears in the airborne camera visual field for a long time, the continuous frame information required by the response reasoning modules will lack the necessary target appearance information. Thus the extensive irrelevant surrounding information will be incorporated to construct a chaotic historical training label and mislead the self-regulated update function in the response reasoning. Finally, the situations will cause the response reasoning to malfunction and induce the tracker to lose the target completely.

H. Onboard Tests

In addition to evaluation on a large-scale established benchmark, the proposed method is further tested on a typical CPU-based onboard platform (Intel NUC8i7HVK with a single i7-8809G CPU) for UAVs to demonstrate the real-time efficiency and robust performance. Figure 11 provides the tracking results of the proposed ReCF on three tests. The results show that the average tracking precision of ReCF in CLE are all less than 20 pixels, and the average precision in

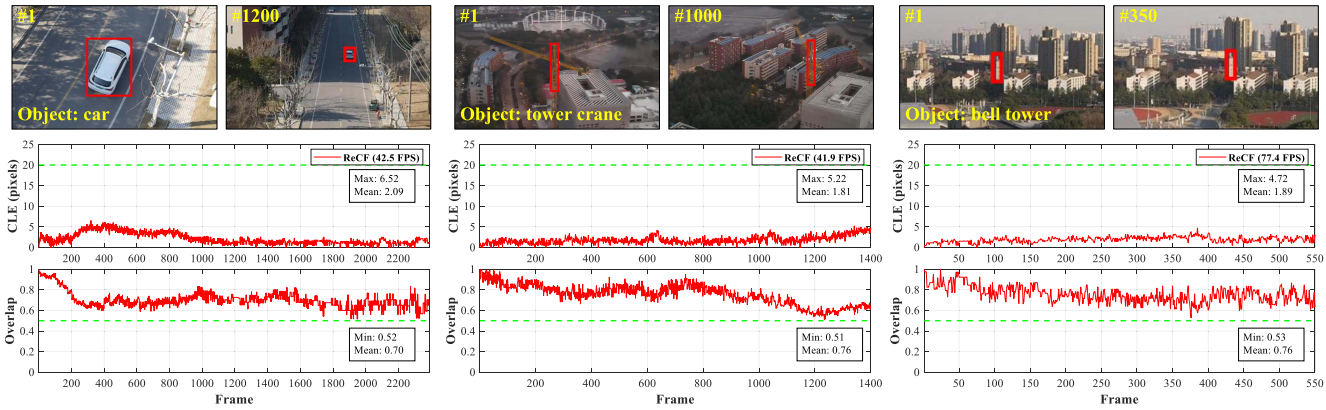


Fig. 11. Onboard tracking tests on different UAV scenarios. The red box denotes the tracked object and the green dashed line denotes the acceptable threshold. The lower the CLE value, the better the tracking precision, and the higher the overlap value, the better the scale estimation accuracy. Experimental results validate the satisfying tracking performance of ReCF.

overlap are all greater than 0.5, indicating the feasibility of ReCF for real-time UAV tracking applications.

V. Conclusion

In this work, a novel adaptive response reasoning tracking approach is presented to leverage the dynamic temporal information for robust filter learning. Beyond the conventional response consistency requirement, the proposed tracking method exploits the historical response regularization to break away from the limitations of the fixed ideal label in the CF framework. Moreover, the combination with the inferred response regularization can help the filter update adaptively, leading to more robustness against the fast motion, viewpoint changes, and other challenging factors. Extensive evaluations on four well-known UAV tracking benchmarks demonstrate that our proposed method is promising for real-time UAV applications, achieving state-of-the-art performance in accuracy, robustness, and efficiency. The proposed method will further extend the development of UAV visual applications, especially in object tracking in intelligent transportation.

References

[1] R. Hartley, B. Kamgar-Parsi, and C. Narber, "Using roads for autonomous air vehicle guidance," *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 12, pp. 3840–3849, Dec. 2018.

[2] H. Zhou, H. Kong, L. Wei, D. Creighton, and S. Nahavandi, "Efficient road detection and tracking for unmanned aerial vehicle," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 1, pp. 297–309, Feb. 2015.

[3] R. Ke, Z. Li, S. Kim, J. Ash, Z. Cui, and Y. Wang, "Real-time bidirectional traffic flow parameter estimation from aerial videos," *IEEE Trans. Intell. Transp. Syst.*, vol. 18, no. 4, pp. 890–901, Apr. 2017.

[4] J. Zhu *et al.*, "Urban traffic density estimation based on ultrahigh-resolution UAV video and deep neural network," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 12, pp. 4968–4981, Dec. 2018.

[5] M. Mueller, G. Sharma, N. Smith, and B. Ghanem, "Persistent aerial tracking system for UAVs," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2016, pp. 1562–1569.

[6] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista, "High-speed tracking with kernelized correlation filters," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 3, pp. 583–596, Mar. 2015.

[7] M. Danelljan, G. Bhat, F. S. Khan, and M. Felsberg, "ECO: Efficient convolution operators for tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6931–6939.

[8] H. K. Galoogahi, A. Fagg, and S. Lucey, "Learning background-aware correlation filters for visual tracking," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1144–1152.

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

[10] C. Wang, L. Zhang, L. Xie, and J. Yuan, "Kernel cross-correlator," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, 2018, vol. 32, no. 1, pp. 4179–4186.

[11] Y. Wu, J. Lim, and M.-H. Yang, "Object tracking benchmark," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1834–1848, Sep. 2015.

[12] M. Kristan *et al.*, "A novel performance evaluation methodology for single-target trackers," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 11, pp. 2137–2155, Nov. 2016.

[13] L. Huang, X. Zhao, and K. Huang, "GOT-10k: A large high-diversity benchmark for generic object tracking in the wild," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 5, pp. 1562–1577, May 2021.

[14] M. Mueller, N. Smith, and B. Ghanem, "A benchmark and simulator for UAV tracking," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 445–461.

[15] S. Li and D.-Y. Yeung, "Visual object tracking for unmanned aerial vehicles: A benchmark and new motion models," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, 2017, vol. 31, no. 1, pp. 4140–4146.

[16] P. Zhu, L. Wen, D. Du, X. Bian, Q. Hu, and H. Ling, "Vision meets drones: Past, present and future," 2020, *arXiv:2001.06303*. [Online]. Available: http://arxiv.org/abs/2001.06303

[17] D. Du *et al.*, "The unmanned aerial vehicle benchmark: Object detection and tracking," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 375–391.

[18] 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 *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Nov. 2019, pp. 4415–4422.

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

[20] Y. Li, C. Fu, F. Ding, Z. Huang, and J. Pan, "Augmented memory for correlation filters in real-time UAV tracking," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2020, pp. 1559–1566.

[21] S. Boyd, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Found. Trends Mach. Learn.*, vol. 3, no. 1, pp. 1–122, 2010.

[22] D. Bolme, J. R. Beveridge, B. A. Draper, and Y. M. Lui, "Visual object tracking using adaptive correlation filters," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 2544–2550.

[23] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista, "Exploiting the circulant structure of tracking-by-detection with kernels," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2012, pp. 702–715.

[24] M. Danelljan, F. S. Khan, M. Felsberg, and J. Van De Weijer, "Adaptive color attributes for real-time visual tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1090–1097.

[25] L. Bertinetto, J. Valmadre, S. Golodetz, O. Miksik, and P. H. S. Torr, "Staple: Complementary learners for real-time tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1401–1409.

[26] N. Wang, W. Zhou, Q. Tian, R. Hong, M. Wang, and H. Li, "Multi-cue correlation filters for robust visual tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4844–4853.

[27] C. Fu, F. Lin, Y. Li, and G. Chen, "Correlation filter-based visual tracking for UAV with online multi-feature learning," *Remote Sens.*, vol. 11, no. 5, pp. 549–571, 2019.

[28] H. K. Galoogahi, T. Sim, and S. Lucey, "Correlation filters with limited boundaries," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 4630–4638.

[29] M. Danelljan, G. Häger, F. S. Khan, and M. Felsberg, "Learning spatially regularized correlation filters for visual tracking," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 4310–4318.

[30] H. Zhu, H. Peng, G. Xu, L. Deng, Y. Cheng, and A. Song, "Bilateral weighted regression ranking model with spatial-temporal correlation filter for visual tracking," *IEEE Trans. Multimedia*, early access, Apr. 28, 2021, doi: 10.1109/TMM.2021.3075876.

[31] Z. Liu, Z. Lian, and Y. Li, "A novel adaptive kernel correlation filter tracker with multiple feature integration," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2017, pp. 254–265.

[32] M. Danelljan, G. Häger, F. S. Khan, and M. Felsberg, "Accurate scale estimation for robust visual tracking," in *Proc. Brit. Mach. Vis. Conf.*, 2014, pp. 1–11.

[33] M. Danelljan, G. Häger, F. S. Khan, and M. Felsberg, "Discriminative scale space tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 8, pp. 1561–1575, Aug. 2017.

[34] F. Li, Y. Yao, P. Li, D. Zhang, W. Zuo, and M.-H. Yang, "Integrating boundary and center correlation filters for visual tracking with aspect ratio variation," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2017, pp. 2001–2009.

[35] A. Lukezic, T. Vojír, L. C. Zajc, J. Matas, and M. Kristan, "Discriminative correlation filter with channel and spatial reliability," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4847–4856.

[36] C. Ma, J.-B. Huang, X. Yang, and M.-H. Yang, "Hierarchical convolutional features for visual tracking," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3074–3082.

[37] M. Danelljan, G. Häger, F. S. Khan, and M. Felsberg, "Convolutional features for correlation filter based visual tracking," in *Proc. IEEE Int. Conf. Comput. Vis. Workshop (ICCVW)*, Dec. 2015, pp. 621–629.

[38] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2005, pp. 886–893.

[39] J. van de Weijer, C. Schmid, J. Verbeek, and D. Larlus, "Learning color names for real-world applications," *IEEE Trans. Image Process.*, vol. 18, no. 7, pp. 1512–1523, Jul. 2009.

[40] T. Xu, Z.-H. Feng, X.-J. Wu, and J. Kittler, "Joint group feature selection and discriminative filter learning for robust visual object tracking," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 7949–7959.

[41] T. Xu, Z.-H. Feng, X.-J. Wu, and J. Kittler, "Learning adaptive discriminative correlation filters via temporal consistency preserving spatial feature selection for robust visual object tracking," *IEEE Trans. Image Process.*, vol. 28, no. 11, pp. 5596–5609, Nov. 2019.

[42] M. Danelljan, A. Robinson, F. S. Khan, and M. Felsberg, "Beyond correlation filters: Learning continuous convolution operators for visual tracking," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 472–488.

[43] M. Danelljan, G. Häger, F. S. Khan, and M. Felsberg, "Adaptive decontamination of the training set: A unified formulation for discriminative visual tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1430–1438.

[44] G. Bhat, J. Johnander, M. Danelljan, F. S. Khan, and M. Felsberg, "Unveiling the power of deep tracking," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 493–509.

[45] A. Bibi, M. Mueller, and B. Ghanem, "Target response adaptation for correlation filter tracking," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 419–433.

[46] C. Ma, J.-B. Huang, X. Yang, and M.-H. Yang, "Adaptive correlation filters with long-term and short-term memory for object tracking," *Int. J. Comput. Vis.*, vol. 126, no. 8, pp. 771–796, Aug. 2018.

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

[48] L. Wang, W. Ouyang, X. Wang, and H. Lu, "Visual tracking with fully convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3119–3127.

[49] Y. Qi *et al.*, "Hedged deep tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4303–4311.

[50] J. Valmadre, L. Bertinetto, J. Henriques, A. Vedaldi, and P. H. S. Torr, "End-to-end representation learning for correlation filter based tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5000–5008.

[51] S. Yun, J. Choi, Y. Yoo, K. Yun, and J. Y. Choi, "Action-decision networks for visual tracking with deep reinforcement learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1349–1358.

[52] T. Zhang, C. Xu, and M.-H. Yang, "Multi-task correlation particle filter for robust object tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4819–4827.

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

[54] K. Dai, D. Wang, H. Lu, C. Sun, and J. Li, "Visual tracking via adaptive spatially-regularized correlation filters," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4665–4674.

[55] X. Li, C. Ma, B. Wu, Z. He, and M.-H. Yang, "Target-aware deep tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1369–1378.

[56] N. Wang, Y. Song, C. Ma, W. Zhou, W. Liu, and H. Li, "Unsupervised deep tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1308–1317.

[57] L. Bertinetto, J. Valmadre, J. F. Henriques, A. Vedaldi, and P. H. S. Torr, "Fully-convolutional Siamese networks for object tracking," in *Proc. Eur. Conf. Comput. Vis. Workshops (ECCVW)*, 2016, pp. 850–865.

[58] Q. Guo, W. Feng, C. Zhou, R. Huang, L. Wan, and S. Wang, "Learning dynamic Siamese network for visual object tracking," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1781–1789.

[59] Z. He, Y. Fan, J. Zhuang, Y. Dong, and H. Bai, "Correlation filters with weighted convolution responses," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2017, pp. 1992–2000.

[60] J. Zhao, K. Dai, D. Wang, H. Lu, and X. Yang, "Online filtering training samples for robust visual tracking," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 1488–1496.

[61] N. Wang, W. Zhou, Y. Song, C. Ma, and H. Li, "Real-time correlation tracking via joint model compression and transfer," *IEEE Trans. Image Process.*, vol. 29, pp. 6123–6135, 2020.

[62] N. Wang, W. Zhou, Y. Song, C. Ma, W. Liu, and H. Li, "Unsupervised deep representation learning for real-time tracking," *Int. J. Comput. Vis.*, vol. 129, no. 2, pp. 400–418, Feb. 2021.

Fuling Lin (Graduate Student Member, IEEE) received the B.Eng. degree in mechanical engineering from Tongji University, Shanghai, China, where he is currently pursuing the M.Sc. degree in mechanical engineering. His research interests include robotics, visual object tracking, and computer vision.

Changhong Fu (Member, IEEE) received the Ph.D. degree in robotics and automation from the Computer Vision and Aerial Robotics (CVAR) Laboratory, Technical University of Madrid, Madrid, Spain, in 2015. During his Ph.D. degree, he held two research positions at Arizona State University, Tempe, AZ, USA, and Nanyang Technological University (NTU), Singapore. After receiving his Ph.D. degree, he was as a Post-Doctoral Research Fellow at NTU. He is currently an Associate Professor with the School of Mechanical Engineering, Tongji University, Shanghai, China. He is leading seven projects related to the vision for unmanned systems (US). He has worked on more than ten projects related to the vision for UAV. In addition, he has published more than 70 journal articles and conference papers, including the *IEEE GRS Magazine*, IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING (TGRS), IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY (TCSVT), IEEE TRANSACTIONS ON MULTIMEDIA (TMM), IEEE TRANSACTIONS ON MECHATRONICS (TMECH), IEEE TRANSACTIONS ON INDUSTRIAL ELECTRONICS (TIE), CVPR, ICCV, ICRA, and IROS, related to the intelligent vision and control for UAV. His research areas are intelligent vision and control for US in a complex environment.

Yujie He is currently pursuing the B.Eng. degree in mechanical engineering with Tongji University, Shanghai, China. His research interests include visual tracking and computer vision.

Weijiang Xiong is currently pursuing the B.Eng. degree in mechanical engineering with Tongji University, Shanghai, China. His research interests include robot perception and machine learning.

Fan Li is currently pursuing the B.Eng. degree in mechanical engineering with Tongji University, Shanghai, China. His research interests include visual tracking for unmanned aerial vehicles and computer vision.