2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 25-29, 2020, Las Vegas, NV, USA (Virtual)

Towards Robust Visual Tracking for Unmanned Aerial Vehicle with Tri-Attentional Correlation Filters

Yujie He[1], Changhong Fu[1,*], Fuling Lin[1], Yiming Li[1], and Peng Lu[2]

***Abstract*— Object tracking has been broadly applied in unmanned aerial vehicle (UAV) tasks in recent years. However, existing algorithms still face difficulties such as partial occlusion, clutter background, and other challenging visual factors. Inspired by the cutting-edge attention mechanisms, a novel object tracking framework is proposed to leverage multi-level visual attention. Three primary attention, *i.e.,* contextual attention, dimensional attention, and spatiotemporal attention, are integrated into the training and detection stages of correlation filter-based tracking pipeline. Therefore, the proposed tracker is equipped with robust discriminative power against challenging factors while maintaining high operational efficiency in UAV scenarios. Quantitative and qualitative experiments on two well-known benchmarks with 173 challenging UAV video sequences demonstrate the effectiveness of the proposed framework. The proposed tracking algorithm favorably outperforms 12 state-of-the-art methods, yielding 4.8% relative gain in UAVDT and 8.2% relative gain in UAV123@10fps against the baseline tracker while operating at the speed of ∼28 frames per second.**

I. Introduction

Visual object tracking plays an essential role in unmanned aerial vehicle (UAV) tasks, including target following [1], flying vehicle tracking [2], and autonomous landing [3]. However, most existing trackers remain vulnerable under challenging environmental conditions. In UAV tracking, the complex working conditions continuously lead to different types of object appearance variations, such as partial occlusion and background clutter, which severely degrade the overall tracking accuracy and robustness.

Under the hard constraints of a real-time vision-based UAV system, an ideal tracker should be efficient to gain more computing capability for sensor fusion, high-level control, *etc*. Compared with other discriminative tracking methods, correlation filter (CF)-based trackers assume circulant shifts of the object sample for filter training. Efficient element-wise operations in the Fourier domain can guarantee real-time performance, which is favorably suitable for UAV tracking applications. Thus, CF-based tracking methods have been extensively studied in recent years [4]–[9]. In this process, however, the circulant shifting operation and extended search region introduce synthetic samples and unwarranted noise separately [10]. Thus, the learned filters are easily corrupted and can lead to unrecoverable drift when background clutter, similar objects, or other challenging factors occurred.

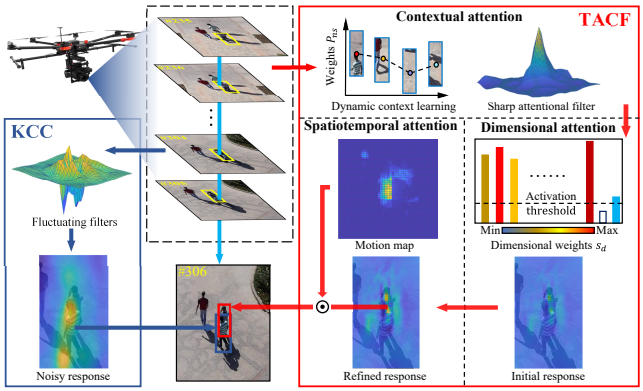

Fig. 1. Comparison between baseline KCC tracker [4] and the proposed TACF tracker. In UAV tracking, TACF incorporates the contextual information into the filter training stage, so that the attentional filters can sense the varying appearance of the object and environment simultaneously. After that, both the dimensional and spatiotemporal attention of the response maps are utilized to enhance the overall map, aiming to repress the noise and improve the tracking performance better. In contrast, the KCC tracker only uses the sample information of the current frame, which makes the filter susceptible to the appearance changes.

Some works have attempted to tackle background clutter or similar object by applying context learning. Background-aware CF (BACF) [11] is proposed to extract more negative samples from the larger background by constructing a cropping matrix. M. Mueller *et al.* [12] managed to expand the receptive fields of filters by utilizing context information. Although these methods are capable of sensing the surrounding environment, their excessive samples lack a reasonable attention mechanism. Therefore, they are not able to employ the limited computational resources to the parts that may interfere with the tracked object.

Besides, the CF-based tracking method mostly ignores the potential information about the multi-dimension response maps itself or from consecutive frames. The responses generated by correlation filters indicate the importance of some particular pixels in the different feature dimensions and locations. What is more, the temporal cues from video data can provide consistency constraints to improve tracking accuracy further. Some algorithms [6], [13] applied the color-based or iterated spatial masks to capture useful information related to tracked objects. However, these methods can hardly cope with the occurrence of cluttered backgrounds and partial occlusion because they merely exploited few levels

*Corresponding author
[1]Yujie He, Changhong Fu, Fuling Lin, and Yiming Li are with the School of Mechanical Engineering, Tongji University, 201804 Shanghai, China changhongfu@tongji.edu.cn
[2]Peng Lu is with the Adaptive Robotic Controls Lab (ArcLab), Hong Kong Polytechnic University (PolyU), Hong Kong, China peng.lu@polyu.edu.hk

of information collaboratively and almost consider only a single frame.

Inspired by the human perception system, the effectiveness of attention mechanisms become increasingly affirmed by many fields, especially in computer vision for robotics [14]–[17]. By intuitively introducing the multi-level visual attention into the training and detection stages of the CF-based framework, the tracker can be enhanced with a comprehensive attention framework. The object-oriented capability can increase the discriminative power against partial occlusion, background clutter, and other challenging factors, which often appear in UAV tracking scenes.

In this work, a novel tracking framework with tri-attention correlation filters is proposed to utilize the importance of different parts more comprehensively for more adaptive filter training and response generation. With contextual attention strategy, this work manages to exploit the dynamic responses by assigning different penalty factors to context patches, so that the filters can increase perception to the environment and discriminative capability to the object simultaneously. When generating response maps, both feature dimension-based attention and spatiotemporal attention are proposed to enhance the quality by relocating the focus on the tracked object-oriented information. As a result, a novel tri-attention correlation filter, *i.e.,* TACF tracker, is proposed to enhance the robustness against partial occlusion, clutter background, and other challenging factors in UAV tracking. The comparison between the proposed tracking method with the baseline tracker is depicted in Fig. 1. TACF shows better performance than the baseline, especially when the target is partially occluded or shares similar visual cues with surrounding objects.

The main contributions of this work are listed as follows:

- A unified tri-attention framework to leverage multi-level visual information, including contextual, spatiotemporal, and dimensional attention, is proposed to improve tracking performance in both robustness and efficiency.
- By introducing a novel criterion to assess the quality of the contextual response map, the contextual attention strategy integrates dynamically varying environmental information into the filter training stage.
- The spatiotemporal and dimensional attention modules are applied to exploit both the internal and external associations of consecutive frames. Therefore, the methods can rule out unwarranted noise along the tracking process by allocating different levels of attention to response maps dynamically.
- Extensive evaluations have conducted on two well-known UAV tracking benchmarks with 173 challenging image sequences. The results have demonstrated that the presented TACF tracker outperforms the other 12 state-of-the-art trackers while operating at the speed of $\sim$28 frames per second (FPS).

II. Related works

A. Tracking with correlation filters

Exploiting CF for object tracking started with the method called the minimum output sum of squared error, *i.e.,* MOSSE tracker [18]. The tracker is constructed and trained using gray-scale samples in the frequency domain for efficiency. Afterward, a variety of works built upon the framework improve the performance by combining multiple features [19], scale estimation [20], kernel trick [21], context learning [11], [12], filter weighing [6], [7], [10], [22], or end-to-end neural network architecture [23]–[25]. Kernel cross-correlator (KCC) [4] provides a novel solution for the CF-based framework with high expandability and brief formulation. However, the lack of context information and further refinement of generated responses limit the tracker's ability to discriminate the object in diversified environments.

Other CF-based methods focus on utilizing features extracted from the convolutional neural network (CNN) to obtain a more comprehensive object representation. Some trackers hierarchically utilize convolutional features or propose an adaptive fusion approach to improve the encoding ability of the model [26]–[29]. Although CF-based trackers have made sound progress, it is still difficult for them to achieve object tracking for UAV with high performance and efficient operation at the same time.

B. Tracking with attention mechanism

Attention, as a critical ability of the human system, enables people to focus on more useful information when processing multi-modal information. Originated from the area of neural machine translation for modeling the contextual information, attention mechanism in machine vision has improved the success of various applications in recent years. It continues to be an omnipresent component in state-of-the-art models, such as image classification [14], image captioning [30], and scene segmentation [17]. In visual tracking, CNN-based methods with attention mechanisms can integrate different visual information to improve tracking accuracy. J. Choi *et al.* [25] proposed an attention network to switch among different features to select the suitable tracking mode. Z. Zhu *et al.* [24] incorporated optical flow based on deep learning into the tracking pipeline. Other tracking methods [31]–[34] used feature maps extracted from deep neural networks to select the appropriate tracking mode for better performance. However, these methods are not suitable for the complex UAV tracking scenarios, *e.g.,* partial occlusion, clutter background, and viewpoint changes.

In this work, contextual attention is first applied in the filter training stage to improve the discrimination of cluttered backgrounds and similar objects in UAV tracking. Subsequent dimensional and spatiotemporal attention in the response generation stage can refine the final tracking result, leading to higher robustness and accuracy of tracker while preserving sufficient tracking speed.

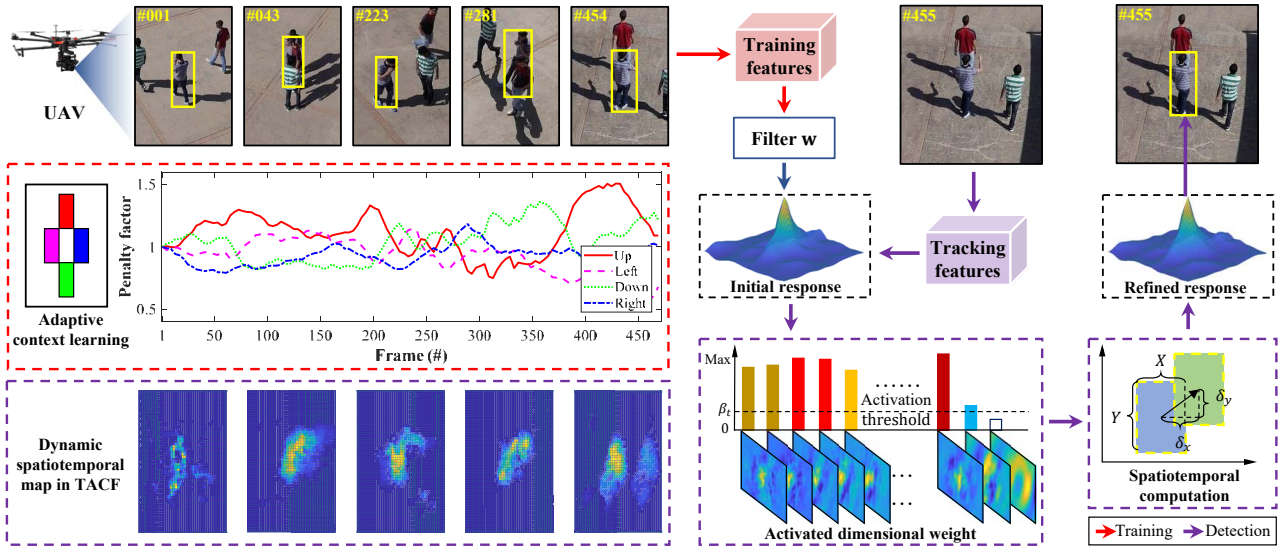

Fig. 2. The main workflow of the TACF tracker. The object and context patches near the object are first extracted and represented as training samples. Then, the contextual attention strategy is implemented to realize adaptive learning of the surrounding environment throughout the online tracking process, thereby notably improving the discriminative power of the tracker. After generating the response, dimensional attention and spatiotemporal attention are applied to refine the response maps and suppress noise. As a result, the proposed method can locate the tracked object efficiently and robustly.

III. Proposed Method

This section first reviews the baseline tracker, *i.e.,* KCC [4]. Then, the proposed TACF tracker is presented in a top-down way: the establishment and solution of the objective function are followed by the introduction of the tri-attention strategy. The main workflow of TACF can be illustrated in Fig. 2.

A. Revisiting KCC

Given the training and testing samples, they are denoted as column vectors $\mathbf{x}$, $\mathbf{z} \in \mathbb{R}^M$ in the subsequent derivation for clarity, which can be extended to the two-dimensional image. With the function $\varphi(\cdot) : \mathbb{R}^M \to \mathbb{R}^H, H \gg M$, the inner product between $\mathbf{x}$ and $\mathbf{z}_i$ can be mapped into high-dimensional space. Accordingly, the kernelized correlator between them is defined as $\kappa(\mathbf{x}, \mathbf{z}_i) = \varphi(\mathbf{x})^\mathsf{T}\varphi(\mathbf{z}_i) \in \mathbb{R}$, where the superscript $(\cdot)^\mathsf{T}$ denotes the transpose operation. Then, the sample-based vector $\mathbf{z}_i \in \mathbb{R}^M$ is computed from the test sample $\mathbf{z}$ with the transform function $\mathcal{T}(\cdot)$ as $\mathbf{z}_i \in \mathcal{T}(\mathbf{z})$. As a result, the sample-based vector set can construct the kernel vector $\mathbf{k^{xz}} = [k_1^{\mathbf{xz}}, \cdots, k_n^{\mathbf{xz}}]^\mathsf{T}$, where $\kappa(\mathbf{x}, \mathbf{z}_i)$ is denoted as $k_i^{\mathbf{xz}}$ for simplicity. Finally, the kernelized cross-correlation can encode the pattern of training samples into filters $\hat{\mathbf{w}}^*$, and the output $\hat{C}(\mathbf{x}, \mathbf{z})$ in frequency domain can be computed as:

$$\hat{C}(\mathbf{x}, \mathbf{z}) = \hat{\mathbf{k}}^{\mathbf{xz}} \odot \hat{\mathbf{w}}^* = \mathcal{F}^{-1}(\mathbf{k^{xz}} \star \mathbf{w}) \ , \tag{1}$$

where $\odot$ denotes element-wise product, and $\star$ stands for cross correlation. The superscript $(\cdot)^*$ and $\hat{\cdot}$ represent complex conjugate operation and discrete Fourier transformation, *i.e.,* $\hat{\mathbf{x}} = \mathcal{F}(\mathbf{x})$. With the ideal response $\mathbf{y}$ and learning samples $\mathbf{x}_i$, the objective function is formulated by minimizing the squared error using ridge regression:

$$\hat{\mathcal{E}} = \sum_{n=1}^{N} \left\| \hat{\mathbf{k}}_n^{\mathbf{xz}} \odot \hat{\mathbf{w}}_n^* - \hat{\mathbf{y}} \right\|_2^2 + \lambda \left\| \hat{\mathbf{w}}^* \right\|_2^2 \ , \tag{2}$$

where $\hat{\mathbf{w}}_n^*$ is the n-th channel of the learned filter.

Because the operations in Eq. (2) can be performed in element-wise, the corresponding $\mathbf{w}^*$ can be solved independently to obtain a closed-form solution. However, KCC utilizes a relatively larger search area, which will bring cluttered information into filters training so that the KCC framework cannot achieve better performance by increasing distinguishing ability against real background information.

B. Tri-attention correlation filters framework

As reviewed in Section III-A, KCC tracker cannot fully exploit the information of surrounding contexts or enhance the critical parts what filters need to pay attention. Thus, attentional contextual information is introduced as negative samples to enhance the training of correlation filters, which is defined as:

$$\sum_{s=1}^{S} \left\| p_{ns} \hat{C}_n(\mathbf{x}_{ns}, \mathbf{x}_{ns}) \right\|_2^2 \ , \tag{3}$$

where S is the number of patches extracted from the up, down, left, and right to the object. They are considered as hard negative samples, so the desired output is zero. Besides, an adaptive penalty factor p_{ns} is proposed to evaluate the importance of the context patches and varies along the tracking process (a detailed explanation is in Section III-C).

Motivated by attention mechanism, spatiotemporal and dimensional attention are integrated into final response generation stage. Accordingly, the tri-attention correlation filters

with N features taken into account can be formed by minimizing the regression target:

$$\hat{\mathcal{E}}(\hat{\mathbf{w}}^*) = \sum_{n=1}^{N} \left(\left\| \hat{C}_n(\mathbf{x}_{n0}, \mathbf{x}_{n0}) - \hat{\mathbf{y}}_n \right\|_2^2 + \lambda_1 \left\| \hat{\mathbf{w}}_n^* \right\|_2^2 + \lambda_2 \sum_{s=1}^{S} \left\| p_{ns} \hat{C}_n(\mathbf{x}_{ns}, \mathbf{x}_{ns}) \right\|_2^2 \right) , \quad (4)$$

where $\mathbf{x}_{n0}$ and $\mathbf{x}_{ns}$ are the representation of image corresponding to the object and context by the n-th feature. Then, $\hat{\mathcal{E}}$ is an error measured by the correlation output of $\mathbf{x}_{n0} \in \mathbb{R}^M$ and desired output $\mathbf{y}_n \in \mathbb{R}^M$. $\hat{\mathbf{w}}_n^* \in \mathbb{C}^M$ denotes the correlation filter for n-th feature in the Fourier domain. λ_1 and λ_2 are regularization factors for the filters and context information learning.

Due to the mutual independence of different features and dimensions, the objective function $\hat{\mathcal{E}}(\hat{\mathbf{w}}^*)$ in Eq. (4) can be reformulated sub-problems $\hat{\mathcal{E}}_n$ that can be obtained as:

$$\hat{\mathcal{E}}_n = \left\| \hat{C}_n(\mathbf{x}_{n0}, \mathbf{x}_{n0}) - \hat{\mathbf{y}}_n \right\|_2^2 + \lambda_1 \left\| \hat{\mathbf{w}}_n^* \right\|_2^2 + \sum_{s=1}^{S} \left\| P_{ns} \hat{C}_n(\mathbf{x}_{ns}, \mathbf{x}_{ns}) \right\|_2^2 , \quad (5)$$

where the regularized factor for each context patch P_{ns} can be computed as $P_{ns} = \sqrt{\lambda_2} p_{ns},\ s = 1, \cdots, S$.

By setting the first derivative of $\hat{\mathbf{w}}_n^*$ to zero, the solution to the optimization problem Eq. (5) can be calculated with element-wise operations as follows:

$$\hat{\mathbf{w}}_n^* = \frac{\hat{\mathbf{K}}^{n0} \odot \hat{\mathbf{y}}_n}{\hat{\mathbf{K}}^{n0} \odot \hat{\mathbf{K}}^{n0*} + \lambda_1 + \sum_{s=1}^{S} \left({P_{ns}}^2 \hat{\mathbf{K}}^{ns} \odot \hat{\mathbf{K}}^{ns*} \right)} , \quad (6)$$

where the fraction operator denotes element-wise division, and $\hat{\mathbf{k}}^{\mathbf{x}_{n0}\mathbf{x}_{n0}}$ and $\hat{\mathbf{k}}^{\mathbf{x}_{ns}\mathbf{x}_{ns}}$ are replaced by $\hat{\mathbf{K}}^{n0}$ and $\hat{\mathbf{K}}^{ns}$ for clarity, respectively.

For contextual patches, learning at each frame can lead to overfitting. Therefore, the context attention is set to update at a frequency of f_c, which can further improve the overall tracking efficiency. When the context patches are not taken into account, Eq. (6) can be reformulated as follows:

$$\hat{\mathbf{w}}_n^* = \frac{\hat{\mathbf{K}}^{n0} \odot \hat{\mathbf{y}}_n}{\hat{\mathbf{K}}^{n0} \odot \hat{\mathbf{K}}^{n0*} + \lambda_1} . \quad (7)$$

C. Contextual attention strategy

When encountering dramatic changes in object appearance, such as occlusion or sudden illumination changes, the constant updating of the model may introduce some noisy negative samples for training, thereby reducing the quality of the filters. Therefore, a novel response quality index, the peak median energy ratio (PME), is proposed for efficient response map evaluation, and give guidance for subsequent filters training, which can be calculated as follows:

$$PME = \frac{|R_{\max} - R_{\text{med}}|^2}{\text{mean}\left[\sum_{x=1}^{W} \sum_{y=1}^{H} \left(R_{(x,y)} - R_{\text{med}} \right)^2 \right]} , \quad (8)$$

where $R_{\max}$ and R_{med} denote the maximum and median score separately, and $R_{(x,y)}$ is pixel-wise value in the map. Accordingly, the difference between $R_{\max}$ and R_{med} can respond to the sharpness of the highest peak. The denominator $R_{(x,y)} - R_{\text{med}}$ can be applied to measure the overall smoothness. Therefore, the sharper peaks and smoother fluctuations in response maps can lead to a higher $R_{\max}$ and smaller R_{med}, so that a higher PME value is obtained to indicate high quality in the response. Accordingly, the challenging factor of surrounding patches against the object patch is defined as follows:

$$c_s = \frac{PME(R_s)}{PME(R_0)} , \quad (9)$$

where R_0 and R_s denote as the response generated from the object and context patch. As a result, the penalty factor to each context patch s can be defined as follows:

$$p_s = \frac{c_s^2}{\sum_{s=1}^{S} c_s^2} . \quad (10)$$

As illustrated in Fig. 2, the PME index exhibits a high sensitivity to the dramatic appearance change. When challenging issues addressed, the quality of the response map is restored, and the index will recover to a reasonable level.

D. Dimensional attention strategy

The various dimensions of the response generated by the different features can be considered as the capture of distinct sub-characteristic of the tracked object. By exploring the interdependence between different dimensions, the semantics of feature expression can be enhanced, and the activation strategy can be used to improve the tracking accuracy further. Given the multi-dimension response $R \in \mathbb{R}^{H \times W \times D}$, the weight of different dimensions is computed as follows:

$$z_d = F_a(R_d) = \frac{1}{HW} \sum_{i=1,j=1}^{H,W} R_d(i,j) + \max(R_d) , \quad (11)$$

where the output of function $F_a(\cdot)$ can be interpreted as an abstract representation of a particular dimension. It is common to use this information in prior feature engineering work. In this paper, global average and maximum operations are selected for fast calculation and accurate evaluation of the response for each dimension.

To fully capture inter-dimension dependencies, a consequent activation is utilized to employ a simple attention mechanism. The operation enhances the ability to learn the non-mutual-exclusive associations among all dimensions since multiple feasible channels and less important ones should be emphasized and filtered at the same time. Thus, a gating function is defined to calculate activated channel weight s_d as follows:

$$s_d = \max(z_d - t, 0) + \beta_t , \quad (12)$$

where β_t is the activation threshold for all weights. Finally, the output of the dimensional attention strategy is obtained

by resigning response of each channel R_d with s_d as follows:

$$R'_d = \sum_{d=1}^{D} s_d R_d \ . \tag{13}$$

As shown in Fig. 2, dimensions with higher reliability are given higher weights, while the ones with lower reliability cannot be activated. Finally, the peak and noise in the refined response can be enhanced and suppressed respectively, resulting in more accurate location.

E. Spatiotemporal attention strategy

Apart from the dimensional attention strategy, an element-wise multiplication with a predefined Hanning window and the pixel-wise sum of response map along the dimension axis is operated before normalization to the range $[0, 1]$ and mean-subtraction. For each pixel $s_{(i,j)}$ more than 0, the value will be activated with an exponential function, which indicates higher importance. Thus, the static spatial attention map S is defined as follows:

$$S = \exp\left[\text{norm}\left(\sum_{d=1}^{D} R_d \odot \mathbf{w}\right)\right] \ . \tag{14}$$

Besides, the spatial information from the object motion is also taken into account. Based on the current target size (X, Y) and the object position changes (δ_x, δ_y) caused by object motion or UAV viewpoint change in previous frame, the motion factor is calculated as follows:

$$\gamma_t = \gamma\sqrt{\frac{\delta_x^2 + \delta_y^2}{X^2 + Y^2}} \ . \tag{15}$$

With the shifting operation $\Delta_{x,y}$ derived from object motion, the dynamic attention map can be computed as:

$$S_d = S + \gamma_t S \Delta_{x,y} \ . \tag{16}$$

Finally, a matrix multiplication to obtain the output is performed on the original response map as $R'' = S_d \odot R$.

IV. Experiments

In this section, the proposed TACF tracker is thoroughly evaluated on two well-known UAV tracking benchmarks with 173 challenging image sequences, *i.e.,* UAV123@10fps [35] and UAVDT [36].

A. Experimental setups

1) ***Implementation details****:* Our TACF is implemented with Matlab 2018a, and all the experiments are evaluated on a PC equipped with Intel i7-8700K CPU (3.7GHz) and NVIDIA GeForce RTX 2080 GPU. The TACF tracker employs two hand-crafted features, *i.e.,* histograms of gradients [37] and color names [19], to represent object and context patches. The regularization parameter λ_1 and λ_2 is set to 5×10^{-5} and 0.0625, respectively. The interval for context learning f_c and the number of context patches S are set as 2 and 4. Details of the TACF tracker can be seen in Algorithm 1. All parameters are fixed for all the experiments.

Algorithm 1: TACF tracker

Input: Frames of the video sequence: $I_1, \cdots, I_K$.
The interval for context learning: f_c.
The number of context patches: S.
Initialize the TACF in the first frame.
Output: Predicted location in frame k.

for $k = 2$ *to end* **do**
  Extract the object patch in the frame k from center location of the object in last frame
  Represent $\mathbf{x}_{n0}$ using hand-crafted features
  Enhance each channel of response maps with dimensional attention operation by Eq. (13)
  Evaluate the object motion and generate the spatiotemporal attention map before fusing each response maps by Eq. (16)
  Calculate the location transformation in frame k by searching the peak on the response maps
  if $\text{mod}\ (k, f_c) == 0$ **then**
    Extract S context patches around the object
    foreach context patch $\mathbf{x}_{ns}$ **do**
      Represent extracted patches using hand-crafted features and calculate penalty factor p_s by Eq. (10)
    Learn new object appearance and update the model $\hat{\mathbf{w}}_k^{(\text{model})}$ by Eq. (6)
  else
    Learn new object appearance and update the model $\hat{\mathbf{w}}_k^{(\text{model})}$ by Eq. (7)
  end
end

Besides, Fig. 6 shows some qualitative tracking results of TACF with the other 12 trackers in challenging UAV video sequences. Related source code and UAV tracking video are available in https://github.com/vision4robotics/TACF-Tracker and https://youtu.be/4IWKLmRoS38.

2) ***Benchmarks and evaluation methodology****:* To validate the effectiveness of the proposed method, extensive experiments on UAV123@10fps [35] and UAVDT [36] benchmarks are conducted to evaluate overall performance.

As the first comprehensive UAV tracking benchmark, UAV123@10fps contains 123 video sequences with more than 37K frames and 12 kinds of challenging visual attributes, making it the most significant object tracking benchmark from an aerial viewpoint. UAVDT benchmark focuses on complex scenarios with 50 representative video sequences, which are fully annotated with bounding boxes with up to 9 kinds of challenging visual attributes.

Success rate (SR) is employed to evaluate the performance of the proposed tracker. SR based on the one-pass evaluation protocol can illustrate the percentage of frames when the overlap ratio between the estimated and the ground truth bounding boxes is higher than a certain threshold. Following [38], the area under curve (AUC) is adopted to rank the success rate of each tracker.

B. Qualitative experiments

1) ***State-of-the-art Comparison****:* In real-world tasks, efficient operations of the tracking algorithm is particu-

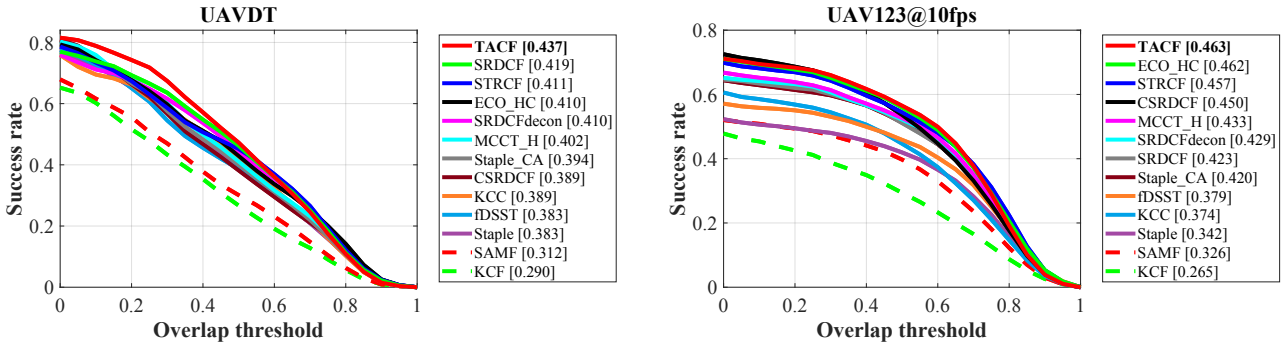

Fig. 3. Success plots of the proposed TACF tracker and other 12 state-of-the-art trackers on UAVDT and UAV123@10fps benchmarks. The experimental results demonstrate that our method yields superior performance on both widely-adopted UAV object tracking benchmarks.

TABLE I

THE AVERAGE FPS OF TACF VERSUS OTHER STATE-OF-THE-ART TRACKERS ON TWO WELL-KNOWN UAV OBJECT TRACKING BENCHMARKS.

	KCF	SAMF	Staple	KCC	fDSST	Staple_CA	SRDCF	SRDCF_decon	MCCT_H	CSRDCF	STRCF	ECO_HC	**TACF**
Avg. FPS	651.1	12.8	65.4	48.9	168.1	58.9	14	7.5	59.7	12.1	28.5	69.3	**28.1**

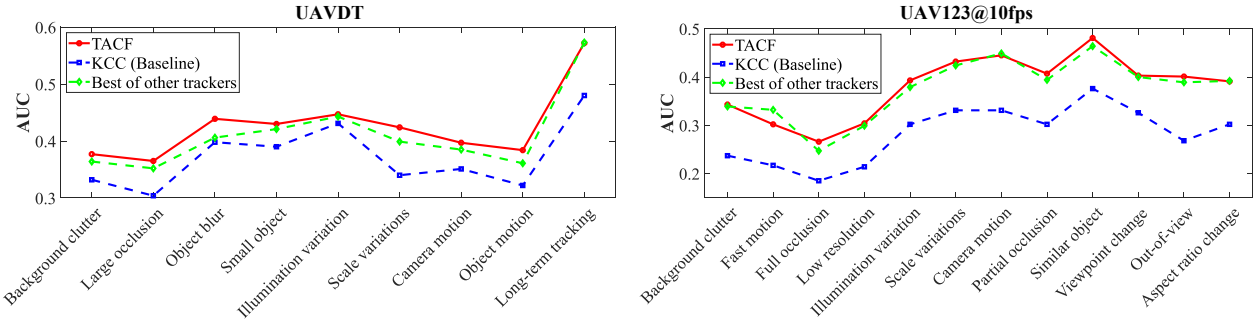

Fig. 4. Attribute-based evaluation between TACF and other state-of-the-art trackers on UAVDT and UAV123@10fps datasets. The AUC scores of different trackers in the success plots are applied to rank trackers. In most cases, the proposed TACF tracker performs favorably against other trackers, achieving significant improvements compared to the baseline tracker.

larly important due to the limited computing resources on UAV platform. Thus, the TACF tracker is compared with other 12 state-of-the-art trackers with hand-crafted features, including MCCT_H [31], Staple_CA [12], SRDCF [10], BACF [11], KCC [4], CSRDCF [13], SAMF [39], Staple [40], KCF [21], SRDCFdecon [41], STRCF [5], fDSST [20], and ECO_HC [28]. The open-source codes of trackers with default parameters provided by the authors are used in the following evaluations.

As shown in Fig. 3, the proposed TACF tracker achieves better performance compared with other trackers in success plots. On the UAVDT dataset, TACF achieves the best score with 0.437, exceeding the second (SRDCF, 0.419) and third-best tracker (STRCF, 0.411) by 4.30% and 6.32%, respectively. On the UAV123@10fps dataset, TACF keeps the best score, outperforming the second (ECO_HC, 0.462) and third-best (STRCF, 0.457) trackers.

In addition to excellent tracking performance, the speed of the proposed TACF tracker (28.1 FPS) is sufficient for UAV real-time tracking, as shown in Table I. Despite that KCF obtains the best tracking speed (651.1 FPS), followed by fDSST (168.1 FPS) and DSST (106.5 FPS), their tracking performance much lower than TACF.

*2) **Attribute-based performance analysis**:* The performance of the TACF tracker and other trackers are also analyzed in different attributes. Fig. 4 shows the scores of different trackers on different challenging attributes and demonstrates TACF exhibits better performance than most of the other trackers except for fast motion. Especially when partial occlusion or background clutter occurs, the proposed TACF tracker has a significant improvement over its baseline, and have achieved state-of-the-art performance in these aspects on these two benchmarks. Usually, in a cluttered background, most CF-based methods tend to learn appearance models from both objects and irrelative noise. By applying the tri-attention strategy, CF can focus on crucial aspects so that the tracker can show better performance against these complex visual factors.

In the future, it is possible to employ a mobile CNN to extract convolutional features for object representation to replace hand-crafted features with limited encoding ability. Features from different layers of CNN can provide semantic information and ensure efficient operation so that the tracker can improve the performance against object fast motion.

*3) **Key parameter analysis**:* To verify the effectiveness of the activation threshold of dimensional attention modules on

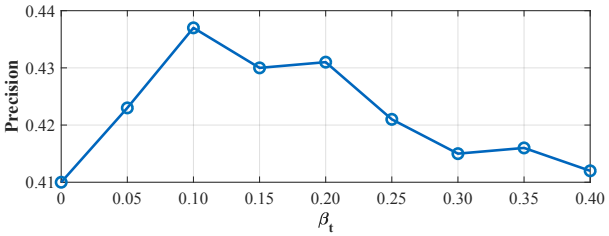

Fig. 5. Effect of activation threshold β_t on the UAVDT benchmark. When $\beta_t > 0$, it contributes to the overall performance of TACF tracker.

the tracking performance, different β_t is further analyzed on the UAVDT dataset. Starting from 0, β_t increases in small increments of 0.05 until 0.4.

The activation threshold β_t determines the degree of focus heightened by the dimensional attention strategy: The higher the β_t is set, the more attention is paid to the relatively more important dimensions. TACF with $\beta_t = 0$ means that the filters trained with the consideration of all noisy dimensions. As shown in Fig. 5, TACF effectively improves the success rate when β_t is set over 0. The success rate reaches the peak (0.718) at $\beta_t = 0.1$. Thus, $\beta_t = 0.1$ is chosen for the best performance on the challenging UAV video sequences.

*4) **Ablation study**:* Here in-depth analysis related to three different modules in the TACF framework, including contextual attention module (CA), dimensional attention (DA), and spatiotemporal attention module (PA), as well as the baseline tracker is performed to verify the effectiveness. Apart from the success rate, milliseconds per frame (MSPF) is applied to evaluate the average operational time cost.

The baseline tracker, KCC, is considered as a particular case of TACF without the tri-attention strategy. The ✓ in each column denotes that the corresponding module is activated while the ✗ means that the module is deactivated. Table II presents that the TACF tracker has significantly superior performance to KCC for 21.1% and 8.4% on UAV123@10fps and UAVDT dataset, respectively. Besides, three different attention strategy integrated into the original tracker has shown satisfying improvements from the baseline.

In terms of the average time cost, each stage of TACF is operated within an acceptable time, *i.e.,* the average running time of CA, SA, and DA is about 11.48, 3.78, and 0.51 ms, respectively. As a result, the TACF tracker operates at an average speed of 28.1 FPS on a single CPU.

Although the superior accuracy and robustness of TACF

TABLE II

PERFORMANCE COMPARISONS BETWEEN TACF WITH DIFFERENT MODULES ON UAVDT AND UAV@10FPS BENCHMARKS.

Tracker	Module			Success rate		MSPF
	CA	DA	SA	UAVDT	UAV@10fps	
KCC	✗	✗	✗	0.389	0.374	20.47
TACF+SA	✗	✗	✓	0.423	0.398	20.98
TACF+DA	✗	✓	✗	0.425	0.407	24.25
TACF+CA	✓	✗	✗	0.432	0.421	32.25
TACF	✓	✓	✓	0.437	0.456	35.59

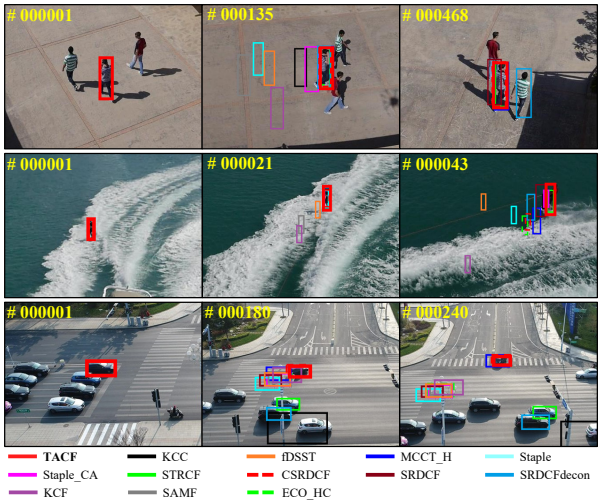

Fig. 6. Examples of UAV tracking results. The first, second, and third row show the *group1_3*, *wakeboard7* from UAV123@10fps benchmark, and *S1606* video sequence from UAVDT benchmark.

come at the cost of extra computational time, it still can meet the real-time performance requirement of UAV object tracking applications. On the one hand, due to the introduction of contextual attention, the processing speed for large object targets is reduced. On the other hand, the current TACF tracker is implemented in MATLAB without additional engineering optimization. Thus, appropriate parallel computing methods operated in the onboard processors can accelerate the operation speed when the proposed method is applied to real-world tracking tasks.

V. CONCLUSIONS

In this work, a novel tracking framework with tri-attention correlation filters for robust UAV object tracking is proposed to achieve high performance in tracking tasks by leveraging multi attention mechanisms. Three types of attention, *i.e.,* contextual, spatiotemporal, and dimensional attention, have been effectively fused into the training and detection stages. Compared with the baseline tracker, the proposed TACF tracker has dramatically improved performance under challenging factors such as partial occlusion and background clutter. Moreover, qualitative and quantitative experiments on two established UAV tracking benchmarks demonstrate that the presented TACF tracker has outperformed other 12 state-of-the-art trackers in terms of accuracy, robustness, and efficiency. We believe, with our proposed tri-attention strategy, the correlation filters with multi-level attention can achieve better tracking performance further, and open the door to more extensive applications and researches for UAV.

ACKNOWLEDGMENT

The work was supported by the National Natural Science Foundation of China (No.61806148).

REFERENCES

[1] H. Cheng, L. Lin, Z. Zheng, Y. Guan, and Z. Liu, "An Autonomous Vision-based Target Tracking System for Rotorcraft Unmanned Aerial Vehicles," in *Proceedings of the IEEE International Conference on Intelligent Robots and Systems (IROS)*, 2017, pp. 1732–1738.

[2] C. Fu, A. Carrio, M. A. Olivares-Mendez, R. Suarez-Fernandez, and P. Campoy, "Robust Real-Time Vision-Based Aircraft Tracking from Unmanned Aerial Vehicles," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2014, pp. 5441–5446.

[3] 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.

[4] C. Wang, L. Zhang, L. Xie, and J. Yuan, "Kernel Cross-Correlator," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018, pp. 1–8.

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

[6] 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. 4670–4679.

[7] 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.

[8] 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 IEEE International Conference on Robotics and Automation (ICRA)*, 2020, pp. 1–7.

[9] 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.

[10] M. Danelljan, G. Hager, F. Shahbaz Khan, and M. Felsberg, "Learning Spatially Regularized Correlation Filters for Visual Tracking," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 4310–4318.

[11] H. Kiani Galoogahi, A. Fagg, and S. Lucey, "Learning Background-Aware Correlation Filters for Visual Tracking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1135–1143.

[12] M. Mueller, N. Smith, and B. Ghanem, "Context-Aware Correlation Filter Tracking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1396–1404.

[13] A. Lukezic, T. Vojir, L. Cehovin 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. 6309–6318.

[14] J. Hu, L. Shen, and G. Sun, "Squeeze-and-Excitation Networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2018, pp. 7132–7141.

[15] S. Woo, J. Park, J.-Y. Lee, and I. So Kweon, "CBAM: Convolutional Block Attention Module," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 1–17.

[16] Z. Chen, L. Liu, I. Sa, Z. Ge, and M. Chli, "Learning Context Flexible Attention Model for Long-Term Visual Place Recognition," *IEEE Robotics and Automation Letters*, vol. 3, no. 4, pp. 4015–4022, 2018.

[17] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual Attention Network for Scene Segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 3146–3154.

[18] 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.

[19] M. Danelljan, F. Shahbaz Khan, M. Felsberg, and J. Van de Weijer, "Adaptive Color Attributes for Real-Time Visual Tracking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 1090–1097.

[20] M. Danelljan, G. Häger, F. Khan, and M. Felsberg, "Accurate Scale Estimation for Robust Visual Tracking," in *Proceedings of the British Machine Vision Conference (BMVC)*, 2014, pp. 1–11.

[21] 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.

[22] C. Fu, Y. He, F. Lin, and W. Xiong, "Robust Multi-Kernelized Correlators for UAV Tracking with Adaptive Context Analysis and Dynamic Weighted Filters," *Neural Computing and Applications*, pp. 1–17, 2020.

[23] J. Valmadre, L. Bertinetto, J. Henriques, A. Vedaldi, and P. H. 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. 2805–2813.

[24] Z. Zhu, W. Wu, W. Zou, and J. Yan, "End-to-End Flow Correlation Tracking with Spatial-Temporal Attention," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 548–557.

[25] J. Choi, H. Jin Chang, S. Yun, T. Fischer, Y. Demiris, and J. Young Choi, "Attentional Correlation Filter Network for Adaptive Visual Tracking," in *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2017, pp. 4807–4816.

[26] C. Ma, J.-B. Huang, X. Yang, and M.-H. Yang, "Hierarchical Convolutional Features for Visual Tracking," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 3074–3082.

[27] 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.

[28] Danelljan, Martin and Bhat, Goutam and Shahbaz Khan, Fahad and Felsberg, Michael, "ECO: Efficient Convolution Operators for Tracking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 6638–6646.

[29] M. Danelljan, G. Bhat, S. Gladh, F. S. Khan, and M. Felsberg, "Deep Motion and Appearance Cues for Visual Tracking," *Pattern Recognition Letters*, vol. 124, pp. 74–81, 2019.

[30] L. Chen, H. Zhang, J. Xiao, L. Nie, J. Shao, W. Liu, and T.-S. Chua, "SCA-CNN: Spatial and Channel-Wise Attention in Convolutional Networks for Image Captioning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 6298–6306.

[31] 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.

[32] S. Pu, Y. Song, C. Ma, H. Zhang, and M.-H. Yang, "Deep Attentive Tracking via Reciprocative Learning," in *Advances in Neural Information Processing Systems*, 2018, pp. 1931–1941.

[33] B. Chen, P. Li, C. Sun, D. Wang, G. Yang, and H. Lu, "Multi Attention Module for Visual Tracking," *Pattern Recognition*, vol. 87, pp. 80–93, 2019.

[34] J. Zhu, H. Yang, N. Liu, M. Kim, W. Zhang, and M.-H. Yang, "Online Multi-Object Tracking with Dual Matching Attention Networks," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 366–382.

[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] D. Du, Y. Qi, H. Yu, Y. Yang, K. Duan, G. Li, W. Zhang, Q. Huang, and Q. Tian, "The Unmanned Aerial Vehicle Benchmark: Object Detection and Tracking," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 370–386.

[37] N. Dalal and B. Triggs, "Histograms of Oriented Gradients for Human Detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 1, 2005, pp. 886–893.

[38] Y. Wu, J. Lim, and M.-H. Yang, "Object Tracking Benchmark," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 9, pp. 1834–1848, 2015.

[39] Y. Li and J. Zhu, "A Scale Adaptive Kernel Correlation Filter Tracker with Feature Integration," in *Proceedings of the European Conference on Computer Vision Workshops*, 2014, pp. 254–265.

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

[41] M. Danelljan, G. Hager, F. Shahbaz Khan, and M. Felsberg, "Adaptive Decontamination of the Training Set: A Unified Formulation for Discriminative Visual Tracking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 1430–1438.