Tri-Attention Correlation Filter for Effective UAV Object Tracking

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

***Abstract*— Object tracking has become a demanding task for booming unmanned aerial vehicle (UAV) applications in recent years. Although correlation filters (CF) based tracking methods show competitive performance in classic tracking benchmarks, those trackers cannot adequately cope with the dramatic appearance changes caused by the rapid flight of UAV in mid-air. What is more, with limited interpretation to training samples, traditional methods might be easily over-fitted, and lack of local attention ability which undermine the performance as well. Unlike previous works using features fusion or dynamic masking, a practical and straightforward tri-attention strategy is proposed to refine responses and filters learning adaptively. Specifically, three attention strategies are proposed on top of CF-based methods, which model the semantic interdependencies in position, dimension, and surrounding context, respectively. Besides, they can be integrated into correlation filter frameworks for more precise tracking because of lightweight and generality. Through extensive experiments on 173 challenging UAV image sequences, the proposed tracker demonstrates competitive tracking accuracy and robustness, achieving better performance than other 12 state-of-the-art trackers under challenging situations.**

I. Introduction

Last decade has witnessed the increasing role of object tracking in unmanned aerial vehicle (UAV) applications. It has been applied in a variety of challenging tasks with complex conditions, including target following [1], flying vehicle tracking [2], autonomous landing [3]. Meanwhile, generic object tracking has achieved astounding progress in recent years, however, the changing working conditions have led to many issues unique to UAV scenarios, such as camera motion, object aspect ratio change, and viewpoint change, which remain challenging to resolve.

In recent years, correlation filter (CF) based tracking approaches have been extensively studied to cope with the continuous appearance variation problems because of its computational efficiency. Compared to other discriminative methods, the CF-based method assumes a circulant shift of the object sample, and therefore it can achieve an efficient operation in Foriour domain to obtain real-time performance and suitability for UAV applications. Nevertheless, due to synthetic information and random noise introduced by circulant shifting operation and extended search region separately, the learned filters are easily weakened and thus difficult to distinguish the dramatic appearance changes in the UAV working environments.

[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

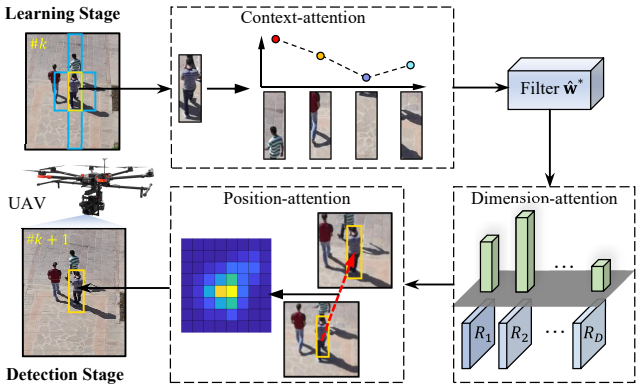

Fig. 1. The overview of TACF tracker, which has applied three sequential attention strategies: context, dimension, and position. The intermediate filters and maps are adaptively refined through the proposed method to achieve high tracking performance efficiently along the tracking process.

Some approaches have been proposed to mitigate the adverse effects. The methods [4], [5] manage to alleviate the problems by utilizing background and context information to expand the receptive fields of filters, while the excessive underutilized training samples are more likely to result in over-fitting and the decrease in accuracy and robustness. Besides, other methods [6]–[9] make use of fixed spatial or temporal regularized operations focusing on the object to improve the accuracy of tracking, but they usually cannot strike a delicate balance between speed and performance.

In object tracking for UAV, the essential information from the multi-dimension response maps and surroundings has received little attention and not been fully exploited. Instead, the context around the object and response generated by correlation filters, to a large extent, indicate the importance of some particular pixels in the current search area, the relevance of the specific feature dimension to the tracked object, and even the different interference caused by the varying environment. Some algorithms [8], [10] apply the color-based or iterated optimization algorithm to achieve improvements. However, these methods are difficult to deal with the occurrence of cluttered backgrounds and similar objects because of the relatively limited considerations in different levels of information.

Inspired by the human perception system, the effectiveness of attention mechanisms become increasingly affirmed by many fields, especially in computer vision for robotics [11]–[14]. Thus, the tri-attention strategy is introduced in this work to fully utilize the importance of different parts under the existing information, dynamically adjust the filters training and the response generation.

With context-attention strategy, this work manages to

exploit the dynamic responses by assigning different penalty factors to surrounding 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-attention and position-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 for UAV tracking, i.e., TACF tracker, is proposed to achieve better tracking performance. The workflow of the proposed tracking method is depicted in Fig. 1.

Contributions of this work can be listed as follows:

- A novel triple attention strategy is proposed to suit for CF frameworks for UAV tracking. It can not only enhance the capability of learned filters to exploit both response dimension- and pixel-wise importance and context inference level, but also help eliminate the adverse effect of excessive information and unwarranted noise. These three attention methods can help selectively emphasized the critical information and dynamically adapt to the varying appearance changes.
- The proposed method is extensively validated on specific UAV tracking datasets. Thorough evaluations and ablation studies have demonstrated that TACF tracker performs favorably against other state-of-the-art trackers and tri-attention strategy is simple yet effective so that it can be integrated into boost other CF-based methods.

II. Related works

A. Tracking with correlation filters

CF has been widely applied in tracking approaches, and it is first introduced in [15], where the minimum output sum of squared error, i.e., MOSSE tracker, and the filters are trained using gray-scale samples to identify the tracking object. Afterwards, several works [4]–[6], [9], [10], [16]–[20] built upon the framework improve the performance by combining multiple features, scale estimation, or kernel trick.

Other CF-based methods focus on utilizing features extracted from the convolutional neural network (CNN) to obtain a more comprehensive object representation [21]–[25]. Some trackers utilize deep features in a hierarchical way [21] or propose an adaptive fusion approach [22], [24] to improve the encoding ability of the model. In addition, techniques for integrating CF framework with CNN architectures have been investigated for visual tracking. The proposed methods in [26]–[28] incorporate correlation filters into a deep network for end-to-end tracking.

Though CF-based trackers have achieved high performance in multiple tracking datasets, it is still difficult to solve the challenging UAV tracking problems.

B. Tracking with attention mechanism

Attention is a critical ability of the human visual system to process multimodal information, enabling the capture of valuable information in the sensing procedure. By relocating limited computing resources to the key part of the image, attention mechanism has proven its versatility and effectiveness in machine vision, including image classification [11], [12], place recognition [13], and scene segmentation [14].

In terms of visual tracking, the context-aware and background-aware CF [4], [5] enhance filters training by using negative samples of the surrounding environment, but these methods may introduce too much noise and behave non-adaptively in the changing environment. Some other tracking methods build adaptive penalty matrix by employing the similarity of color histograms [8] or ADMM iteration [10] to achieve attention to essential parts. Other CNN-based tracking methods [27]–[29] use feature/response maps or combined optical flow methods to select the appropriate tracking mode. Despite that, the high computation complexity of convolutional operations lacks feasibility on UAV onboard processors.

In this work, the proposed tri-attention strategy fully exploit position, dimension, and context-wise information based on an efficient architecture, and its effectiveness is fully verified in mainstream UAV tracking datasets [30]–[32].

III. Proposed method

This section first reviews the baseline tracker, kernel cross-correlator (KCC) [33] which is a generalized CF framework with high expandability and brief formation. 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.

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 a feature mapping function $\varphi(\cdot) : \mathbb{R}^M \rightarrow \mathbb{R}^H, H \gg M$, the inner product between $\mathbf{x}$ and $\mathbf{z}_i$ thus can be represented in high-dimensional space, thus, 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. Besides, the sample-based vector $\mathbf{z}_i \in \mathbb{R}^M$ is generated from the test sample $\mathbf{z}$ with the transform function $\mathcal{T}(\cdot)$, which is computed as $\mathbf{z}_i \in \mathcal{T}(\mathbf{z})$. Then, the sample-based vectors 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. Hence, 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 pixel-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:

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

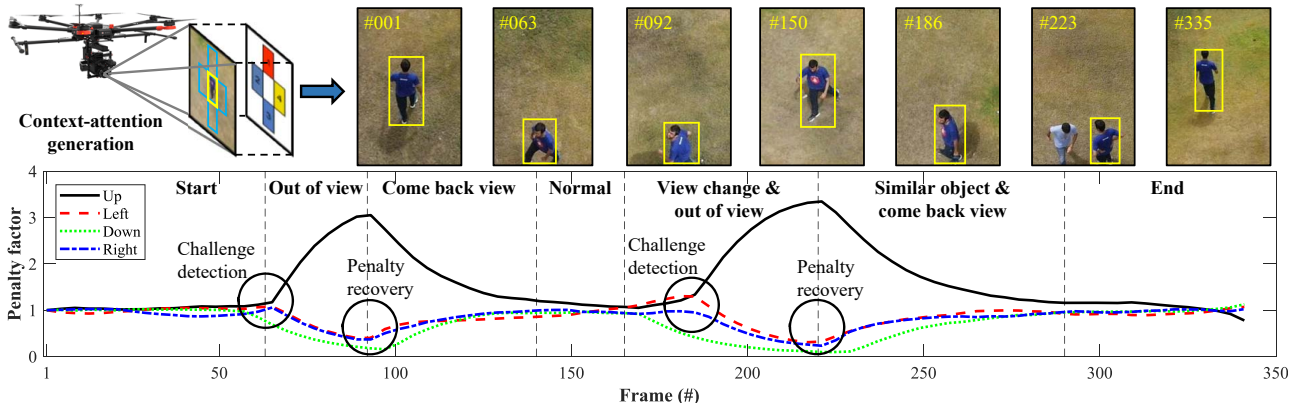

Fig. 2. The dynamic process of contextual attention mechanisms. The line graph shows the penalty factor corresponding to each part is continuously varied on *person10* sequence from UAV123@10fps. Challenging issues, such as out-of-view and similar object, are detected sensitively, so that the proposed method can achieve robust tracking and avoid drifting.

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. What is more, due to its customizable property, it can generalize the existing CFs [16], [18] well. 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).

Additionally, motivated by attention mechanism, position- and dimension-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:

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

where $\mathbf{x}_{n0}$ and $\mathbf{x}_{ns}$ are the representation of image corresponding to the object and context, respectively in 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 to control the filters and context information learning, respectively.

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

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

where the regularized factor for each context patch P_{ns} can be computed as follows:

$$P_{ns} = \sqrt{\lambda_2} p_{ns} \ (s = 1, \cdots, S) . \tag{6}$$

Then, the solution to the optimization problem Eq. (5) can be calculated by setting the first derivative of $\hat{\mathbf{w}}_n^*$ to zero.

Since the operations can be performed element-wise, a close-form solution to $\hat{\mathbf{w}}_n^*$ can be achieved by:

$$\hat{\mathbf{w}}_n^* = \frac{\hat{\mathbf{K}}^{n0} \odot \hat{\mathbf{y}}_n}{\hat{\mathbf{K}}^{n0} \odot \hat{\mathbf{K}}^{n0*} + \lambda_1 + \sum_{s=1}^{S} \left({P_{ns}}^2 \hat{\mathbf{K}}^{ns} \odot \hat{\mathbf{K}}^{ns*} \right)} . \tag{7}$$

where the fraction operator denotes element-wise division.

Besides, for the context patches, learning at each frame may lead to the over-fitting phenomenon. Therefore, learning with context-attention updates at an interval f_c. When the surrounding patches are not taken into account, context term is set to zero so that Eq. (7) can be reformulated as follows:

$$\hat{\mathbf{w}}_n^* = \frac{\hat{\mathbf{K}}^{n0} \odot \hat{\mathbf{y}}_n}{\hat{\mathbf{K}}^{n0} \odot \hat{\mathbf{K}}^{n0*} + \lambda_1} . \tag{8}$$

C. Context-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

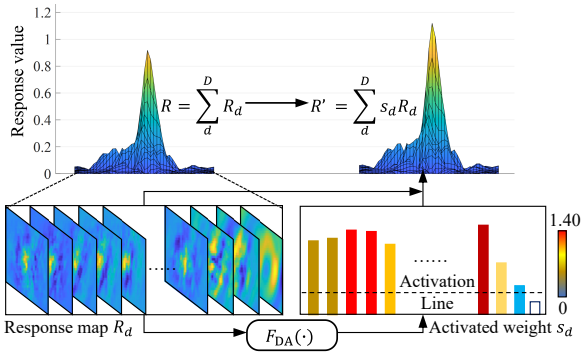

Fig. 3. The response enhancement after adaptive dimension attention.

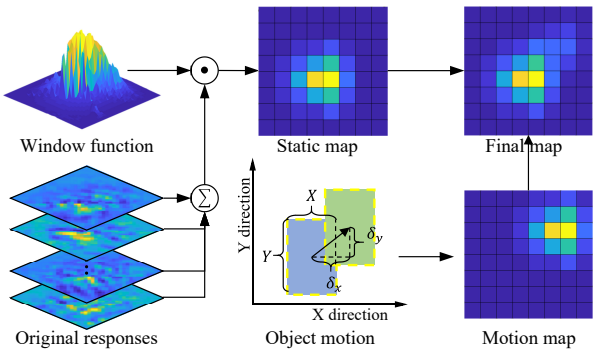

Fig. 4. Diagram of the generation process of the position attention map.

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:

$$PME = \frac{|R_{\max} - R_{\mathrm{med}}|^2}{\mathrm{mean}\left[\sum_{x=1}^{W}\sum_{y=1}^{H}\left(R_{(x,y)} - R_{\mathrm{med}}\right)^2\right]} , \tag{9}$$

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 $R_{(x,y)} - R_{\mathrm{med}}$ 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.

Thus, the challenging factor of surrounding patches against the object patch is defined as follows:

$$c_s = \frac{PME(R_s)}{PME(R_0)} , \tag{10}$$

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:

$$p_s = \frac{c_s^2}{\sum_{s=1}^{S} c_s^2} . \tag{11}$$

Figure 2 shows that the PME index exhibits a higher sensitivity to the dramatic appearance change than the conventional index, such as peak-sidelobe ratio. When challenging issues addressed, the quality of the response map is restored, and the index will recover to a reasonable level.

D. Dimension-attention strategy

Unlike the contextual attention method, 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:

$$z_d = F_a(R_d) = \frac{1}{HW}\sum_{i=1,j=1}^{H,W} R_d(i,j) + \max(R_d) , \tag{12}$$

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, an 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:

$$s_d = \sigma(z_d) = \max(\mathrm{z_d} - \mathrm{t}, 0) + \mathrm{t} , \tag{13}$$

where t is the activation threshold for all weights. Finally, The output of the dimension-attention strategy is obtained by resigning response of each channel R_d with s_d as follows:

$$R'_d = \sum_d^D s_d R_d . \tag{14}$$

Figure 3 presents that dimensions with higher reliability are given higher weights, while the ones with lower reliability cannot be activated, so the peak and noise in the refined response will be enhanced and suppressed respectively, resulting in more accurate location.

E. Position-attention strategy

Different from the dimension-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 position attention map S can is defined as follows:

$$S = \exp\left[\mathrm{norm}\left(\sum_{\mathrm{d}=1}^{\mathrm{D}} \mathrm{R_d} \odot \mathbf{w}\right)\right] . \tag{15}$$

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 pateches: 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 dimension-attention operation by Eq. (14)
  Evaluate the object motion and generate the position-attention map before fusing each response maps by Eq. (17)
  Calculate the location transformation in frame k by searching the peak on the response maps
  if $\mod(k, f_c) == 0$ **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. (11)
    Learn new object appearance and update the model $\hat{\mathbf{w}}_k^{(\text{model})}$ by Eq. (7)
  else
    Learn new object appearance and update the model $\hat{\mathbf{w}}_k^{(\text{model})}$ by Eq. (8)
  end
end

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:

$$\gamma_t = \gamma \sqrt{\frac{\delta_x^2 + \delta_y^2}{X^2 + Y^2}} \ . \tag{16}$$

As shown in Fig. 4, combined with the object motion, the dynamic attention map can be generated as follows:

$$S_d = S + \gamma_t S \Delta_{x,y} \ . \tag{17}$$

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

IV. Experiments

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 (HOG) [34] and color names [17], to represent object and context patches. The regularization parameter λ_1 and λ_2 is set to 5×10^{-5} and 0.0625, respectively. Besides, the interval for context learning f_c and the number of context patches S are set as 2 and 4. Details of TACF tracker can be seen in Algorithm 1. All hyper-parameters are fixed for all the experiments. The source code and related UAV tracking video are available in https://github.com/vision4robotics/TACF-Tracker and https://youtu.be/IUJpFgXKCvc.

2) Evaluation methodology: To validate the effectiveness of the proposed method, extensive experiments on UAV tracking benchmark, i.e., UAVDT [31] and UAV123@10fps [30] are conducted to evaluate performance. In terms of success rate, intersection over union (IoU) is used to measure the tracking performance, which is calculated as the ratio between the estimated and the ground truth bounding boxes. In this work, the area under the curve (AUC) is adopted to rank trackers.

B. Qualitative experiments

1) State-of-the-art Comparison: In real-world tasks, economical and efficient operations on the UAV platform is required, while operating with limited hardware capabilities. Thus, the TACF tracker is evaluated on 173 challenging images sequences from the aforementioned benchmarks compared with other state-of-the-art trackers with hand-crafted features, including MCCT_H [29], Staple_CA [5], SRDCF [6], BACF [4], KCC [33], CSRDCF [8], SAMF [35], Staple [7], KCF [18], SRDCFdecon [36], STRCF [9], fDSST [16], and ECO_HC [23]. The open-source codes of these trackers with default parameters provided by the authors have been used in the following evaluation and experiments.

Figure 5 shows the proposed tracker achieves superiority compared with other trackers in success plots. On UAVDT dataset, the TACF tracker 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 UAV123@10fps dataset, the TACF keeps the best score, outperforming the second (ECO_HC, 0.462) and third (STRCF, 0.457) best tracker with the same hand-crafted features.

2) Attribute-based performance analysis: The performance of the TACF tracker and other trackers are also analyzed in different attributes. Figure 6 shows the scores of different trackers on different attributes and demonstrates TACF exhibits better performance than most of the other trackers except for fast motion. Especially when background clutter or occlusion occurs, the proposed TACF 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 tri-attention strategy, CF can focus on crucial aspects so that the tracker can show better performance in these complex situations.

In this paper, position-attention is introduced to some extent to alleviate the problems resulted from fast motion. In the future, it is possible to employ a lightweight CNN to extract convolutional features for object representation to replace hand-crafted features with limited encoding ability. Deep features can provide semantic information and ensure

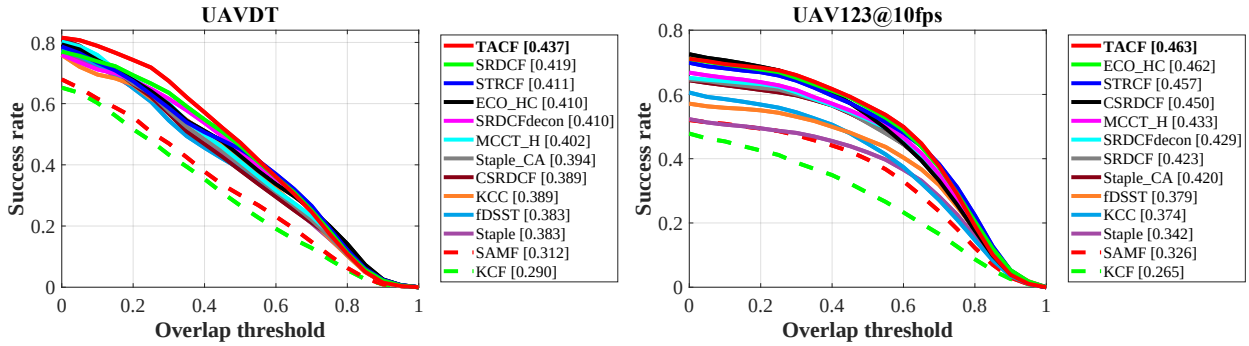

Fig. 5. Success plots of the proposed TACF tracker and other 12 state-of-the-art trackers on UAVDT (left) and UAV123@10fps (right) datasets.

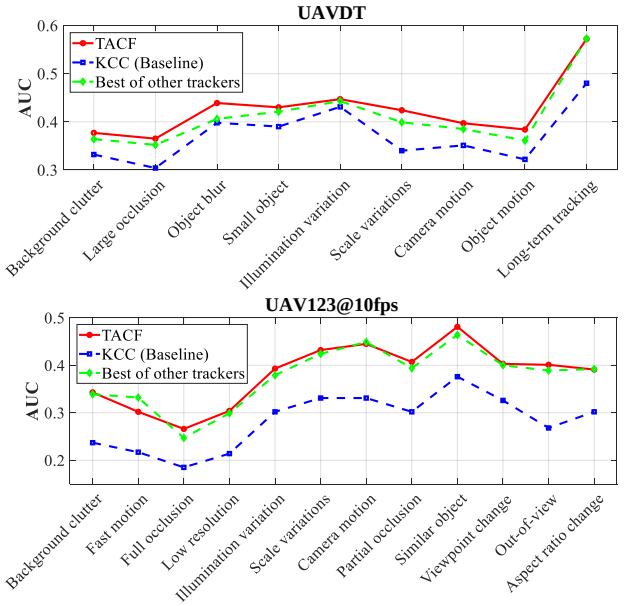

Fig. 6. Success scores for 9 and 12 attribute-based experiments on the UAVDT and UAV123@10fps datasets, respectively. In most cases, the proposed TACF tracker shows superiority over the best of other trackers, achieving significant improvements compared to the baseline.

efficient operation so that the tracker can improve the performance against object fast motion.

3) Ablation study: In this section, the performance of baseline, the permuted and complete versions of the proposed TACF trackers is evaluated extensively. Apart from the success score, frame per second (FPS) and millisecond per frame (MSPF) are applied to evaluate the operation speed of the tracker. The baseline tracker, KCC, is considered as a special case of TACF with HOG feature and no context information. TACF1, TACF2, and TACF3 are the proposed method with the context-attention, dimension-attention, and position-attention only.

Table I presents that TACF tracker has hugely 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.

Although the proposed method increases the average processing time compared to the baseline, the accuracy and robustness are greatly improved, which still can meet the practical implementation for UAV object tracking. On the one hand, due to the introduction of context-attention, the processing speed for large object targets is reduced. Accordingly, proper image down-sampling can be considered to speed up while ensuring accuracy. 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 applying to real-world tasks.

TABLE I

COMPARISONS BETWEEN TACF WITH DIFFERENT MODULES ON UAVDT AND UAV@10FPS. **BOLD** FONT INDICATES THE BEST PERFORMANCE IN SUCCESS SCORES. ALL TESTS ARE EVALUATED SOLELY ON CPU MODE.

Tracker	Success		FPS	MSPF
	UAVDT	UAV123@10fps		
KCC	0.389	0.374	48.85	20.47
TACF1	0.432	0.421	31.01	32.25
TACF2	0.425	0.407	35.17	28.43
TACF3	0.423	0.398	47	20.98
TACF	**0.437**	**0.456**	24.2	41.32

V. CONCLUSIONS

In this work, a novel tri-attention correlation filters for effective UAV object tracking, i.e., TACF tracker, is proposed to achieve high performance in tracking applicants. The practical and brief tri-attention strategy can make contributions to enhance response maps and improve filters training simultaneously. Three types of attention, i.e., position, dimension, and context-attention, have been presented and verified its effectiveness. Moreover, extensive and in-depth experiments on two challenging UAV image datasets show that the presented TACF tracker has performed favorably against 12 trackers, which are the state-of-the-art tracking methods in the literature, in terms of accuracy, robustness, and efficiency. Besides, the proposed method can be integrated into other CF-based methods. We believe, with our proposed tri-attention strategy, the correlation filters can achieve performances further.

ACKNOWLEDGMENT

The work was supported by the National Natural Science Foundation of China (No.61806148) and the Fundamental Research Funds for the Central Universities (No.22120180009).

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 *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2017, pp. 1732–1738.

[2] R. Opromolla, G. Fasano, and D. Accardo, "A vision-based approach to uav detection and tracking in cooperative applications," *Sensors*, vol. 18, no. 10, p. 3391, 2018.

[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] 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*, 2017, pp. 1135–1143.

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

[6] 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*, 2015, pp. 4310–4318.

[7] 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*, 2016, pp. 1401–1409.

[8] 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*, 2017, pp. 6309–6318.

[9] 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*, 2018, pp. 4904–4913.

[10] 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*, 2019, pp. 4670–4679.

[11] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[12] 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*, 2018, pp. 3–19.

[13] 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.

[14] 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*, 2019, pp. 3146–3154.

[15] 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*, 2010, pp. 2544–2550.

[16] M. Danelljan, G. Häger, F. Khan, and M. Felsberg, "Accurate scale estimation for robust visual tracking," in *British Machine Vision Conference*, 2014.

[17] 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*, 2014, pp. 1090–1097.

[18] 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.

[19] 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*, 2019.

[20] 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 *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2019.

[21] 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*, 2015, pp. 3074–3082.

[22] 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*, 2016, pp. 472–488.

[23] 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*, 2017, pp. 6638–6646.

[24] G. Bhat, J. Johnander, M. Danelljan, F. Shahbaz Khan, and M. Felsberg, "Unveiling the power of deep tracking," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 483–498.

[25] 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.

[26] 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*, 2017, pp. 2805–2813.

[27] 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*, 2018, pp. 548–557.

[28] 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*, 2017, pp. 4807–4816.

[29] 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*, 2018, pp. 4844–4853.

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

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

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

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

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

[35] Y. Li and J. Zhu, "A scale adaptive kernel correlation filter tracker with feature integration," in *ECCV workshops*, 2014, pp. 254–265.

[36] 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*, 2016, pp. 1430–1438.