Robust Visual Object Tracking via Adaptive Attribute-Aware Discriminative Correlation Filters

Xue-Feng Zhu, Xiao-Jun Wu, *Member, IEEE* Tianyang Xu, Zhen-Hua Feng, *Member, IEEE*
and Josef Kittler, *Life Member, IEEE*

***Abstract*—In recent years, attention mechanisms have been widely studied in Discriminative Correlation Filter (DCF) based visual object tracking. To realise spatial attention and discriminative feature mining, existing approaches usually apply regularisation terms to the spatial dimension of multi-channel features. However, these spatial regularisation approaches construct a shared spatial attention pattern for all multi-channel features, without considering the diversity across channels. As each feature map (channel) focuses on a specific visual attribute, a shared spatial attention pattern limits the capability for mining important information from different channels. To address this issue, we advocate channel-specific spatial attention for DCF-based trackers. The key ingredient of the proposed method is an Adaptive Attribute-Aware spatial attention mechanism for constructing a novel DCF-based tracker (A^{3}DCF). To highlight the discriminative elements in each feature map, spatial sparsity is imposed in the filter learning stage, moderated by the prior knowledge regarding the expected concentration of signal energy. In addition, we perform a post processing of the identified spatial patterns to alleviate the impact of less significant channels. The net effect is that the irrelevant and inconsistent channels are removed by the proposed method. The results obtained on a number of well-known benchmarking datasets, including OTB2015, DTB70, UAV123, VOT2018, LaSOT, GOT-10K and TrackingNet, demonstrate the merits of the proposed A^{3}DCF tracker, with improved performance compared to the state-of-the-art methods.**

***Index Terms*—Visual object tracking, discriminative correlation filter, visual attribute, spatial attention**

I. Introduction

Visual object tracking is a fundamental research topic in computer vision, with the aim to precisely and continuously estimate the state of a target of interest in a video. It is a very challenging task to achieve robust and efficient tracking under unconstrained scenarios, due to a wide spectrum of appearance variations of the target in a video. To improve the performance of a visual object tracker, various innovative ideas have been explored and a significant progress has been made in recent years. Among existing visual tracking algorithms, the Discriminative Correlation Filter (DCF) based tracking algorithms have exhibited encouraging performance and drawn widespread attention.

A DCF-based tracker formulates the learning objective as a ridge regression problem with a circulant matrix structure, which simplifies the filter optimisation [1], [2] and thus results in the use of more sophisticated features [3], [4] and regularised filter learning [5], [6], [7]. Most existing DCF trackers use multiple types of features, such as Histogram of Oriented Gradient (HOG) [8], Colour Names (CN) [9] and Convolutional Neural Network (CNN) features [10], [11], [12]. For each feature type, multiple channel features capturing different properties of the visual content are extracted for filter learning. Besides the exploitation of powerful features, the regularisation of the estimated filters also plays an important role in robust visual tracking by endowing tracking systems with an attention mechanism, which helps to highlight the most discriminative and important visual information of a target. To this end, both fixed [5], [6] and adaptive spatial regularisation [13], [14], [15], [16] have been explored in the DCF paradigm to achieve attention-based discriminative spatial appearance modelling. With powerful feature representations and spatial attention mechanisms, the DCF-based trackers have witnessed a continuous performance enhancement.

In spite of the success of DCF, many aspects such as the relevance of multi-channel feature maps, regularised learning formulation, spatial attention mechanisms and discriminative data fitting, have not been adequately explored [4], [13], [14], [17], [18]. First, the extracted multi-channel features maps, which may exceed thousands of channels, include irrelevant and redundant information [14], [4]. The filters trained with such feature maps often contain negligible energy and may degrade the performance of a DCF tracker [4], [16]. So far there are only a very few works focusing on reducing the information redundancy of the feature maps in DCF-based tracking [13], [14]. Besides, for spatial regularisation, existing DCF trackers exploit a shared spatial attention pattern across all the feature channels, which neglects the descriptive diversity and discriminative competitiveness of the features in different channels. To address the above issue, we propose to adaptively optimise the spatial configuration pattern for each feature map (channel), reflecting the intrinsic link between visual attributes and attention. In essence, we propose to learn Adaptive Attribute-Aware Discriminative Correlation Filters

X.-F Zhu and X.-J Wu (corresponding author) are with the School of Artificial Intelligence and Computer Science, Jiangnan University, Wuxi, P.R. China. (e-mail: xuefeng_zhu95@163.com, wu_xiaojun@jiangnan.edu.cn)

T. Xu is with the School of IoT, Jiangnan University, Wuxi, P.R. China and the Centre for Vision, Speech and Signal Processing, University of Surrey, Guildford, GU2 7XH, UK. (e-mail: tianyang_xu@163.com)

Z.-H Feng is with the Department of Computer Science and the Centre for Vision, Speech and Signal Processing, University of Surrey, Guildford, GU2 7XH, UK. (e-mail: z.feng@surrey.ac.uk)

J. Kittler is with the Centre for Vision, Speech and Signal Processing, University of Surrey, Guildford, GU2 7XH, UK. (e-mail: j.kittler@surrey.ac.uk)

DOI: https://doi.org/10.1109/TMM.2021.3050073

(A^{3}DCF) for robust visual tracking.

We start by noting that each feature map in a tensor representation potentially emphasises a unique visual attribute (pattern), and different feature maps focus on different visual attributes. For example, each channel of CN features reflects one specific colour attribute [9] and each channel of HOG features corresponds to different gradient orientations [8]. Similarly, the semantic information captured by different deep CNN feature channels is typically distinct [19]. Based on the above observation, we propose to perform adaptive channel-wise spatial attention learning for correlation filters, based on the discrimination relevance of each feature channel and its corresponding filter. Specifically, we simultaneously optimise the filter coefficients and channel-wise binary spatial attention pattern in our paradigm. Through this attribute-aware scheme, the discriminative elements in each feature map are highlighted in support of enhanced parsimony and compactness. Moreover, by virtue of the advocated attention mechanism, the proposed A^{3}DCF method accomplishes adaptive feature suppression by enforcing a prior constraint on the filters so that the energy is concentrated on the central region of a search window. In consequence, we eliminate irrelevant and interfering information in the learning stage, enhancing the discriminative capability of the trained model and resulting in better tracking performance.

We illustrate the learned attention patterns of our A^{3}DCF method in Fig. 1. In the figure, the first column is the search window centred on the position of the target. The second column contains the multi-channel CN feature maps of the search window. The third column presents the initial spatial patterns which are adaptively learned from the input features according to the visual attribute of each channel. The fourth column shows the final learned attribute-related spatial patterns obtained by performing a post processing of the preliminary spatial patterns. Specifically, the fourth column is generated by imposing shared spatial patterns, that is a prior constraint, on the third column. The final column displays the corresponding correlation filters optimised with the attribute-related spatial patterns. As some feature channels provide ambiguous appearance information, they can be considered as less representative or irrelevant channels. Through the spatial attention learned by the proposed adaptive attribute-aware mechanism, these irrelevant channels can be eliminated. Therefore, the adaptive spatial regularisation and feature selection across channels are realised simultaneously, resulting in enhanced discrimination of the learned filters.

To summarise, the main innovations of the proposed A^{3}DCF method include:

- A new adaptive attribute-aware scheme is proposed to emphasise channel-specific discriminative features, invariably related to the corresponding attribute it represents. With the learned adaptive spatial attention patterns, irrelevant information of multi-channel features is significantly reduced and the boundary effect is alleviated.
- The proposed post processing of the initial attribute-related spatial patterns preserves only the feature channels representing discriminative attributes, while irrelevant and inconsistent channels are suppressed.

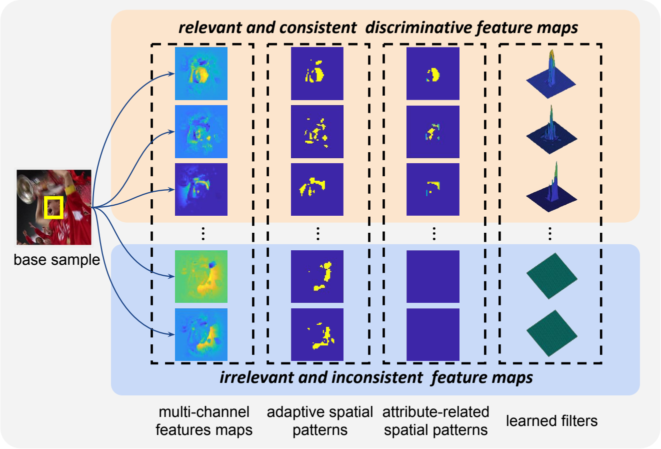

relevant and consistent discriminative feature maps
base sample
irrelevant and inconsistent feature maps
multi-channel features maps
adaptive spatial patterns
attribute-related spatial patterns
learned filters

Fig. 1. Illustration of the adaptive attribute-related spatial patterns. Noting that each feature map (channel) focuses on one exclusive attribute, the proposed adaptive attribute-aware spatial configurations are optimised jointly by discriminative data fitting.

- An extensive evaluation is performed on a number of well-known benchmarks, *i.e.* OTB2015 [20], DTB70 [21], UAV123 [22], VOT2018 [23], LaSOT [24], GOT-10K [25] and TrackingNet [26]. The experimental results demonstrate the superior performance of the proposed A^{3}DCF method over the state-of-the-art algorithms, in terms of both effectiveness and robustness.

II. Related Work

There are numerous reputable studies in visual object tracking. In this section, we briefly discuss the recent filter based approaches, *i.e.*, Siamese networks and Discriminative Correlation Filters. The main objective of Siamese networks is to learn an offline mapping function producing similar features for naturally varying target appearance, while suppressing the surroundings. The filters are generated by embedding a target instance (template) into an appearance variation preserved feature space, where the target's centre corresponds to high response [27]. In view of the diversity of target categories, CFnet [28] was proposed to generate adaptive filters with an additional structure to reflect the template appearance for different sequences better. To further improve the precision of the target centre location, other approaches were developed to support more complicated network constructions [29], [30], [31]. To accurately predict the bounding box, the SiameseRPN framework [32], [33], [34], [35] was designed to simultaneously optimise the classification loss and a bounding box regression loss. Though advanced efficiency can be achieved by Siamese networks via offline learning the mapping function, the relevance and causality between the embedding features and the response generators are ambiguous and of limited interpretability.

Regarding DCF approaches, thanks to the seminal work of MOSSE [36], DCF has received much attention in visual object tracking, especially since the development of the CSK tracker [1] that embeds kernel tricks in the circulant structure [37]. Considering the importance of robust feature representation [38] and the unexpected boundary effect produced by the implicit assumption of periodicity of the input signal induced by circulant samples [39], a variety of advanced DCF-based trackers have been proposed [40], [41], [42]. As robust feature extraction is crucial to advanced DCF training, Henriques *et al.* adopted HOG features in KCF [2] and Danelljan *et al.* proposed to employ CN features [43], resulting in improved performance, as compared with the trackers using greyscale representations. Recently, due to the remarkable performance of Convolutional Neural Networks (CNN), deep features have been widely used in DCF-based trackers and shown to be instrumental in achieving advanced tracking performance [44]. Modern DCF trackers usually use multiple features, such as HOG, CN and CNN, collaboratively for robust feature extraction [45], [4], [13], [14]. However, these DCF-based trackers simply assemble all the features for learning discriminative filters, without considering the effect of the potentially inconsistent and redundant information they may convey, which may degrade the performance of the learned filters. These deficiencies motivate the development of our A^{3}DCF which is designed to reduce the redundancy and interference of feature representations in the filter learning stage, which leads to improved robustness and better discriminative ability of the learned model.

To cope with the issue of spatial boundary effect, Danelljan *et al.* proposed the spatially regularised DCF method that introduced a fixed opposite Gaussian-shaped spatial weighting mask for correlation filters learning [39], concentrating the energy of the learned filters on the central target region. Similarly, CSRDCF [6] proposed to generate a mask using colour-histogram-based image segmentation to suppress the background area. In the same spirit, BACF [5] employs a predefined binary matrix to crop valid training samples. In contrast to a fixed spatial mask, an adaptive spatial regularisation is proposed in LADCF [13] to learn an adaptive DCF via spatial feature selection. The above trackers effectively alleviate the problem of boundary effect and achieve convincing results in recent benchmarks and competitions [20], [46], [23], [47], [48]. Nevertheless, these methods enforce the same attention mechanism on all filter channels and therefore cannot adapt to the diversity of visual attributes linked to different feature maps. We argue, therefore, that it is absolutely essential to identify an attribute-related spatial pattern for each filter channel separately, so as to enhance discrimination and simultaneously alleviate the boundary effect.

III. Adaptive Attribute-Aware DCF

A. Standard DCF Formulation

We first briefly revisit the standard DCF tracking approach that follows a 2-stage tracking-learning framework for online object tracking. Let us assume that the location of a target in the t-th frame is known. Given this information, multi-channel correlation filters $\mathbf{H}$ are trained by minimising the following objective function in the learning stage:

$$E(\mathbf{H}) = \frac{1}{2}\left\|\sum_{k=1}^{K}\mathbf{X}_k \star \mathbf{H}_k - \mathbf{Y}\right\|_F^2 + \frac{\lambda}{2}\sum_{k=1}^{K}\|\mathbf{H}_k\|_F^2, \quad (1)$$

where $\mathbf{X}_k \in \mathbb{R}^{N\times N}$ is the k-th channel feature representation of an image patch (search window) centred on the position of the predicted target, $\mathbf{H}_k \in \mathbb{R}^{N\times N}$ denotes the corresponding k-th channel filter. $\mathbf{Y} \in \mathbb{R}^{N\times N}$ stands for the desired detector response map [1] of Gaussian shape. K is the number of feature channels and λ is a regularisation parameter. $\|\cdot\|_F$ means Frobenius norm and $\star$ is the circular convolution operator [2]. With the circulant structure [37] and Fourier transform [49], the optimisation of Eqn. (1) can efficiently be solved in the frequency domain [2].

To prevent their temporal degradation, the final filters are updated online after obtaining the trained discriminative correlation filters $\mathbf{H}$ in the t-th frame:

$$\mathbf{H}^t = (1-\eta)\mathbf{H}^{t-1} + \eta\mathbf{H}, \quad (2)$$

where $\eta \in (0,1)$ is the online updating rate. Then the updated filters $\mathbf{H}^t$ are used to localise the target in the (t+1)-th frame:

$$\mathbf{R} = \mathscr{F}^{-1}\left(\sum_{k=1}^{K}\hat{\mathbf{Z}}_k^{t+1} \odot \hat{\mathbf{H}}_k^t\right), \quad (3)$$

where $\mathscr{F}^{-1}$ denotes the inverse Discrete Fourier Transform (DFT), $\hat{}$ denotes DFT and $\odot$ denotes point-wise multiplication. $\mathbf{Z}_k^{t+1} \in \mathbb{R}^{N\times N}$ is the k-th channel of multi-channel features extracted from the search region in the (t+1)-th frame and K is the number of feature channels. $\mathbf{R} \in \mathbb{R}^{N\times N}$ is the response map, in which the location of the maximal value is considered as the predicted target location.

B. Adaptive Attribute-Aware DCF Formulation

In the classical DCF paradigm, the filters are learned from multi-channel features extracted from a search window enclosing the target. However, these multi-channel features are often irrelevant and may contain inconsistent information that degrades the discriminative capability of the learned model. To address this issue, attention mechanisms realised by a spatial regularisation have been widely studied in recent advanced DCF-based trackers [39], [5], [6], [14]. In this work, noting that each feature map (channel) reflects one specific visual attribute, we propose an adaptive attribute-aware scheme in the filter learning stage to highlight the discriminative part of each channel. As illustrated in Fig. 2, the adaptive attribute-aware correlation filters can be learned by optimising the objective:

$$\begin{aligned} E(\mathbf{H},\mathbf{P}) = \frac{1}{2}\left\|\sum_{k=1}^{K}\mathbf{X}_k \star (\mathbf{H}_k \odot \mathbf{P}_k) - \mathbf{Y}\right\|_F^2 \\ + \frac{\lambda_1}{2}\sum_{k=1}^{K}\|\mathbf{H}_k\|_F^2 + \frac{\lambda_2}{2}\sum_{k=1}^{K}\|\mathbf{P}_k - \mathbf{P}_k^r\|_F^2, \end{aligned} \quad (4)$$

where $\mathbf{X}_k \in \mathbb{R}^{N\times N}$ is the k-th channel feature map, $\mathbf{H}_k \in \mathbb{R}^{N\times N}$ is the corresponding k-th correlation filter and $\mathbf{Y} \in \mathbb{R}^{N\times N}$ is the desired Gaussian shaped response map. $\mathbf{P}_k \in \mathbb{R}^{N\times N}$ is the spatial attention pattern, denoted as a binary matrix, corresponding to the k-th channel of filters. $\mathbf{P}_k^r \in \mathbb{R}^{N\times N}$ is a predefined binary matrix with the values of one for the target and zero for others. λ_1 and λ_2 are regularisation parameters. In Eqn. (4), the first term is the data fitting term, where the adaptive spatial attention patterns

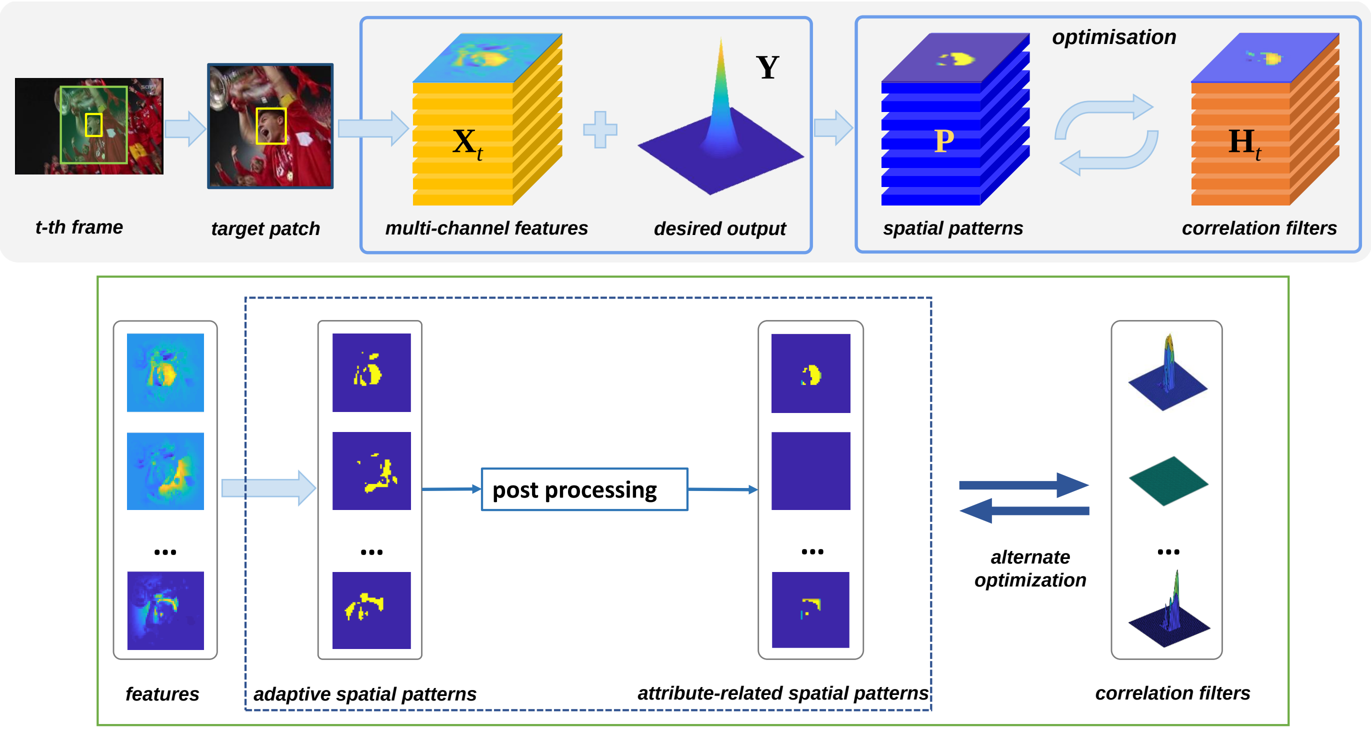


Fig. 2. Illustration of the pipeline of the proposed A^{3}DCF. The adaptive attribute-related spatial patterns and attribute-aware discriminative filters are simultaneously optimised in our design. In the green box, the specific process of optimising the adaptive attribute-related spatial patterns and attribute-aware filters is presented. The diversity across different feature maps is considered in our attention mechanisms to support enhanced discrimination and to alleviate irrelevant appearance information.

$\mathbf{P}$ and filters $\mathbf{H}$ are optimised simultaneously. The second term is a regularisation term to prevent over-fitting. The third term incorporates the prior requirement for the filters to concentrate their energy on the central region of the search window. This is expressed in the form of a binary mask, $\mathbf{P}^r$. It should be noted that each channel of $\mathbf{P}^r$ shares the same binary values.

C. Optimisation

According to the existing DCF design methodology, the optimisation step is usually performed in the Fourier domain for efficient filter learning and target detection. To obtain an efficient solution of Eqn. (4), the attention constraint can be formalised as $\mathbf{H} \equiv \mathbf{P} \odot \mathbf{H}$. Additionally, an auxiliary variable, $\mathbf{G}$, with the constraint $\mathbf{G} - \mathbf{H} = 0$, is introduced. We can now employ the following Augmented Lagrange Method [50] in the frequency domain:

$$\begin{aligned}\mathcal{L}(\hat{\mathbf{G}},\hat{\mathbf{H}},\mathbf{P},\hat{\mathbf{\Gamma}}) &= \frac{1}{2}\left\|\sum_{k=1}^{K}\hat{\mathbf{X}}_k \odot \hat{\mathbf{G}}_k - \hat{\mathbf{Y}}\right\|_F^2 \\ &+ \frac{\lambda_1}{2}\sum_{k=1}^{K}\left\|\hat{\mathbf{H}}_k\right\|_F^2 + \frac{\lambda_2}{2}\sum_{k=1}^{K}\|\mathbf{P}_k - \mathbf{P}_k^r\|_F^2 \\ &+ \sum_{k=1}^{K} tr\left(\hat{\mathbf{\Gamma}}_k^T(\hat{\mathbf{G}}_k - \hat{\mathbf{H}}_k)\right) + \frac{\mu}{2}\sum_{k=1}^{K}\left\|\hat{\mathbf{G}}_k - \hat{\mathbf{H}}_k\right\|_F^2,\end{aligned} \tag{5}$$

where $\mathbf{\Gamma}$, μ, and tr denote the Lagrangian multiplier, penalty factor, and trace calculation, respectively. It can be reformulated as follows by combining the last two terms:

$$\begin{aligned}\mathcal{L}(\hat{\mathbf{G}},\hat{\mathbf{H}},\mathbf{P},\hat{\mathbf{\Gamma}}) &= \frac{1}{2}\left\|\sum_{k=1}^{K}\hat{\mathbf{X}}_k \odot \hat{\mathbf{G}}_k - \hat{\mathbf{Y}}\right\|_F^2 \\ &+ \frac{\lambda_1}{2}\sum_{k=1}^{K}\left\|\hat{\mathbf{H}}_k\right\|_F^2 + \frac{\lambda_2}{2}\sum_{k=1}^{K}\|\mathbf{P}_k - \mathbf{P}_k^r\|_F^2 \\ &+ \frac{\mu}{2}\sum_{k=1}^{K}\left\|\hat{\mathbf{G}}_k - \hat{\mathbf{H}}_k + \frac{\hat{\mathbf{\Gamma}}_k}{\mu}\right\|_F^2,\end{aligned} \tag{6}$$

where $\mathbf{\Gamma}$ is the Lagrange multiplier and μ is a penalty factor. The Lagrangian in Eqn. (6) can be minimised by the Alternating Direction Method of Multipliers (ADMM) by solving the following sub-problems:

Sub-problem G: Given $\hat{\mathbf{H}}$, $\mathbf{P}$ and $\hat{\mathbf{\Gamma}}$, $\hat{\mathbf{G}}$ can be updated by solving:

$$\begin{aligned}\hat{\mathbf{G}} = \arg\min_{\hat{\mathbf{G}}} &\frac{1}{2}\left\|\sum_{k=1}^{K}\hat{\mathbf{X}}_k \odot \hat{\mathbf{G}}_k - \hat{\mathbf{Y}}\right\|_F^2 \\ &+ \frac{\mu}{2}\sum_{k=1}^{K}\left\|\hat{\mathbf{G}}_k - \hat{\mathbf{H}}_k + \frac{\hat{\mathbf{\Gamma}}_k}{\mu}\right\|_F^2.\end{aligned} \tag{7}$$

To solve this problem efficiently, we decompose it to process simultaneously all channels of each spatial unit. Decomposing

Eqn. (7) into *NN* sub-problems as :

$$\mathcal{V}_j(\hat{\mathbf{G}}) = \arg\min_{\mathcal{V}_j(\hat{\mathbf{G}})} \frac{1}{2}\left\|\mathcal{V}_j(\hat{\mathbf{X}})^T\mathcal{V}_j(\hat{\mathbf{G}}) - \hat{\mathbf{Y}}_j\right\|_2^2 + \frac{\mu}{2}\left\|\mathcal{V}_j(\hat{\mathbf{G}}) - \mathcal{V}_j(\hat{\mathbf{H}}) + \frac{1}{\mu}\mathcal{V}_j(\hat{\mathbf{\Gamma}})\right\|_2^2, \tag{8}$$

where $\mathcal{V}_j(\hat{\mathbf{G}}) \in \mathbb{R}^{K\times 1}$ denotes the vector containing the *j*-th spatial unit of $\hat{\mathbf{G}}$ with elements across all the channels, we can apply the Sherman-Morrison formula to obtain the closed-form solution to Eqn. (8) as:

$$\mathcal{V}_j(\hat{\mathbf{G}}) = \frac{1}{\mu}(\mathbf{I} - \frac{\mathcal{V}_j(\hat{\mathbf{X}})\mathcal{V}_j(\hat{\mathbf{X}})^T}{\mu + \mathcal{V}_j(\hat{\mathbf{X}})^T\mathcal{V}_j(\hat{\mathbf{X}})})\mathbf{q}, \tag{9}$$

where $\mathbf{q} = \mathcal{V}_j(\hat{\mathbf{X}})\hat{\mathbf{Y}}_j + \mu\mathcal{V}_j(\hat{\mathbf{H}}) - \mathcal{V}_j(\hat{\mathbf{\Gamma}})$.

Sub-problem H: Given $\hat{\mathbf{G}}$ and $\hat{\mathbf{\Gamma}}$, each channel of $\hat{\mathbf{H}}$ can be updated by solving the following sub-problem:

$$\hat{\mathbf{H}}_k = \arg\min_{\hat{\mathbf{H}}_k} \frac{\lambda_1}{2}\left\|\hat{\mathbf{H}}_k\right\|_F^2 + \frac{\mu}{2}\left\|\hat{\mathbf{G}}_k - \hat{\mathbf{H}}_k + \frac{\hat{\mathbf{\Gamma}}_k}{\mu}\right\|_F^2. \tag{10}$$

This sub-problem can be easily solved by setting the derivative to zero, and the closed-form solution is given by:

$$\hat{\mathbf{H}}_k = \frac{\mu\hat{\mathbf{G}}_k + \hat{\mathbf{\Gamma}}_k}{\lambda_1 + \mu}. \tag{11}$$

Sub-problem P: Since solving $\mathbf{P}$ is a binary optimisation problem, we relax and embed it into filters $\mathbf{H}$. First, we introduce a variable $\mathbf{J}$ as $\mathbf{J}_k = \mathbf{H}_k \odot \mathbf{P}_k$ and a variable $\mathbf{J}^r$ as $\mathbf{J}^r_k = \mathbf{H}_k \odot \mathbf{P}^r_k$. Then given $\hat{\mathbf{G}}$, $\mathbf{H}$ and $\hat{\mathbf{\Gamma}}$, $\hat{\mathbf{J}}$ can be solved by optimising:

$$\hat{\mathbf{J}} = \arg\min_{\hat{\mathbf{J}}} \frac{1}{2}\left\|\sum_{k=1}^{K}\hat{\mathbf{X}}_k \odot \hat{\mathbf{J}}_k - \hat{\mathbf{Y}}\right\|_F^2 + \frac{\lambda_2}{2}\sum_{k=1}^{K}\left\|\hat{\mathbf{J}}_k - \hat{\mathbf{J}}^r_k\right\|_F^2. \tag{12}$$

Note that Eqn. (12) can be solved in the same way as Eqn. (7). Hence we get the analytical solution as :

$$\mathcal{V}_j(\hat{\mathbf{J}}) = \frac{1}{\lambda_2}(\mathbf{I} - \frac{\mathcal{V}_j(\hat{\mathbf{X}})\mathcal{V}_j(\hat{\mathbf{X}})^T}{\lambda_2 + \mathcal{V}_j(\hat{\mathbf{X}})^T\mathcal{V}_j(\hat{\mathbf{X}})})(\mathcal{V}_j(\hat{\mathbf{X}})\hat{\mathbf{Y}}_j + \lambda_2\mathcal{V}_j(\hat{\mathbf{J}}^r)). \tag{13}$$

In Eqn. (13), $\mathbf{J}^r$ can be computed directly with predefined $\mathbf{P}^r$. Once $\mathbf{J}$ is found in the spatial domain by the inverse DFT of $\hat{\mathbf{J}}$, we subject it to a binarisation process to produce the binary matrix $\mathbf{P}$. To apply binarisation, we set the first $r\%$ values in each channel to one and all the others to zero, according to the absolute value of $\mathbf{J}$. In fact, in the *t*-th frame (except the first frame), for temporal consistency, we apply binarisation to $\mathbf{J}^t$ to obtain $\mathbf{P}$ and $\mathbf{J}^t = (1-\gamma)\mathbf{J}^{t-1} + \gamma\mathbf{J}$, where γ is a predefined constant. The $\mathbf{P}$ is the initial attribute-related spatial patterns, as exhibited in the third column of Fig. 1. To obtain the final attribute-related spatial patterns, a post processing operation using the identified spatial patterns, namely the prior constraint $\mathbf{P}^r$, is performed to update $\mathbf{P}$. The specific steps are summarised as follows.

Post Processing: After obtaining the initial attribute-related spatial attention patterns, $\mathbf{P}$, as illustrated in the third column of Fig. 1, a post processing operation is applied to $\mathbf{P}$. Specifically, we update $\mathbf{P}$ by enforcing a prior constraint as:

$$\mathbf{P}_k = \mathbf{P}_k \odot \mathbf{P}^r_k. \tag{14}$$

The updated $\mathbf{P}$ are the final attribute-related spatial attention patterns as shown in the fourth column of Fig. 1. By doing this, the spatial attention patterns focus more on the central region and alleviate the impact of less representative channels.

Update H: After updating channel-specific spatial patterns $\mathbf{P}$, we update the filters $\mathbf{H}$ by enforcing the attention patterns on filters $\mathbf{H}$ as:

$$\mathbf{H}_k = \mathbf{H}_k \odot \mathbf{P}_k. \tag{15}$$

It should be noted that the filters $\mathbf{H}$ are updated in the spatial domain and should be transformed into the frequency domain to update the Lagrangian multipliers.

Update Lagrangian Multiplier: Given $\hat{\mathbf{G}}$, $\hat{\mathbf{H}}$ and μ, the Lagrangian multipliers $\hat{\mathbf{\Gamma}}$ can be updated as :

$$\hat{\mathbf{\Gamma}} = \hat{\mathbf{\Gamma}} + \mu(\hat{\mathbf{G}} - \hat{\mathbf{H}}). \tag{16}$$

After each iteration, the step size parameter μ is updated as:

$$\mu = \min(\mu_{max},\ \rho\mu), \tag{17}$$

where μ_{max} denotes the predefined maximum value of μ, and $\rho > 1$ is a factor that controls the convergence speed.

Therefore, the optimisation of our proposed A^{3}DCF can be processed by iteratively employing the above steps. The total complexity of our optimisation framework is $O(LKN^2log(N))$, where L is the number of iterations. After optimisation, the filters $\mathbf{H}$ are used to update the filter model and localise the target in the next frame.

D. Online Update and Object Detection

For adaptation to appearance variations, the model is online updated by utilising the same updating strategy as other DCF-based trackers, as presented in Eqn. (2). Additionally, similar to many other DCF-based trackers, we follow the fast Discriminative Scale Space Tracking (fDSST) method [51] to achieve position localisation and scale estimation simultaneously. When a new frame becomes available, multiple scales a^s of the search region $\{\mathbf{X_s}\}_{\mathbf{s}\in\{\lfloor\frac{\mathbf{1-S}}{\mathbf{2}}\rfloor,\ldots,\lfloor\frac{\mathbf{S-1}}{\mathbf{2}}\rfloor\}}$ are analysed to extract features, where S is the number of scales. Then a filter, $\mathbf{H}_s$, is designed for each $\mathbf{X}_s$ in the Fourier domain to obtain multi-channel response maps by using Eqn. (3). The position and scale of target p_t and s_t are predicted according to the maximum of these response maps.

IV. Experiments

A. Implementation Details

The proposed A^{3}DCF method is implemented in MATLAB2019a on a platform with one Intel i7-9700 3.00GHZ CPU and a single NVIDIA GeForce GTX 1660Ti GPU. For feature extraction, we use three hand-crafted features including

TABLE I
TRACKING RESULTS WITH DIFFERENT FEATURES ON OTB2015 IN TERMS OF OP SCORES AND FPS.

Feature	HOG				HOG+CN			HOG+CN+ResNet-50		
Method	SRDCF	CSRDCF	BACF	A^{3}DCF	C-COT	ECO	A^{3}DCF	UPDT	GFSDCF	A^{3}DCF
OP	71.1%	70.5%	77.6%	**78.1%**	75.7%	78.0%	**80.2%**	88.7%	89.0%	**89.4%**
FPS	5.2	7.4	38.3	42.6	1.5	26.1	26.9	5.4	4.8	4.2

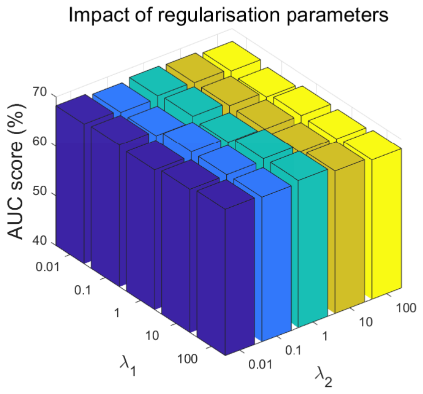


Fig. 3. The AUC score of A^{3}DCF on OTB2015 dataset with different regularisation parameters λ_1 and λ_2, ranging from 0.01 to 100.

TABLE II
A COMPARISON OF THE BASELINE, BASELINE_AAA, BASELINE_PP, BASELINE_GAM, BASELINE_ALL ON THE OTB2015 DATASET.

Method	AUC	DP
BaseLine	63.9%	84.8%
BaseLine_AAA	67.5%	89.9%
BaseLine_PP	67.9%	91.9%
BaseLine_GAM	68.2%	91.7%
BaseLine_ALL	69.6%	92.5%

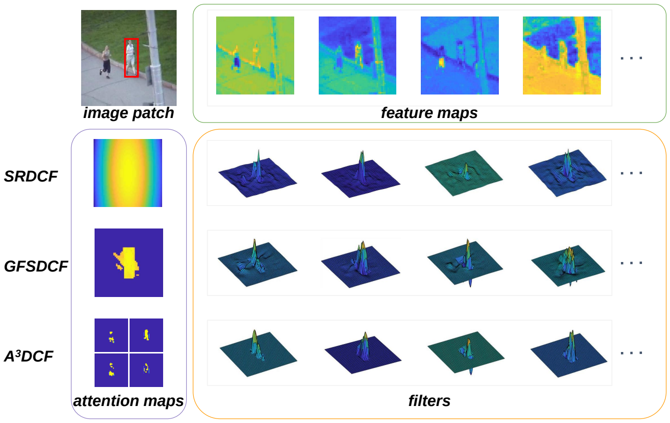


Fig. 4. Visualisation of attention maps and filters using the frame #28 of *Jogging-2* in OTB2015. We visualise 4 channels of the CN feature maps, attention maps and corresponding learned filters.

HOG, CN and Intensity Channels, and the *Res4e* layer of the pre-trained ResNet-50 model as deep CNN features. We set λ_1, λ_2 in Eqn. (4) and γ as $\lambda_1 = 100$, $\lambda_2 = 1$ and $\gamma = 0.05$. The number of iterations L is set to $L = 2$. For hand-crafted features, we set the corresponding parameters as $\eta = 0.01$, $r = 5$. For deep features, we set the corresponding parameters as $\eta = 0.003$, $r = 20$. The source code will be made publicly available.

B. Evaluation Settings

We extensively evaluate our A^{3}DCF tracker on four benchmarks including OTB2015 [20], DTB70 [21], UAV123 [22], VOT2018 [23], LaSOT [24], GOT-10K [25] and TrackingNet [26], compared with numerous state-of-the-art visual object tracking approaches, such as ASRCF [16], GFSDCF [14], BACF [5], SRDCF [39], Staple [52], STAPLE_CA [7], MDNet [53], GOTURN [54], CF2 [55] CSRDCF [6], C-COT [3], ECO [4], CREST [56], MCPF [57], STRCF [18], LADCF [13], SiamFC [27], DSiam [30], CFNet [28], StructSiam [58], LSART [59], DRT [60], MFT [23], SiamRPN [32], GCT [61], GradNet [62], UPDT [17], TADT [63] and fdKCF [64] respectively.

For OTB2015, DTB70, UAV123 and LaSOT, we use the precision plot (measure centre location error) and success plot (measure bounding box overlap ratio) as metrics. Besides, four numerical criteria, *i.e.*, Distance Precision (DP), Overlap Precision (OP), Area Under Curve (AUC) and Frames Per Second (FPS) are used in the evaluation. For VOT2018, Expected Average Overlap (EAO), Accuracy (A) and Robustness (R) are used as criteria for performance evaluation. For GOT-10K, Average Overlap (AO) and Success Rate (SR) are used to evaluate the performance of a tracker. For TrackingNet, Precision, Normalized Precision and Success are employed to quantitatively compare the performance of the algorithms. Specifically, DP is the percentage of location errors within a threshold of 20 pixels. OP is the percentage of overlap ratios surpassing a threshold of 0.5. AUC is the value of area under the curve of success plot. FPS measures the speed of a tracker. EAO is the main evaluation metric in VOT challenges which considers both bounding box overlap and failures (robustness). The A measure denotes the accuracy value of a tracker. The R measure quantifies the robustness of a tracker (the lower the better). The AO metric is the average of overlap rates between the tracking results and groundtruth over all the video frames. SR is the percentage of successfully tracked frames where the overlap rates are above a threshold. Precision is the same as Distance Precision (DP) and Success is the same as Area Under Curve (AUC). Normalized Precision is measured by normalizing the Precision over the size of the ground truth bounding box.

C. Self Analysis

Parameters Analysis: In this section, we first analyse the sensitivity of our A^{3}DCF tracker to parameters λ_1, λ_2 in Eqn. (4). In Fig. 3, we report the variation of the AUC scores

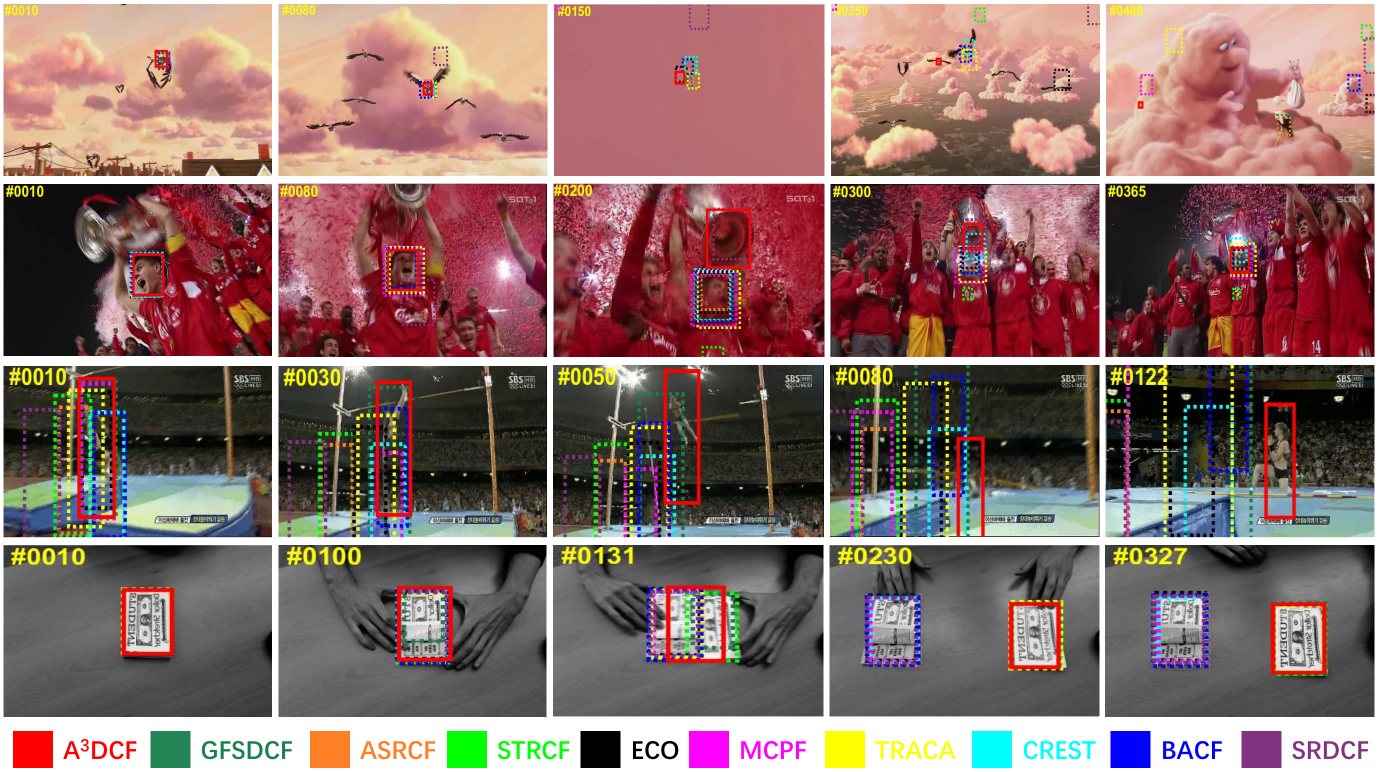


Fig. 5. Showing some failed cases on challenging video sequences including *Bird1, Soccer, Jump, Coupon*.The colour bounding boxes are the corresponding results of A^{3}DCF, GFSDCF, ASRCF, STRCF, ECO, MCPF, TRACA, CREST, BACF and SRDCF respectively.

TABLE III
EVALUATION OF 11 ADVANCED TRACKERS ON THE VOT2018 DATASET IN TERMS OF EAO, A, AND R. THE TOP THREE RESULTS ARE SHOWN IN RED, BLUE AND GREEN COLOURS RESPECTIVELY.

	ECO	LSART	LADCF	MFT	SiamRPN	DRT	STRCF	UPDT	ASRCF	GFSDCF	A^{3}DCF
EAO	0.280	0.323	0.389	0.385	0.383	0.356	0.345	0.378	0.328	0.397	0.406
A	0.484	0.495	0.503	0.505	0.586	0.519	0.523	0.536	0.494	0.511	0.548
R	0.276	0.218	0.159	0.140	0.276	0.201	0.215	0.184	0.234	0.143	0.162

TABLE IV
EVALUATION OF 8 ADVANCED TRACKERS ON THE GOT-10K DATASET IN TERMS OF AO, SR$_{0.5}$, AND SR$_{0.75}$. THE TOP THREE RESULTS ARE SHOWN IN RED, BLUE AND GREEN COLOURS RESPECTIVELY.

	MDNet	CF2	ECO	CCOT	GOTURN	SiamFC	CFNet	A^{3}DCF
AO	0.299	0.315	0.316	0.325	0.342	0.348	0.374	0.427
SR$_{0.5}$	0.303	0.297	0.309	0.328	0.372	0.353	0.404	0.467
SR$_{0.75}$	0.099	0.088	0.111	0.107	0.107	0.098	0.144	0.138

TABLE V
EVALUATION OF 7 ADVANCED TRACKERS ON THE TRACKINGNET DATASET IN TERMS OF SUCCESS, PRECISION, AND NORMALIZED PRECISION. THE TOP THREE RESULTS ARE SHOWN IN RED, BLUE AND GREEN COLOURS RESPECTIVELY.

	STAPLE_CA	ECO	SiamFC	CFNet	MDNet	GFSDCF	A^{3}DCF
Success (%)	53.6	56.1	57.1	57.8	61.4	60.9	62.5
Precision (%)	46.7	48.9	53.3	53.3	55.5	56.6	58.0
Norm.Prec. (%)	60.8	62.1	66.3	65.4	71.0	71.8	72.8

TABLE VI
THE OP SCORES OF TRACKERS ON THE OTB2013, UAV123 AND LASOT DATASETS. THE TOP THREE RESULTS ARE SHOWN IN RED, BLUE AND GREEN COLOURS RESPECTIVELY.

		Staple	CSRDCF	BACF	SRDCF	ECO	STRCF	ASRCF	GFSDCF	A^{3}DCF
Avg. OP(%)	OTB2015	70.2	70.5	77.6	71.1	84.9	84.6	87.6	89.0	89.4
	UAV123	53.7	49.9	53.7	54.9	63.7	60.9	61.7	67.8	67.6
	LaSOT	24.0	22.4	26.3	24.5	32.9	32.5	36.6	41.4	42.3

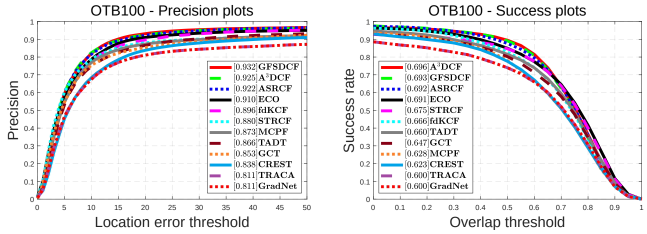


Fig. 6. The experimental results on OTB2015. This figure shows the precision and success plots in terms of the OPE protocol. The DP and AUC score of each tracker is shown in the legend.

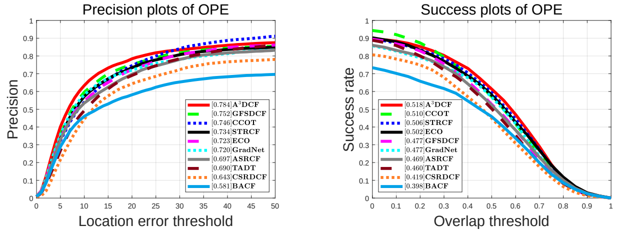


Fig. 7. The experimental results on DTB70. This figure shows the precision and success plots in terms of the OPE protocol. The DP and AUC score of each tracker is shown in the legend.

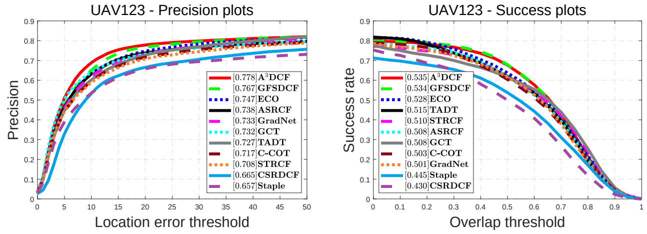


Fig. 8. The experimental results on UAV123. This figure shows the precision and success plots in terms of the OPE protocol. The DP and AUC score of each tracker is shown in the legend.

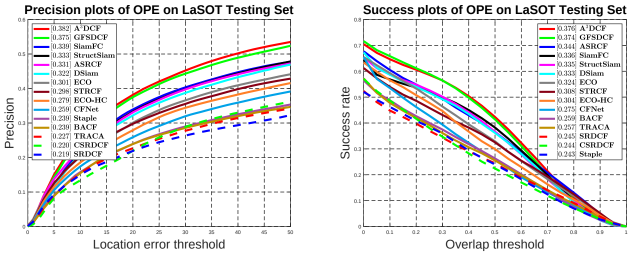


Fig. 9. The experimental results on LaSOT. This figure shows the precision and success plots in terms of the OPE protocol. The DP and AUC score of each tracker is shown in the legend.

obtained by A^{3}DCF on OTB2015 as a function of different values of λ_1 and λ_2. Our A^{3}DCF achieves stable performance (within 2% in terms of AUC) with $\lambda_1, \lambda_2 \in [0.01, 100]$.

Feature Configurations Analysis: Additionally, we employ different feature configurations to test our method and compare the results obtained by the state-of-the-art trackers using the same features. In Table I, the OP scores and tracking speed on OTB2015 of different trackers are provided. Our A^{3}DCF performs better than the other trackers regardless of the features adopted, demonstrating the merits of the proposed formulation.

Ablation Experiments and Analysis: We further perform ablation studies to demonstrate the effectiveness of the key components of the proposed A^{3}DCF tracker, including the attribute-aware attention (AAA) and post processing (PP) modules. Besides, we conduct an experiment to demonstrate the benefits of learning the attention maps for feature channels separately over the learning of a general attention map (GAM) for all the channels jointly. The baseline tracker is the original DCF equipped with the same hand-crafted and deep features, and the same updating rate as our A^{3}DCF. We construct 5 trackers, including BaseLine, BaseLine_AAA, BaseLine_PP, BaseLine_GAM, BaseLine_ALL (A^{3}DCF).

According to the results reported in Table. II, in general, the proposed attribute-aware attention and post processing modules improve the performance of the original DCF (BaseLine). In comparison with the BaseLine tracker, the attribute-aware attention and post processing modules improve the performance by 5.6%/ 6.3% and 6.0%/ 7.2% in terms of AUC and DP respectively. Note that the contributions of the proposed attribute-aware attention and post processing mechanisms to our A^{3}DCF are not linearly additive. The proposed post processing stage enables the learned channel-wise attribute-aware attention maps to be more discriminative by concentrating the energy around the target centre. Besides, thanks to post processing, some irrelevant channels in the filters are removed. In this way, implicit feature selection across the channels is realised and the performance of BaseLine_ALL is significantly improved. Moreover, all the configurations of the BaseLine_GAM tracker are the same as the BaseLine_ALL tracker, except that BaseLine_GAM learns a general attention map for all the channels. As shown in the table, the performance of BaseLine_ALL is better than the performance of BaseLine_GAM. This demonstrates that learning attention for each channel separately is better than learning a global attention map for all the channels jointly. These results verify the effectiveness of the proposed spatial attention mechanism implemented via our adaptive attribute-aware strategy.

Merit Analysis: To intuitively demonstrate the superiority of the proposed attention mechanism, we visualize the attention maps and the corresponding filters of our A^{3}DCF and the other two trackers in Fig. 4. The existing DCF-based algorithms usually adopt a general spatial regularisation pattern for all the feature channels, which is not optimal for individual feature channels. In the figure, SRDCF and GFSDCF employ a fixed attention map for all the feature maps, while our A^{3}DCF learns attention maps for each channel adaptively. It is apparent that the filters trained by our algorithm concentrate more on the target area, thus enhancing the prominence of the most discriminative part of each input feature channel. As a result, the proposed attention mechanism is able to increase the discriminative capability and robustness of the learned model.

Tracking Failures and Demerit Analysis: We present several failed tracking cases in Fig. 5, and analyse the demerits of our A^{3}DCF. From the figure, our method is incapable of dealing with the situations when the target is absent for many frames,

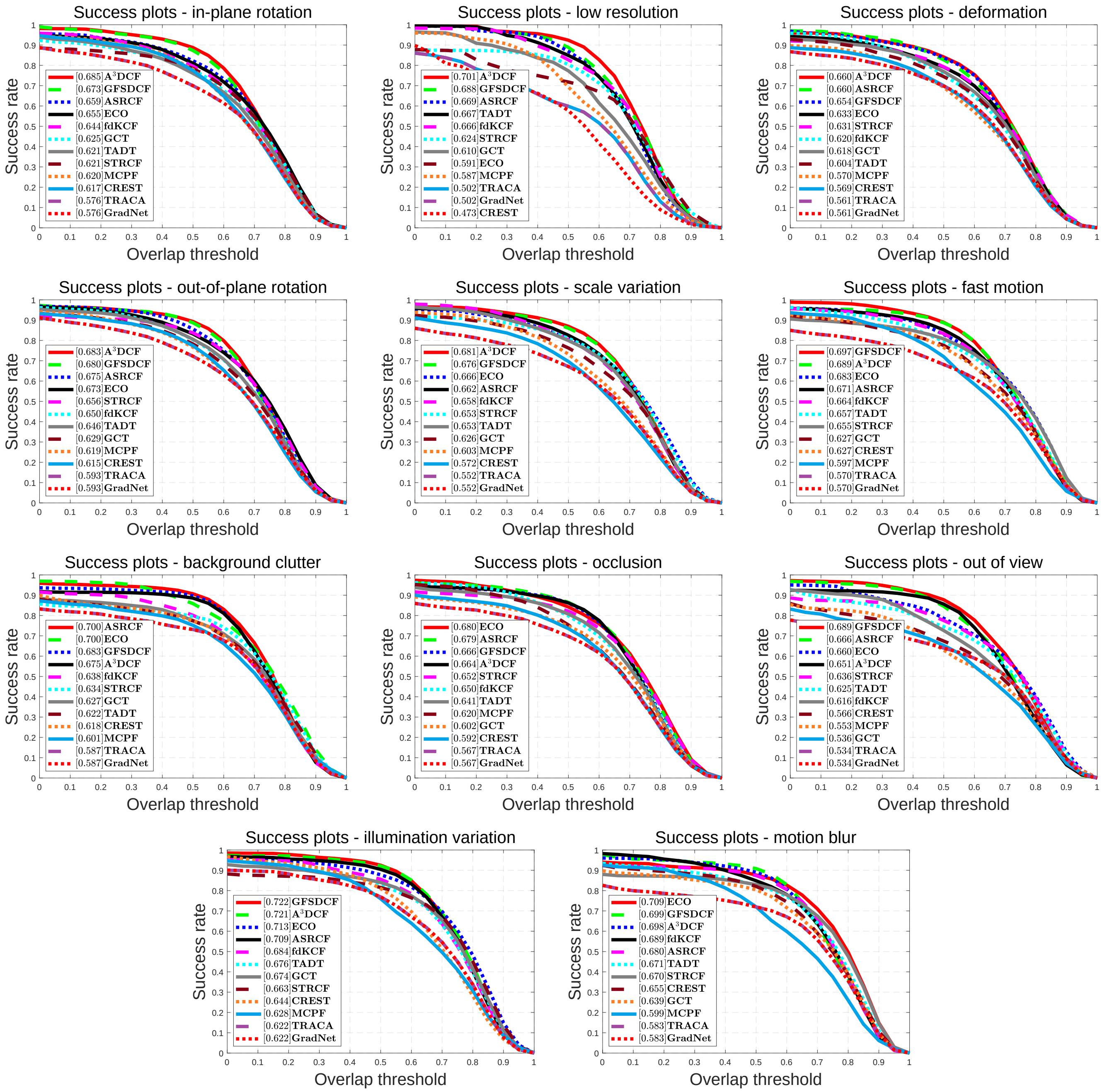


Fig. 10. Success plots of 14 trackers on the OTB2015 dataset in terms of 11 challenging attributes. The AUC scores are shown in the legend.

which causes the tracking failures in the videos such as *Bird1* and *Soccer* of dataset OTB2015. Additionally, DCF-based trackers are unable to estimate the size of the object precisely, when the scale of the target changes rapidly, especially when the aspect ratio of the target varies significantly over short time. As can be seen in the video sequence *Jump*, the aspect ratio of the target changes abruptly. Consequently, the bounding boxes predicted by all the exhibited DCF-based trackers are inaccurate. Besides, due to the strong semantics of the deep CNN features and their dominating influence on target localisation, the interference of similar objects to the target may cause tracking failures. As shown in video sequence *Coupon*, almost all the trackers using deep CNN features misinterpret the distractor as the target, while the trackers that use only shallow hand-crafted features such as BACF and SRDCF are able to track successfully. In summary, although our A^{3}DCF has achieved the state-of-the-art performance, there is still a scope for improvement to rectify the imperfections due to the inherent limitations of the DCF framework, especially compared with the recent trackers that are offline trained with large amounts of video sequences.

D. Comparison with SOTA Methods

VOT2018 has 60 challenging videos. We report the results on VOT2018 in Table III. Our A^{3}DCF achieves the best EAO score, 0.406, outperforming recent advanced DCF trackers.

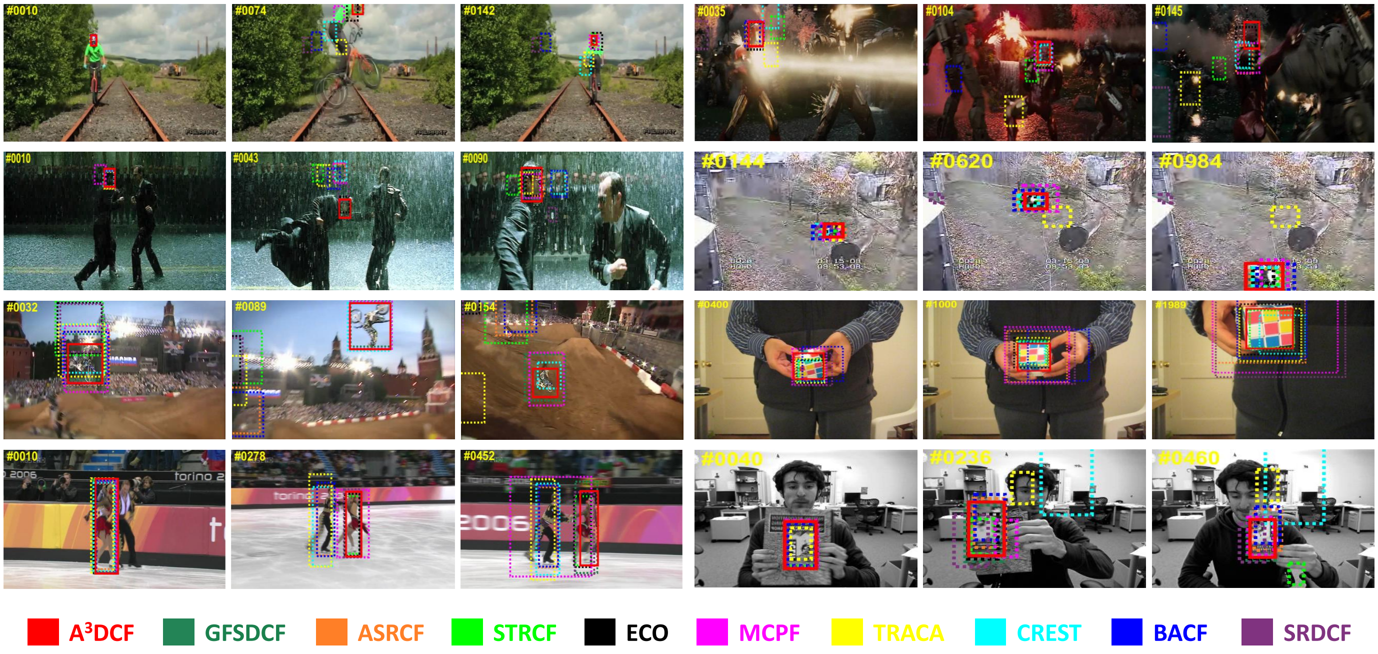


Fig. 11. Illustration of qualitative experimental results on challenging sequences including *Biker, Ironman, Matrix, Panda, MotorRolling, Rubik, Skating2-1 and ClifBar*. The colour bounding boxes are the corresponding results of A^3DCF, GFSDCF, ASRCF, STRCF, ECO, MCPF, TRACA, CREST, BACF and SRDCF respectively.

Besides, A^3DCF achieves advanced performance in terms of Accuracy (0.548) and Robustness (0.162) respectively.

OTB2015 contains 100 video sequences with 11 challenging attributes. We compare our method with other 10 trackers with deep structures/features on OTB2015. The precision and success plots with DP and AUC scores are presented in Fig. 6. Our A^3DCF achieves 92.5% in DP and 69.6% in AUC. Despite the DP score of A^3DCF falling 0.75% behind the best tracker GFSDCF, our method achieves the top ranking AUC score, which is the same as the tracker LADCF.

DTB70 is a benchmark dataset of high diversity consisting of 70 videos captured by drone cameras. The precision and success plots with DP and AUC scores are exhibited in Fig. 7. Our A^3DCF achieves the best DP and AUC scores, namely 78.4% and 51.8% respectively. Compared with the second best tracker, our method achieves gains of 4.3% and 1.5% in terms of DP and AUC respectively.

UAV123 is a dataset composed of 123 challenging video sequences. We report the precision and success plot in Fig. 8. The DP and AUC scores are shown in the figure legend. In terms of DP and AUC scores, our A^3DCF outperforms all the other trackers, achieving 53.5% and 77.8% respectively. Compared with the second best tracker GFSDCF, A^3DCF achieves gains of 1.4% in DP.

LaSOT is a dataset consisting of 1400 sequences in 70 categories with an average length of more than 2500 frames in each sequence. In this paper, we use a simplified official version of LaSOT that contains 280 video sequences, 4 videos of each category. The precision and success plots are presented in Fig. 9. From the plots, A^3DCF performs the best, achieving 38.2% and 37.6% in terms of DP and AUC scores, respectively.

GOT-10K is a large benchmarking dataset for visual object tracking. It has 180 test video sequences in 84 categories. We report the results obtained on GOT-10K in Table IV. $SR_{0.5}$ and $SR_{0.75}$ are the SR metric with a threshold of 0.5 and 0.75 respectively. According to the table, A^3DCF achieves the best AO and $SR_{0.5}$ scores of 0.427 and 0.467 respectively, outperforming the second best tracker CFNet by 14.2% and 16.0%.

TrackingNet is a large-scale tracking dataset that consists of 511 test videos. The evaluation of 7 advanced trackers on TrackingNet is presented in Table V. Our A^3DCF surpasses other trackers on all three evaluation metrics. In detail, A^3DCF outperforms the second best tracker, GFSDCF, with an improvement of 2.6%, 2.5% and 1.4% on Success, Precision and Normalized Precision respectively.

Furthermore, we present the evaluation results on OTB2015, UAV123 and LaSOT in terms of the OP criterion in Table VI. The proposed A^3DCF achieves 89.4%, 67.6% and 42.3% on OTB2015, UAV123 and LaSOT respectively. This top ranking performance achieves gains of 0.4% and 0.9% over the second best tracker GFSDCF on OTB2015 and LaSOT. On UAV123, A^3DCF achieves the second best OP score, falling behind the best tracker GFSDCF by 0.2%. Compared with the state-of-the-art DCF-based tracker GFSDCF, we have realised adaptive feature selection across spatial and channel in another way. Overall, as a DCF-based tracker, the performance of our A^3DCF is competitive.

Attribute-based Evaluation: We report an attribute-based evaluation of 14 trackers on OTB2015. Fig. 10 shows the success plots with AUC scores in terms of 11 video attributes including in-plane rotation, low resolution, deformation, out-of-plane rotation, scale variation, fast motion, background clutter, occlusion, out of view, illumination variation and motion blur. From the AUC scores shown in the legend, our A^{3}DCF outperforms other trackers in 5 attributes. In other challenging attributes, our method is among the top four performers. The merits of our attribute-aware mechanism, which adaptively selects the most discriminative and relevant information from a search region and ignores irrelevant and interfering information, is particularly evident when the target encounters appearance variations in rotation, deformation, and scale.

Qualitative Comparison: To intuitively demonstrate the advantages of our method, in Fig. 11, we provide a qualitative comparison of 10 trackers including A^{3}DCF, GFSDCF, ASRCF, STRCF, ECO, MCPF, TRACA, CREST, BACF and SRDCF respectively, on several video sequences. Although these sequences are challenging with severe appearance variations, our A^{3}DCF performs accurately and steadily. The model trained with the attention mechanism accomplished by the proposed adaptive attribute-aware strategy is able to focus on the discriminative part of the feature input, while paying less attention to irrelevant information. Evidently, our A^{3}DCF can successfully handle these complicated scenarios, delivering desired performance.

V. Conclusion

In this work, to achieve robust visual object tracking, we developed an adaptive attribute-aware scheme for discriminative correlation filter learning. The major innovation of the proposed adaptive attribute-aware tracking method is to enable a channel specific regularisation, which has the ability to identify the discriminative information present in each feature channel. This information is typically linked to different visual attributes of the tracked target. Combined with an appropriate post processing in the filter learning stage, irrelevant channels are suppressed and inconsistent channels are removed by the proposed method. By alleviating the impact of irrelevant information, our model becomes more discriminative and robust in dealing with complex tracking situations. The results of extensive experimental studies on OTB2015, DTB70, UAV123, VOT2018, LaSOT, GOT-10K and TrackingNet, demonstrate the superiority of our A^{3}DCF method over the state-of-the-art trackers.

Acknowledgement

This work was supported in part by the National Key Research and Development Program of China under Grant 2017YFC1601800, the National Natural Science Foundation of China (Grant No.61672265, U1836218, 62020106012), the 111 Project of Ministry of Education of China (Grant No.B12018), the EPSRC Programme Grant (FACER2VM) EP/N007743/1 and the EPSRC/MURI/Dstl project under Grant EP/R018456/1.

References

[1] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista, "Exploiting the circulant structure of tracking-by-detection with kernels," in *European Conference on Computer Vision*, 2012, pp. 702–715.

[2] 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, 2014.

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

[4] M. Danelljan, G. Bhat, F. Shahbaz Khan, and M. Felsberg, "Eco: Efficient convolution operators for tracking," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 6638–6646.

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

[6] A. Lukezic, T. Vojir, L. Cehovin Zajc, J. Matas, and M. Kristan, "Discriminative filter with channel and spatial reliability," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 6309–6318.

[7] 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.

[8] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2005, pp. 01–08.

[9] J. Van De Weijer, C. Schmid, J. Verbeek, and D. Larlus, "Learning color names for real-world applications," *IEEE Transactions on Image Processing*, vol. 18, no. 7, pp. 1512–1523, 2009.

[10] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[11] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proceedings of the International Conference on Learning Representations*, 2015, pp. 01–14.

[12] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[13] 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 Transactions on Image Processing*, vol. 28, no. 11, pp. 5596–5609, 2019.

[14] 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 *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 7950–7960.

[15] T. Xu, Z.-H. Feng, X.-J. Wu, and J. Kittler, "Learning low-rank and sparse discriminative correlation filters for coarse-to-fine visual object tracking," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 01–13, 2019.

[16] 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.

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

[18] 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.

[19] B. Jiang, M. Wang, W. Gan, W. Wu, and J. Yan, "Stm: Spatiotemporal and motion encoding for action recognition," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 2000–2009.

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

[21] S. Li and D.-Y. Yeung, "Visual object tracking for unmanned aerial vehicles: A benchmark and new motion models," in *Thirty-First AAAI Conference on Artificial Intelligence*, 2017, pp. 01–07.

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

[23] M. Kristan, A. Leonardis, J. Matas, M. Felsberg, R. Pflugfelder, L. Cehovin Zajc, T. Vojir, G. Bhat, A. Lukezic, A. Eldesokey *et al.*, "The sixth visual object tracking vot2018 challenge results," in *European Conference on Computer Vision*, 2018, pp. 01–52.

[24] H. Fan, L. Lin, F. Yang, P. Chu, G. Deng, S. Yu, H. Bai, Y. Xu, C. Liao, and H. Ling, "Lasot: A high-quality benchmark for large-scale single object tracking," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2019, pp. 5374–5383.

[25] L. Huang, X. Zhao, and K. Huang, "Got-10k: A large high-diversity benchmark for generic object tracking in the wild," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, p. 01–17, 2019. [Online]. Available: http://dx.doi.org/10.1109/TPAMI.2019.2957464

[26] M. Muller, A. Bibi, S. Giancola, S. Alsubaihi, and B. Ghanem, "Trackingnet: A large-scale dataset and benchmark for object tracking in the wild," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 300–317.

[27] L. Bertinetto, J. Valmadre, J. F. Henriques, A. Vedaldi, and P. H. Torr, "Fully-convolutional siamese networks for object tracking," in *European Conference on Computer Vision*. Springer, 2016, pp. 850–865.

[28] 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.

[29] H. Fan and H. Ling, "Sanet: Structure-aware network for visual tracking," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2017, pp. 42–49.

[30] Q. Guo, W. Feng, C. Zhou, R. Huang, L. Wan, and S. Wang, "Learning dynamic siamese network for visual object tracking," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 1763–1771.

[31] Q. Wang, Z. Teng, J. Xing, J. Gao, W. Hu, and S. Maybank, "Learning attentions: residual attentional siamese network for high performance online visual tracking," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4854–4863.

[32] B. Li, J. Yan, W. Wu, Z. Zhu, and X. Hu, "High performance visual tracking with siamese region proposal network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8971–8980.

[33] B. Li, W. Wu, Q. Wang, F. Zhang, J. Xing, and J. Yan, "Siamrpn++: Evolution of siamese visual tracking with very deep networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2019, pp. 4282–4291.

[34] Z. Zhu, Q. Wang, B. Li, W. Wu, J. Yan, and W. Hu, "Distractor-aware siamese networks for visual object tracking," in *European Conference on Computer Vision*, 2018, pp. 101–117.

[35] H. Fan and H. Ling, "Siamese cascaded region proposal networks for real-time visual tracking," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2019, pp. 7952–7961.

[36] 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.

[37] R. M. Gray *et al.*, "Toeplitz and circulant matrices: A review," *Foundations and Trends in Communications and Information Theory*, vol. 2, no. 3, pp. 155–239, 2006.

[38] N. Wang, J. Shi, D.-Y. Yeung, and J. Jia, "Understanding and diagnosing visual tracking systems," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 3101–3109.

[39] 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.

[40] W. Ruan, J. Chen, Y. Wu, J. Wang, C. Liang, R. Hu, and J. Jiang, "Multi-correlation filters with triangle-structure constraints for object tracking," *IEEE Transactions on Multimedia*, vol. 21, no. 5, pp. 1122–1134, 2018.

[41] K. Chen and W. Tao, "Learning linear regression via single-convolutional layer for visual object tracking," *IEEE Transactions on Multimedia*, vol. 21, no. 1, pp. 86–97, 2018.

[42] B. Huang, T. Xu, S. Jiang, Y. Chen, and Y. Bai, "Robust visual tracking via constrained multi-kernel correlation filters," *IEEE Transactions on Multimedia*, pp. 01–13, 2020.

[43] 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.

[44] Y. Qi, S. Zhang, L. Qin, H. Yao, Q. Huang, J. Lim, and M.-H. Yang, "Hedged deep tracking," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4303–4311.

[45] T. Xu, X.-J. Wu, and J. Kittler, "Non-negative subspace representation learning scheme for correlation filter based tracking," in *International Conference on Pattern Recognition*, 2018, pp. 1888–1893.

[46] M. Kristan, A. Leonardis, J. Matas, M. Felsberg, R. Pflugfelder, L. Cehovin Zajc, T. Vojir, G. Hager, A. Lukezic, A. Eldesokey *et al.*, "The visual object tracking vot2017 challenge results," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 1949–1972.

[47] M. Kristan, J. Matas, A. Leonardis, M. Felsberg, R. Pflugfelder, J.-K. Kamarainen, L. Cehovin Zajc, O. Drbohlav, A. Lukezic, A. Berg *et al.*, "The seventh visual object tracking vot2019 challenge results," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2019, pp. 01–36.

[48] D. Du, P. Zhu, L. Wen, X. Bian, H. Ling, Q. Hu, J. Zheng, T. Peng, X. Wang, Y. Zhang *et al.*, "Visdrone-sot2019: The vision meets drone single object tracking challenge results," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2019, pp. 01–14.

[49] I. T. Young, N. Malik, A. V. Oppenheim, A. S. Willsky, and S. H. Nawab, *Signals and Systems: Solutions Manual*. Prentice-Hall, 1983.

[50] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends® in Machine learning*, vol. 3, no. 1, pp. 1–122, 2011.

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

[52] 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.

[53] H. Nam and B. Han, "Learning multi-domain convolutional neural networks for visual tracking," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4293–4302.

[54] D. Held, S. Thrun, and S. Savarese, "Learning to track at 100 fps with deep regression networks," in *European Conference on Computer Vision*. Springer, 2016, pp. 749–765.

[55] 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.

[56] Y. Song, C. Ma, L. Gong, J. Zhang, R. W. Lau, and M.-H. Yang, "Crest: Convolutional residual learning for visual tracking," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2555–2564.

[57] T. Zhang, C. Xu, and M.-H. Yang, "Learning multi-task correlation particle filters for visual tracking," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 2, pp. 365–378, 2018.

[58] Y. Zhang, L. Wang, J. Qi, D. Wang, M. Feng, and H. Lu, "Structured siamese network for real-time visual tracking," in *European Conference on Computer Vision*, 2018, pp. 351–366.

[59] C. Sun, D. Wang, H. Lu, and M.-H. Yang, "Learning spatial-aware regressions for visual tracking," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8962–8970.

[60] C. Sun, D. Wang, H. Lu, and M.-H. Yang, "Correlation tracking via joint discrimination and reliability learning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 489–497.

[61] J. Gao, T. Zhang, and C. Xu, "Graph convolutional tracking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4649–4659.

[62] P. Li, B. Chen, W. Ouyang, D. Wang, X. Yang, and H. Lu, "Gradnet: Gradient-guided network for visual object tracking," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 6162–6171.

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

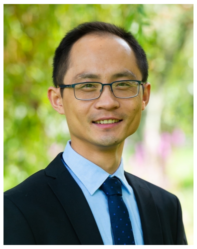
[64] L. Zheng, M. Tang, Y. Chen, J. Wang, and H. Lu, "Fast-deepkcf without boundary effect," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 4020–4029.

Xue-Feng Zhu received the B.Eng. degree in internet of things engineering from Chongqing University of Posts and Telecommunications, Chongqing, China, in 2017. He is a PhD student at the School of Artificial Intelligence and Computer Science, Jiangnan University, Wuxi, China. His research interests include computer vision and machine learning.

Xiao-Jun Wu received the B.Sc. degree in mathematics from Nanjing Normal University, Nanjing, China, in 1991, and the M.S. degree and Ph.D. degree in pattern recognition and intelligent system from the Nanjing University of Science and Technology, Nanjing, in 1996 and 2002, respectively. From 1996 to 2006, he taught at the School of Electronics and Information, Jiangsu University of Science and Technology, where he was promoted to Professor.

He has been with the School of Information Engineering, Jiangnan University since 2006, where he is a Professor of pattern recognition and computational intelligence. He was a Visiting Researcher with the Centre for Vision, Speech, and Signal Processing (CVSSP), University of Surrey, U.K. from 2003 to 2004. He has published over 300 papers in his fields of research. His current research interests include pattern recognition, computer vision, fuzzy systems, neural networks and intelligent systems. He was a Fellow of the International Institute for Software Technology, United Nations University, from 1999 to 2000. He was a recipient of the Most Outstanding Postgraduate Award from the Nanjing University of Science and Technology.

Tianyang Xu received the B.Sc. degree in electronic science and engineering from Nanjing University, Nanjing, China, in 2011. He received the PhD degree at the School of Internet of Things Engineering, Jiangnan University, Wuxi, China, in 2019. He is currently a research fellow at the Centre for Vision, Speech and Signal Processing (CVSSP), University of Surrey, Guildford, United Kingdom. His research interests include visual tracking and deep learning. He has published several scientific papers, including International Conference on Computer Vision, IEEE Transactions on Image Processing, IEEE Transactions on Circuits and Systems for Video Technology, Pattern Recognition etc. He achieved top 1 tracking performance in the VOT2018 public dataset (ECCV2018), VOT2020 RGBT challenge (ECCV2020), and Anti-UAV challenge (CVPR2020).



Zhen-Hua Feng (S'13-M'16) received the Ph.D. degree from the Centre for Vision, Speech and Signal Processing (CVSSP), University of Surrey, Guildford, U.K. in 2016. He is currently a Lecturer in Computer Vision and Machine Learning at the Department of Computer Science, University of Surrey. Before this, he was a senior research fellow at CVSSP. His research interests include pattern recognition, machine learning and computer vision. He has published more than 50 original research papers in top tier conferences and journals such as CVPR, ICCV, IJCAI, TPAMI, IJCV, TCYB, TIP, TIFS, TCSVT, TBIOM, ACM TOMM, Information Sciences, Pattern Recognition, etc. He has received the 2017 European Biometrics Industry Award from the European Association for Biometrics (EAB) and the Best Paper Award for Commercial Applications from the AMDO2018 conference.

Josef Kittler (M'74-LM'12) received the B.A., Ph.D., and D.Sc. degrees from the University of Cambridge, in 1971, 1974, and 1991, respectively. He is a distinguished Professor of Machine Intelligence at the Centre for Vision, Speech and Signal Processing, University of Surrey, Guildford, U.K. He conducts research in biometrics, video and image database retrieval, medical image analysis, and cognitive vision. He published the textbook Pattern Recognition: A Statistical Approach and over 700 scientific papers. His publications have been cited more than 66,000 times (Google Scholar).

He is series editor of Springer Lecture Notes on Computer Science. He currently serves on the Editorial Boards of Pattern Recognition Letters, Pattern Recognition and Artificial Intelligence, Pattern Analysis and Applications. He also served as a member of the Editorial Board of IEEE Transactions on Pattern Analysis and Machine Intelligence during 1982-1985. He served on the Governing Board of the International Association for Pattern Recognition (IAPR) as one of the two British representatives during the period 1982-2005, President of the IAPR during 1994-1996.