Learning Adaptive Discriminative Correlation Filters via Temporal Consistency preserving Spatial Feature Selection for Robust Visual Tracking

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

***Abstract*—With efficient appearance learning models, Discriminative Correlation Filter (DCF) has been proven to be very successful in recent video object tracking benchmarks and competitions. However, the existing DCF paradigm suffers from two major problems, *i.e.* spatial boundary effect and temporal filter degeneration. To mitigate these challenges, we propose a new DCF-based tracking method. The key innovations of the proposed method include adaptive spatial feature selection and temporal consistent constraints, with which the new tracker enables joint spatio-temporal filter learning in a lower dimensional discriminative manifold. More specifically, we apply structured sparsity constraints to multi-channel filers. Consequently, the process of learning spatial filters can be approximated by the lasso regularisation. To encourage temporal consistency, the filter model is restricted to lie around its historical value and updated locally to preserve the global structure in the manifold. Last, a unified optimisation framework is proposed to jointly select temporal consistency preserving spatial features and learn discriminative filters with the augmented Lagrangian method. Qualitative and quantitative evaluations have been conducted on a number of well-known benchmarking datasets such as OTB2013, OTB50, OTB100, Temple-Colour and UAV123. The experimental results demonstrate the superiority of the proposed method over the state-of-the-art approaches.**

***Index Terms*—Visual tracking, correlation filter, feature selection, temporal consistency**

I. Introduction

Object tracking is an important research topic in computer vision, image understanding and pattern recognition. Given the initial state (centre location and scale) of a target in a video sequence, the aim of visual tracking is to automatically obtain the states of the object in the subsequent video frames. With the rapid development of the research area during the past decade, a variety of tracking algorithms have been proposed and shown to deliver promising results [1], [2], [3], [4], [5], [6]. The advances opened a wide spectrum of applications in practical scenarios, such as intelligent surveillance, robot perception, medical image processing and other visual intelligence systems. Despite the great success, robust and real-time visual tracking remains a challenging task, especially in unconstrained scenarios in the presence of illumination variation, changing background, camera jitter, blur, non-rigid deformation and partial occlusion.

Effective and reliable modelling of the appearance for a target and its surroundings is one of the most important keys to robust visual tracking under unconstrained scenarios. To this end, both generative and discriminative methods have been extensively studied. A generative method usually uses a parametric model, *e.g.* the probability density function, to describe target appearance,. The most plausible candidate is selected as the tracking result by maximising its similarity to a generative model or minimising its reconstruction error. In contrast, a discriminative method exploits the background information to improve the representation capacity of an appearance model. Discriminative methods consider a tracking task as classification or regression problem hence directly infer the output of a candidate by estimating the conditional probability distribution of labels for given inputs. The optimal candidate with the highest response/score is selected as the tracking result. Recently, considering the joint circular structure of sliding candidates, Discriminative Correlation Filter (DCF-) based tracking methods have achieved outstanding performance in many challenging benchmarks and competitions [5], [7], [8]. The main advantages of DCF include the use of circulant structure of original samples and the formulation of the tracking problem as ridge regression. Besides, DCF employs the Fast Fourier Transform (FFT) to accelerate the computation of closed-form solutions for all circularly shifted candidates in the frequency domain.

Despite the effectiveness and efficiency of DCF, its performance is affected by two problems: spatial boundary effect and temporal filter degeneration. As the circular shift of an image patch results in discontinuity around original boundaries. Such a boundary effect causes spatial distortion hence decreases the quality of training examples. In addition, filter degeneration reduces the effectiveness for lack of integrating historical appearance information. To eliminate these problems, we advocate the joint use of temporal information and spatial regularisation for adaptive target appearance modelling in a DCF-based framework.

The key to solving the first issue, *i.e.* spatial boundary effect, is to enhance the spatial appearance model learning framework. DCF-based tracking methods aim to optimise the correlation filter so that it is able to efficiently locate a target from surroundings. To use the cyclic structure [9], the DCF paradigm typically represents the target by extended

T. Xu is with the School of Internet of Things Engineering, 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@surrey.ac.uk)

Z.-H. Feng and J. Kittler are with the Centre for Vision, Speech and Signal Processing, University of Surrey, Guildford, GU2 7XH, UK. (e-mail: {z.feng, j.kittler}@surrey.ac.uk)

X.-J. Wu is with the School of Internet of Things Engineering, Jiangnan University, Wuxi, P.R. China. (e-mail: wu_xiaojun@jiangnan.edu.cn)

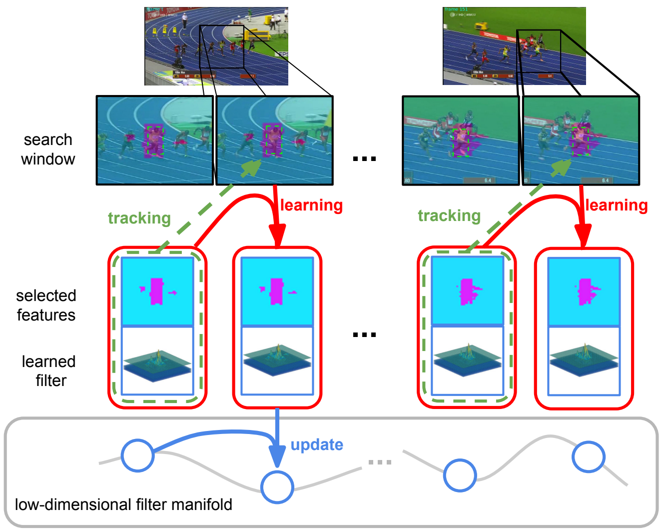

Fig. 1. Illustration of the proposed temporal consistency preserving spatial feature selection embedded tracking model in sequence *Bolt*. Only selected spatial features of filters are active in the tracking stage to achieve accurate results (selected spatial features are annotated by purple). Both sparse constraint and temporal consistency are jointly considered in the learning stage. Filters are robustly updated to form a low-dimensional discriminative manifold, preserving the smoothness of filters over time.

padding from the target box to the entire search region. As the correlation operator is defined by the inner product, features from different locations in the search region contribute to the final response. A large search region contributes more clutter from background while a small search region may lead to drift and target underdetection. To address this issue, spatial regularisation has been explored [10], [11], [12], [13]. However, existing spatial regularisation methods only regularise the filter with simple pre-defined constraints without considering the diversity and redundancy of the entire feature input.

In this work, a more informative spatial appearance leaning formulation is proposed to perform adaptive spatial feature selection and filter learning jointly. Spatial features are regularised and selected with the lasso regularisation that adaptively preserves the structure of the discriminative manifold which encapsulates the variation of a target and its background. With both, the centre and surrounding regions being exploited to support discrimination, spatial features with positive reference values are selected, including those from the background with similar and stable motions with respect to that of the target. As illustrated in Fig. 1, our feature selection strategy enables compress sensing by adaptively selecting an optimal discriminative spatial mask, avoiding boundary distortion and restraining the impact of the distractive information inherent in the original representations.

For the second issue, *i.e.* temporal filter degradation, traditional DCF methods are vulnerable to the vagaries of extensive appearance variations of a target caused by the use of unstable temporal updating models for filters. This instability is precipitated by two factors: i) single frame independent learning, and ii) fixed-rate model update. Although numerical visual features are integrated with appearance models, the existing tracking methods are not able to represent and store dynamic appearance of targets as well as surroundings. In contrast, a single frame learning scheme is used in the traditional DCF paradigm to accelerate the optimisation process.

On the other hand, a fixed-rate model update of the moving average form, *i.e.* $\theta = (1-\alpha)\theta + \alpha\theta'$, ignores the variation between different frames hence reduces the adaptability to appearance variations. In this work, temporal consistency is forced to interact with spatial feature selection in appearance modelling to generate a manifold structure capturing the global dynamic appearance information (as shown in Fig. 1). In particular, we propose an online temporal consistency preserving model to construct a generative manifold space that identifies an effective feature configuration and filtering function. Such an online adaptation strategy is capable of preventing filter degeneration and enhances temporal smoothness.

Last, we present a unified optimisation framework that can efficiently perform spatial feature selection and discriminative filter learning. To be more specific, the augmented Lagrangian method is used to optimise the variables in an iterative fashion. We perform feature selection in the spatial domain and filter learning in the frequency domain. The transformation between different domains is realised by FFT. It should be highlighted that, for feature extraction, hand-crafted and deep neural network features can be used in conjunction with our method. The experimental results obtained on OTB2013 [3], OTB50 [5], OTB100 [5], Template-Colour [14], and UAV123 [15] benchmarks demonstrate the efficiency and superiority of the proposed method in learning adaptive discriminative correlation filters (LADCF) via temporal consistency preserving spatial feature selection, over the state-of-the-art approaches. The main contributions of our LADCF method include:

- A new appearance model construction technique with adaptive spatial feature selection. In our experiments, only about 5% hand-crafted and 20% deep features are selected for filter learning, yet achieving better performance. The proposed method also addresses the issues of spatial boundary effect and background clutter.
- A novel methodology for designing a low-dimensional discriminative manifold space by exploiting temporal consistency, which realises reliable and flexible temporal information compression, alleviating filter degeneration and preserving appearance diversity.
- A unified optimisation framework is proposed to achieve efficient filter learning and feature selection using the augmented Lagrangian method.

The paper is organised as follows. Section II discusses the prior work relevant to our tracking framework. In Section III, we introduce the classical DCF tracking formulation. The proposed temporal consistency preserving spatial feature selection method is introduced in Section IV. The experimental results are reported and analysed in Section V. Conclusion is given in Section VI.

II. Related work

For a comprehensive review of existing tracking methods the reader can refer to recent surveys [1], [2], [3], [4], [5], [6]. In this section, we focus on the most relevant techniques that define the baseline for the research presented in this paper. The review includes basic generative and discriminative tracking methods, DCF-based tracking methods and embedded feature selection methods.

Generative methods: Since the proposal of the Lucas-Kanade algorithm [16], [17], generative methods have become widely used techniques in many computer vision tasks, including visual tracking. Subsequently, mean-shift [18] was proposed for visual tracking with iterative histogram matching. Accordingly, the spatial appearance information is summarised in the form of histogram for a target in its candidate locations, and the tracking task is achieved by minimising the Bhattacharyya distance. To improve the robustness of visual tracking with generative models, Adam *et. al.* proposed fragments-based tracking by matching an ensemble of patches [19], in which the robustness of the use of histograms is improved by exploiting exhaustive appearance information conveyed by collaborative fragments. Later, the adoption of subspace-based tracking approaches offered a better explanation for the appearance model and provided an incremental learning method for appearance model update [20]. More recently, trackers based on sparse and low-rank representations also achieved robust results by considering structural constraints and manifold information [21], [22], [23], [24], [25]. Though generative models achieve considerable success in constrained scenarios by faithfully modelling the target appearance, they are vulnerable to extensive appearance variations and unpredictable movement. Thus, more attention has been paid to discriminative approaches.

Discriminative methods: Unlike generative methods that focus on exploring the similarities between candidates and a target model, discriminative methods aim to construct a classifier or regressor to distinguish the target from its background. To this end, an on-line boosting tracker [26] was proposed by Grabner *et. al.* by fusing multiple weak classifiers. In addition, a multi-instance learning tracker [27] was proposed to learn a discriminative classifier from the extracted positive and negative samples. To address temporal appearance variations of a target, the tracking-learning-detection method [28] was proposed to handle short-term occlusion using a double tracking strategy. In order to exploit the labelled and unlabelled data more effectively, Struck [29] proposed to train a structural supervised classifier for visual tracking.

More recently, taking the advantage of offline-learning and online-tracking deep models, end-to-end learning approaches have been introduced to visual tracking with GPU acceleration. In this category, GOTURN [30] uses consecutive sample pairs for regression model training and achieves efficient tracking results. For robust comparison, SINT [31] considers the tracking task as a verification matching problem solved by a Siamese network. The same strategy was applied in SiamFC [32] and CFNet [33]. SiamFC established a fully-convolutional Siamese network by cross-correlating deep features, while CFNet learned correlation filter as a differentiable layer in deep architecture and achieved good results on standard benchmarks [5]. In addition, residual learning for spatial and temporal appearance was proposed by CREST [34], in which feature extraction, response calculation and model updating were fused in a single layer of a Convolutional Neural Network (CNN).

DCF based tracking methods have attracted wide attention recently. DCF employs circulant structure to solve a ridge regression problem in the frequency domain. Based on the proposals of Normalised Cross Correlation (NCC) [35] and Minimum Output Sum of Squared Error (MOSSE) filter [36], Henriques *et al.* improved MOSSE by introducing circulant structure [9], which enables efficient calculation of filter learning with element-wise operations. Other improvements in DCF focus on exploring robust feature representations, scale detection and spatial regularisation.

For feature representation, contextual feature information was exploited in [37] to achieve spatio-temporal learning. Colour names [38] were fused with the correlation filter framework by Danelljan *et. al.* [39] to better represent an input. Staple [40] employed colour histograms from foreground and background to generate a response map, which improves the reliability of the final response. Later, CNNs have been widely used to provide better feature representation, as in deepSRDCF [11].

For scale detection, SAMF [41] and DSST [42] were proposed to handle scale variations by performing scale selection in a scale pool after the tracking stage. On the other hand, fDSST [43] proposed to perform scale detection in the tracking stage. This improves the efficiency by joint scale and location estimation.

For spatial regularisation, a pre-defined filter weighting strategy was proposed in SRDCF [11], concentrating the filter energy in the central region. In CSRDCF [12], the filter was equipped with a colour based segmentation mask, with which only the discriminative target region is activated. A similar approach was employed in CFLB [10] and BACF [13], force the parameters corresponding to background to be exactly zero. In addition, DCF-based tracking methods have also been extended to support long-term memory [44], multi-kernel method [45], structural constraints [46], support vector representation [47], sparse representation [48] and enhanced robustness [49], [50], [51], [52], [53], [54]. Furthermore, adaptive decontamination of the training set [55] was proposed to achieve adaptive multi-frame learning in the DCF paradigm, which improves the generalisation performance. Danelljan *et al.* proposed sub-grid tracking by learning continuous convolution operators (C-COT) [56]. Efficient Convolution Operators (ECO) [57] were proposed to achieve a light-weight version of C-COT with a generative sample space and dimension reduction mechanism.

Embedded feature selection methods: Feature selection has been proven to be very effective in processing high-dimensional data. Thus it has been a widely studied topic in many pattern recognition and computer vision applications, *e.g.* image classification [58] and image compression [59]. Embedded feature selection methods are regularisation models with an optimisation objective that simultaneously minimises classification (regression) errors and forces the variables to satisfy specific prior properties, *e.g.* lying in a ℓ_p ball around 0 [60], [61]. These approaches enhance model generalisation by reducing over-fitting with the advantage of interpretability.

Basic DCF based trackers [9], [62] utilise ℓ_2-norm to regularise the coefficients. Such a penalty shrinks all the filter coefficients towards zero, but it does not set any of them exactly to zero. A weighted ℓ_2-norm regularisation was exploited in SRDCF, C-COT and ECO, in which the coefficients in the background shrink more than those in the target region. Besides, the exact zero-one feature selection method has been applied in the DCF paradigm by CFLB, CSRDCF and BACF. Only the coefficients in the target region are activated in CFLB and BACF. CSRDCF utilised a two-stage feature selection regression method to first pre-define the selected region by discriminative colour information and then train the mask constrained filters. In addition to the ℓ_2-norm regularisation, lasso regularisation, based on the ℓ_1-norm, is a popular method to achieve embedded feature selection in classification (regression). It has been widely explored in sparse representation based tracking approaches [63], [23], [25]. The use of lasso regularisation is able to perform sparse estimation with only a small number of features activated. However, the basic lasso regularisation methods ignore the structure information as they assume the variables are independent. To this end, structured feature selection methods [64], [65] have been proposed to integrate group knowledge, improving the accuracy and robustness.

Our method employs the DCF paradigm to formulate the tracking problem, with temporal consistency preserving spatial feature selection embedded appearance model. The proposed DCF learning scheme and spatial feature selection achieve efficient discriminative filter learning and enhanced generalisation. Adaptive spatial features are activated by lasso regularisation, that can be efficiently optimised by iterative threshold shrinkage. Last, dynamic appearance is modelled with temporal consistency to enhance the robustness of the selected features over time.

III. Tracking Formulation

Consider a base sample, an $n \times n$ image patch $\boldsymbol{x} \in \mathbb{R}^{n^2 \times 1}$. The circulant matrix for this sample is generated by collecting its full cyclic shifts, $\boldsymbol{X}^\top = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_{n^2}]^\top \in \mathbb{R}^{n^2 \times n^2}$ with the corresponding gaussian shaped regression labels $\boldsymbol{y} = [y_1, y_2, \ldots, y_{n^2}]$ [62]. Our goal is to learn a discriminative function $f(\boldsymbol{x}_i; \boldsymbol{\theta}) = \boldsymbol{\theta}^\top \boldsymbol{x}_i$ to distinguish the target from background, where $\boldsymbol{\theta}$ denotes the target model in the form of DCF. Such a data augmentation method is widely used in the DCF paradigm with the calculation convenience in the frequency domain:

$$f(\boldsymbol{X}; \boldsymbol{\theta}) = \boldsymbol{\theta}^\top \boldsymbol{X} = \boldsymbol{\theta} \circledast \boldsymbol{x} = \mathcal{F}^{-1}\left(\hat{\boldsymbol{\theta}} \odot \hat{\boldsymbol{x}}^*\right), \tag{1}$$

where $\mathcal{F}^{-1}$ denotes the Inverse Discrete Fourier Transform (IDFT), and $\hat{\boldsymbol{x}}^*$ is the complex conjugate of $\hat{\boldsymbol{x}}$ in the frequency domain. $\hat{\boldsymbol{x}}$ is the Fourier representation of $\boldsymbol{x}$, *i.e.* $\mathcal{F}(\boldsymbol{x}) = \hat{\boldsymbol{x}}$. $\circledast$ denotes the circular convolution operator and $\odot$ denotes the element-wise multiplication operator. It should be noted that the computational complexity of $f(\boldsymbol{X})$ is significantly decreased from $\mathcal{O}(n^4)$ to $\mathcal{O}(n^2 \log n)$ with the help of convolution theorems. We use the following tracking-learning-updating framework to formulate our tracking method.

Tracking: Given the model parameter $\boldsymbol{\theta}_{\text{model}}$ estimated from previous frames, we aim to find the optimal candidate that maximises the discriminative function in the current frame $\boldsymbol{I}$:

$$\boldsymbol{x}_* = \arg\max_{\boldsymbol{x}_i} f(\boldsymbol{x}_i; \boldsymbol{\theta}_{\text{model}}), \tag{2}$$

where the candidates are generated by the circulant structure of a base sample $\boldsymbol{x}$, which is the image patch centred around the tracking result in the previous frame. Consequently, the results can be efficiently calculated in the frequency domain.
Learning: After the tracking stage, a new model is trained by minimising the regularised loss function:

$$\boldsymbol{\theta}_* = \arg\min_{\boldsymbol{\theta}} \mathcal{E}(\boldsymbol{\theta}, \boldsymbol{\mathcal{D}}) + \mathcal{R}(\boldsymbol{\theta}), \tag{3}$$

where $\mathcal{E}()$ is the objective and $\mathcal{R}()$ is the regularisation term. $\boldsymbol{\mathcal{D}} = (\boldsymbol{X}, \boldsymbol{y})$ represents the labelled training samples generated by the circulant matrix with the base sample $\boldsymbol{x}$ centred around the tracking result in the current frame. In the traditional DCF paradigm, the quadratic loss and ℓ_2-norm penalty are used in the learning stage to form a ridge regression problem, *i.e.* $\mathcal{E}(\boldsymbol{\theta}, \boldsymbol{\mathcal{D}}) = \|\boldsymbol{\theta}^\top \boldsymbol{X} - \boldsymbol{y}\|_2^2$ and $\mathcal{R}(\boldsymbol{\theta}) = \|\boldsymbol{\theta}\|_2^2$ [62].
Updating: Consider the diversity of target appearance, an incremental model update strategy [9] is used in DCF:

$$\boldsymbol{\theta}_{\text{model}} = (1-\alpha)\boldsymbol{\theta}_{\text{model}} + \alpha\boldsymbol{\theta}_*, \tag{4}$$

where $\alpha \in [0, 1]$ is the learning rate for the purpose of trade-off between the current and historical information.

IV. The Proposed Algorithm

In this section, we first present our temporal consistency preserving spatial feature selection method for appearance modelling using single-channel signals. Then we extend the method to multi-channel features. The optimisation process is designed using the alternating direction method of multipliers (ADMM). Last, we depict the proposed LADCF tracking algorithm in more details.

A. Temporal Consistency Preserving Spatial Feature Selection Model

Our feature selection process aims at selecting several specific elements in the filter $\boldsymbol{\theta} \in \mathbb{R}^{n^2 \times 1}$ to preserve discriminative and representative information. It is formulated as:

$$\boldsymbol{\theta}_{\boldsymbol{\phi}} = \mathbf{diag}(\boldsymbol{\phi})\,\boldsymbol{\theta}, \tag{5}$$

where $\mathbf{diag}(\boldsymbol{\phi})$ is the diagonal matrix generated from the indicator vector of selected features $\boldsymbol{\phi}$. Unlike traditional dimensionality reduction methods, such as the principal component analysis (PCA) and locally linear embedding (LLE), the indicator vector $\boldsymbol{\phi}$ enables dimensionality reduction as well as spatial structure preservation. The elements in $\boldsymbol{\phi}$ are either 0 or 1, disabling or enabling the corresponding element. Our feature selection enhanced filter design simultaneously selects spatial features and learns discriminative filters. It should be noted that the selected spatial features are implicitly shared by input $\boldsymbol{x}$, $\boldsymbol{\theta}_{\boldsymbol{\phi}}^\top \boldsymbol{x} = \boldsymbol{\theta}_{\boldsymbol{\phi}}^\top \boldsymbol{x}_{\boldsymbol{\phi}}$, which reveals that only relevant features are activated for each training sample, forming a low-dimensional and compact feature representation. Thus, the spatial feature selection embedded learning stage can be formulated as:

$$\begin{aligned} &\underset{\boldsymbol{\theta},\boldsymbol{\phi}}{\operatorname{argmin}} \left\|\boldsymbol{\theta} \circledast \boldsymbol{x} - \boldsymbol{y}\right\|_2^2 + \lambda_1 \left\|\boldsymbol{\phi}\right\|_0 \\ &s.t.\ \boldsymbol{\theta} = \boldsymbol{\theta}_{\boldsymbol{\phi}} = \mathbf{diag}\left(\boldsymbol{\phi}\right)\boldsymbol{\theta}, \end{aligned} \tag{6}$$

where the indicator vector $\boldsymbol{\phi}$ can potentially be represented by $\boldsymbol{\theta}$ and $\left\|\boldsymbol{\phi}\right\|_0 = \left\|\boldsymbol{\theta}\right\|_0$. As ℓ_0-norm is non-convex, its convex envelope ℓ_1-norm is widely used to approximate the sparsity [66]. On the other hand, in order to emphasise temporal consistency during tracking, the indicator vectors from successive video frames are assumed to be on a low-dimensional manifold. More specifically, we propose to restrict our estimate to lie in a ℓ_0-norm ball around the current template, *i.e.* $\left\|\boldsymbol{\theta} - \boldsymbol{\theta}_{\text{model}}\right\|_0 < t$. Therefore, such temporal consistency enables that the selected spatial features are changed locally, preserving the discriminative layout. We formulate the temporal consistency preserving spatial feature selection model via ℓ_1-norm relaxation as:

$$\arg\min_{\boldsymbol{\theta}} \left\|\boldsymbol{\theta} \circledast \boldsymbol{x} - \boldsymbol{y}\right\|_2^2 + \lambda_1 \left\|\boldsymbol{\theta}\right\|_1 + \lambda_2 \left\|\boldsymbol{\theta} - \boldsymbol{\theta}_{\text{model}}\right\|_1, \tag{7}$$

where λ_1 and λ_2 are tuning parameters and $\lambda_1 << \lambda_2$. As shown in Fig. 2, an intuitive explanation is that we perform a stronger constraint on the temporal consistency term than that on the spatial feature selection term. In addition, the temporal consistency term promotes the sparsity of $\boldsymbol{\theta}$ by enhancing the similarity between the estimate $\boldsymbol{\theta}$ and the template $\boldsymbol{\theta}_{\text{model}}$. Note that as $\left\|\boldsymbol{\theta} - \boldsymbol{\theta}_{\text{model}}\right\|_2 \leq \left\|\boldsymbol{\theta} - \boldsymbol{\theta}_{\text{model}}\right\|_1 \leq n \left\|\boldsymbol{\theta} - \boldsymbol{\theta}_{\text{model}}\right\|_2$ and $\lambda_1 << \lambda_2$, we perform ℓ_2-norm relaxation for the temporal consistency term to further simplify the objective as:

$$\arg\min_{\boldsymbol{\theta}} \left\|\boldsymbol{\theta} \circledast \boldsymbol{x} - \boldsymbol{y}\right\|_2^2 + \lambda_1 \left\|\boldsymbol{\theta}\right\|_1 + \lambda_2 \left\|\boldsymbol{\theta} - \boldsymbol{\theta}_{\text{model}}\right\|_2^2, \tag{8}$$

where the spatial features $\boldsymbol{\phi}$ are selected by lasso regularisation controlled by λ_1, but the number of non-zero entries in $\boldsymbol{\theta}$ cannot be guaranteed. We define this number by forcing $\left\|\boldsymbol{\phi}\right\|_0 = M$. The filter $\boldsymbol{\theta}$ optimised from this objective function can adaptively highlight the spatial configuration as well as preserve discriminative information. Specific spatial features corresponding to target and background region can be simultaneously activated to form a robust pattern. In addition, according to the fact that discriminative learning depends heavily on the reliability of supervised information, the quality of the training samples is significant to tracking performance. Therefore, we promote temporal consistency by imposing smooth variation between consecutive frames with the help of the filter template $\boldsymbol{\theta}_{\text{model}}$ in Equ. 8. In this way the diversity of dynamic and static appearance can be extracted and preserved by our temporal consistency preserving spatial feature selection enhanced appearance model learning.

B. Generalising to Multi-channel Feature Representations

Multi-channel feature representations, *e.g.* HOG [62], Colour-Names [38] and deep features [57], have been widely used in object tracking. To this end, the traditional DCF paradigm [62] explores each channel independently with equal weight. In contrast, CSRDCF [12] proposes a weighting strategy for each channel at the decision stage. ECO [57] applies a projection matrix to realise channel compression.

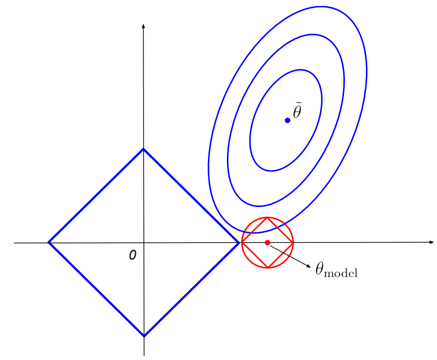

Fig. 2. Illustration of ℓ_2-norm relaxation for the temporal consistency constraint. $\bar{\boldsymbol{\theta}}$ is the optimal least-square solution and $\boldsymbol{\theta}_{\text{model}}$ is the template point. As $\boldsymbol{\theta}_{\text{model}}$ is sparse, its ℓ_2-norm ball shares the same property, guiding the estimate to lie on a low-dimensional manifold.

In our approach we aim to characterise our appearance model from single-channel signals to multi-channel feature representations. As multi-channel features share the same spatial layout, $\boldsymbol{\phi}$, a group structure is considered in our multi-channel learning model [64], [61]. We denote the multi-channel input as $\boldsymbol{\mathcal{X}} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_L\}$ and the corresponding filters as $\boldsymbol{\theta} = \{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_L\}$, where L is the number of channels. For the ith channel, $\boldsymbol{x}_i = \left[x_i^1, x_i^2, \ldots, x_i^{D^2}\right]^\top \in \mathbb{R}^{D^2 \times 1}$, the spatial size of the feature map is $D \times D$, D^2 is the number of spatial features, and x_i^j is the element corresponding to the jth spatial feature in the ith channel. As the index vector $\boldsymbol{\phi}$ is applied in all channels to configure a global spatial layout, the objective function in Equ. 8 can be extended to multi-channel signals with structured sparsity:

$$\begin{aligned} \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^{L} \left\|\boldsymbol{\theta}_i \circledast \boldsymbol{x}_i - \boldsymbol{y}\right\|_2^2 + \lambda_1 \left\|\sqrt{\sum_{i=1}^{L} \boldsymbol{\theta}_i \odot \boldsymbol{\theta}_i}\right\|_1 \\ + \lambda_2 \sum_{i=1}^{L} \left\|\boldsymbol{\theta}_i - \boldsymbol{\theta}_{\text{model}\ i}\right\|_2^2, \end{aligned} \tag{9}$$

where the structured spatial feature selection term calculates ℓ-2 norm values for each spatial location and then performs ℓ-1 norm calculation to realise joint sparsity. Such group sparsity enables robust feature selection that reflects the joint contribution of feature maps from all the channels. In addition, different input channels are potentially weighted by structured sparsity, unifying the entire input during feature selection adaptively.

C. Optimisation

In order to optimise Equ. 9, we introduce slack variables to construct the following objective based on convex optimisation [67]:

$$\begin{aligned} &\underset{\boldsymbol{\theta},\boldsymbol{\theta}'}{arg\min} \sum_{i=1}^{L} \|\boldsymbol{\theta}_i \circledast \boldsymbol{x}_i - \boldsymbol{y}\|_2^2 + \lambda_1 \left\| \sqrt{\sum_{i=1}^{L} \boldsymbol{\theta}'_i \odot \boldsymbol{\theta}'_i} \right\|_1 \\ &\quad + \lambda_2 \sum_{i=1}^{L} \|\boldsymbol{\theta}_i - \boldsymbol{\theta}_{\text{model } i}\|_2^2 , \\ &\quad s.t.\ \boldsymbol{\theta} = \boldsymbol{\theta}'. \end{aligned} \tag{10}$$

Exploiting augmented Lagrange multipliers to combine the equality constraint into the criterion function, our objective can be formulated to minimise the following Lagrange function:

$$\begin{aligned} \mathcal{L} = &\sum_{i=1}^{L} \|\boldsymbol{\theta}_i \circledast \boldsymbol{x}_i - \boldsymbol{y}\|_2^2 + \lambda_1 \left\| \sqrt{\sum_{i=1}^{L} \boldsymbol{\theta}'_i \odot \boldsymbol{\theta}'_i} \right\|_1 \\ &+ \lambda_2 \sum_{i=1}^{L} \|\boldsymbol{\theta}_i - \boldsymbol{\theta}_{\text{model } i}\|_2^2 + \frac{\mu}{2} \sum_{i=1}^{L} \left\| \boldsymbol{\theta}_i - \boldsymbol{\theta}'_i + \frac{\boldsymbol{\eta}_i}{\mu} \right\|_2^2 , \end{aligned} \tag{11}$$

where $\boldsymbol{\mathcal{H}} = \{\boldsymbol{\eta}_1, \boldsymbol{\eta}_2, \dots, \boldsymbol{\eta}_L\}$ are the Lagrange multipliers and $\mu > 0$ is the corresponding penalty parameter. As $\mathcal{L}$ is convex, ADMM is exploited iteratively to optimise the following sub-problems with guaranteed convergence [68]:

$$\begin{cases} \boldsymbol{\theta} = \arg\min_{\boldsymbol{\theta}} \mathcal{L}\left(\boldsymbol{\theta}', \boldsymbol{\mathcal{H}}, \mu\right) \\ \boldsymbol{\theta}' = \arg\min_{\boldsymbol{\theta}'} \mathcal{L}\left(\boldsymbol{\theta}, \boldsymbol{\mathcal{H}}, \mu\right) \\ \boldsymbol{\mathcal{H}} = \arg\min_{\boldsymbol{\mathcal{H}}} \mathcal{L}\left(\boldsymbol{\theta}, \boldsymbol{\theta}', \mu\right) \end{cases} . \tag{12}$$

Updating $\boldsymbol{\theta}$. Given $\boldsymbol{\theta}'$, $\boldsymbol{\mathcal{H}}$ and μ, optimising $\boldsymbol{\theta}$ is similar to the DCF learning scheme. We utilise the circulant structure and Parseval's formula to transform it into the frequency domain [62], which requires solving the following convex optimisation problem for each channel independently:

$$\begin{aligned} \hat{\boldsymbol{\theta}}_i = &\arg\min_{\hat{\boldsymbol{\theta}}_i} \left\| \hat{\boldsymbol{\theta}}_i^* \odot \hat{\boldsymbol{x}}_i - \hat{\boldsymbol{y}} \right\|_2^2 \\ &+ \lambda_2 \left\| \hat{\boldsymbol{\theta}}_i - \hat{\boldsymbol{\theta}}_{\text{model } i} \right\|_2^2 + \frac{\mu}{2} \left\| \hat{\boldsymbol{\theta}}_i - \hat{\boldsymbol{\theta}}'_i + \frac{\hat{\boldsymbol{\eta}}_i}{\mu} \right\|_2^2 \end{aligned} \tag{13}$$

that admits a closed-form optimal solution for $\hat{\boldsymbol{\theta}}_i$:

$$\hat{\boldsymbol{\theta}}_i = \frac{\hat{\boldsymbol{x}}_i \odot \hat{\boldsymbol{y}}^* + \lambda_2 \hat{\boldsymbol{\theta}}_{\text{model } i} + \frac{1}{2}\mu \hat{\boldsymbol{\theta}}'_i - \frac{1}{2}\hat{\boldsymbol{\eta}}_i}{\hat{\boldsymbol{x}}_i \odot \hat{\boldsymbol{x}}_i^* + \lambda_2 + \frac{1}{2}\mu}. \tag{14}$$

The division in Equ. 14 is an element-wise operation, as element $\hat{\theta}_i^j$ is only determined by $\hat{x}_i^j$, $\hat{y}^j$, $\hat{\theta}_{\text{model } i}^j$, $\hat{\theta}_i'^j$ and $\hat{\eta}_i^j$ with parameters λ_2 and μ.

Updating $\boldsymbol{\theta}'$. Given $\boldsymbol{\theta}$, $\boldsymbol{\mathcal{H}}$ and μ, optimising $\boldsymbol{\theta}'$ involves solving the following optimisation problem:

$$\boldsymbol{\theta}' = \arg\min_{\boldsymbol{\theta}'} \lambda_1 \left\| \sqrt{\sum_{i=1}^{L} \boldsymbol{\theta}'_i \odot \boldsymbol{\theta}'_i} \right\|_1 + \frac{\mu}{2} \sum_{i=1}^{L} \left\| \boldsymbol{\theta}_i - \boldsymbol{\theta}'_i + \frac{\boldsymbol{\eta}_i}{\mu} \right\|_2^2 . \tag{15}$$

We rewrite the objective by changing the summation index from channels to spatial features,

$$\boldsymbol{\theta}' = \arg\min_{\boldsymbol{\theta}'} \lambda_1 \sum_{j=1}^{D^2} \left\| \boldsymbol{\theta}'^j \right\|_2 + \frac{\mu}{2} \sum_{j=1}^{D^2} \left\| \boldsymbol{\theta}^j - \boldsymbol{\theta}'^j + \frac{\boldsymbol{\eta}^j}{\mu} \right\|_2^2 \tag{16}$$

and the closed-form optimal solution can be efficiently calculated for each spatial feature [69]:

$$\boldsymbol{\theta}'^j = \max\left(0, 1 - \frac{\lambda_1}{\mu \left\| \boldsymbol{g}^j \right\|_2}\right) \boldsymbol{g}^j \tag{17}$$

with $\boldsymbol{g}^j = \boldsymbol{\theta}^j + \frac{\boldsymbol{\eta}^j}{\mu}$. It is clear that $\boldsymbol{\theta}'^j$ tends to shrink to zero by collaboratively integrating the constraints imposed by all channels.

Updating $\boldsymbol{\mathcal{H}}$. Given $\boldsymbol{\theta}$, $\boldsymbol{\theta}'$ and μ, $\boldsymbol{\mathcal{H}}$ can be updated by:

$$\boldsymbol{\mathcal{H}} = \boldsymbol{\mathcal{H}} + \mu\left(\boldsymbol{\theta} - \boldsymbol{\theta}'\right), \tag{18}$$

and we update the penalty μ after each iteration as, $\mu = \min(\rho\mu, \mu_{\max})$, where $\rho > 1$ controls the convergence speed and $\mu_{\max}$ avoids the choice of excessive values. We pre-define K as the maximum number of iterations.

Complexity. The sub-problem of $\boldsymbol{\theta}$ needs FFT and inverse FFT in each iteration, which can be solved in $\mathcal{O}\left(LD^2\log(D)\right)$. The remaining element-wise operations can be solved in $\mathcal{O}(1)$ each. The total complexity of our optimisation framework is $\mathcal{O}\left(KLD^2\log(D)\right)$.

Algorithm 1 LADCF tracking algorithm.

Input Image frame $\boldsymbol{I}_t$, filter model $\boldsymbol{\theta}_{\text{model}}$, target center coordinate p_{t-1} and scale size $w \times h$ from frame $t-1$;
Tracking:
Extract search windows with S scales from $\boldsymbol{I}_t$ at p_{t-1};
Obtain corresponding feature representations $[\boldsymbol{\mathcal{X}}(s)]_{s=1}^{S}$;
Calculate response scores $\boldsymbol{f}$ using Equ. 19;
Set p_t and scale size $w \times h$ using Equ. 20;
Learning:
Obtain multi-channel feature representations $\boldsymbol{\mathcal{X}}$ based on current target bounding box;
Optimise $\boldsymbol{\theta}$ using Equ. 12 for K iterations;
Updating:
Update filter model $\boldsymbol{\theta}_{\text{model}}$ using Equ. 21;
Output Target bounding box (centre coordinate p_t and current scale size $w \times h$). Updated filter model $\boldsymbol{\theta}_{\text{model}}$ for frame t.

D. Tracking Framework

We propose the LADCF tracker based on learning adaptive discriminative correlation filters incorporating temporal consistency-preserving spatial feature selection. The tracking framework is summarised in Algorithm 1.

Position and scale detection. We follow fDSST [43] to achieve target position and scale detection simultaneously. When the new frame $\boldsymbol{I}_t$ becomes available, we extract a search window set $\left[\boldsymbol{I}_t^{\text{patch}}\{s\}\right]$ with multiple scales, $s = 1, 2, \dots, S$, with S denoting the number of scales. For each scale s, the search window patch is centred around the target centre position p_{t-1} with a size of $a^N n \times a^N n$ pixels, where a is the scale factor and $N = \lfloor \frac{2s-S-1}{2} \rfloor$. We resize each patch to $n \times n$ pixels with bilinear interpolation, where $n \times n$ is the basic search window size which is determined by the target size $w \times h$ and padding parameter ϱ as: $n = (1+\varrho)\sqrt{w \times h}$. Then we collect multi-channel feature representations for each scale search window as $\boldsymbol{\mathcal{X}}(s) \in \mathbb{R}^{D^2 \times L}$. Given the filter template $\boldsymbol{\theta}_{\text{model}}$, the response scores can be efficiently calculated in the frequency domain as:

$$\hat{\boldsymbol{f}}(s) = \hat{\boldsymbol{x}}(s) \odot \hat{\boldsymbol{\theta}}(s)^{*}_{\text{model}} \,. \tag{19}$$

Having performed IDFT of $\hat{\boldsymbol{f}}(s)$ for each scale, the relative position and scale are obtained by the maximum value of $\boldsymbol{f} \in \mathbb{R}^{D^2 \times S}$ as $\boldsymbol{f}(\Delta p, s^*)$. Then the resulting target bounding box parameters (centre p_t and scale size $w \times h$) are set as:

$$\begin{cases} p_t = p_{t-1} + \dfrac{n}{D}\Delta p \\ w = a^{\lfloor \frac{2s^* - S - 1}{2} \rfloor} w \\ h = a^{\lfloor \frac{2s^* - S - 1}{2} \rfloor} h \end{cases} . \tag{20}$$

Learning, updating and initialisation. It should be noted that in the learning stage, the multi-channel input $\boldsymbol{\mathcal{X}}$ in Equ. 9 forms the feature representation of the padded image patch centred at p_t with size $n \times n$. We follow our multi-channel temporal consistency preserving spatial feature selection embedded appearance leaning and optimisation framework as depicted in subsection IV-B and subsection IV-C to obtain the filter $\boldsymbol{\theta}$. In order to control the number of selected spatial features, we sort the ℓ-2 norm of each spatial feature vector $\left\|\boldsymbol{\theta}^j\right\|_2$ in the descending order and only preserve the largest M vectors. We adopt the same updating strategy as the traditional DCF method [9]:

$$\boldsymbol{\theta}_{\text{model}} = (1-\alpha)\boldsymbol{\theta}_{\text{model}} + \alpha\boldsymbol{\theta}, \tag{21}$$

where α is the updating rate. More specifically, as $\boldsymbol{\theta}_{\text{model}}$ is not available in the learning stage for the first frame, we use a pre-defined mask with only target region activated to optimise $\boldsymbol{\theta}$ as in BACF [13], and then initialise $\boldsymbol{\theta}_{\text{model}} = \boldsymbol{\theta}$ after the learning stage of the first frame.

V. Performance Evaluation

We perform qualitative and quantitative experimental evaluations to validate our method. In this section, we first describe the implementation details, including features and parameters of our tracking method. We then present the benchmark datasets and evaluation metrics used in our experiments, as well as several state-of-the-art trackers for comparison. We analyse the experimental results on different datasets and discuss the advantages of our method. In addition, an analysis in different components and parameters of our LADCF method is carried out to explore the sensitivity and effect of specific parameters.

A. Implementation Details

Feature representations: We use both hand-crafted and deep features in our method. It has been demonstrated that robust feature representation plays the most essential role in high-performance visual tracking [70], [71]. We equip LADCF with only hand-crafted features to compare with the trackers not using deep features and construct LADCF* with both features to facilitate a fair comparison with the trackers using deep structures. Table I shows the detailed setting of the features used for the evaluation. For CNN, we use the middle (Conv-3) convolutional layer in the VGG network[1] powered by the MatConvNet toolbox [2] [72].

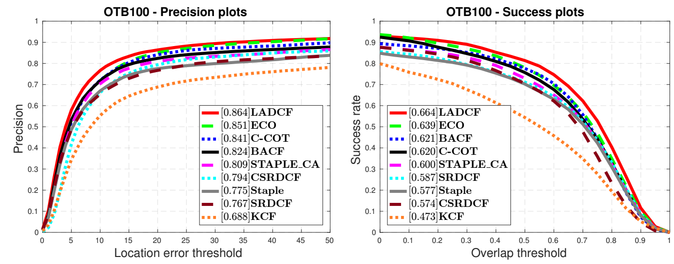

Fig. 3. The experimental results of trackers using hand-crafted features on OTB100. The precision plots (left) and the success plots (right) are presented.

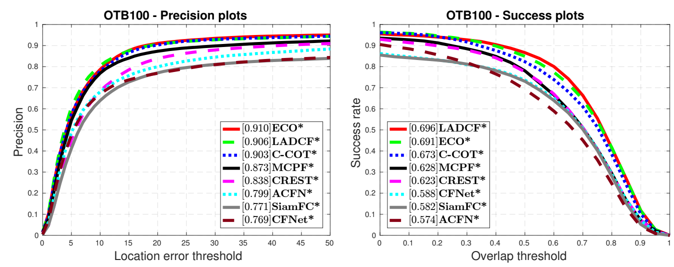

Fig. 4. The experimental results of trackers using deep features/structures on OTB100. The precision plots (left) and the success plots (right) are presented.

TABLE I
FEATURE REPRESENTATIONS FOR THE PROPOSED METHOD.

Feature Type	Hand-crafted		Deep
Feature	HOG	Colour-Names	CNN
Channels	31	10	512
Cell Size	4×4	4×4	-
LADCF	yes		no
LADCF*	yes		yes

Parameters setup: We set the regularisation parameters λ_1 and λ_2 in Equ. 11 as $\lambda_1 = 1$ and $\lambda_2 = 15$. The initial penalty parameter $\mu = 1$ and maximum penalty $\mu_{\max} = 20$ reached with the amplification rate $\rho = 5$. We set the maximum number of iterations as $K = 2$, the padding parameter as $\varrho = 4$, the scale factor as $a = 1.01$ and the number of scales as $S = 5$. The number of selected spatial features is set as $M = D^2 \times r$, where D^2 is the number of spatial features, determined by the target size and feature cell size in each sequence, and r is the selection ratio. For hand-crafted feature based LADCF, we set $r = 5\%$, and for deep feature based LADCF* we set $r = 20\%$. The learning rate α in Equ. 21 is set to 0.95 and 0.13 for LADCF and LADCF*, respectively. The parameter settings remain the same for all the experiments.

Experimental platform: We implement our LADCF and LADCF* in MATLAB 2016a on an Intel i5 2.50 GHz CPU

[1]http://www.vlfeat.org/matconvnet/models/imagenet-vgg-m-2048.mat
[2]http://www.vlfeat.org/matconvnet/

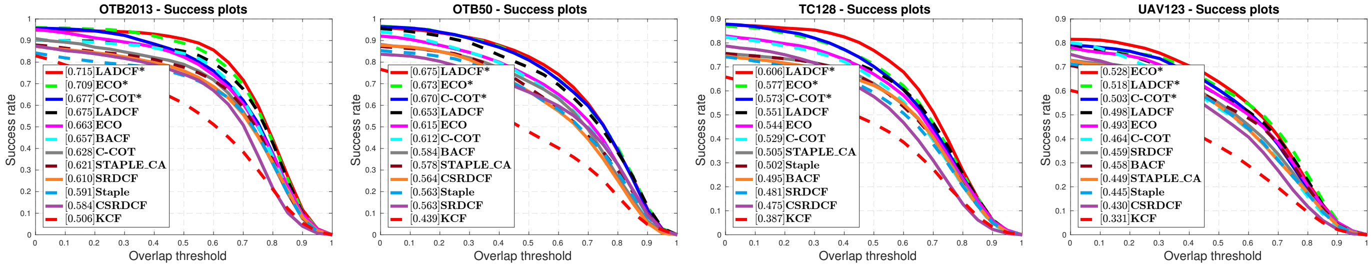

Fig. 5. Success plots of tracking performance on OTB2013, OTB50, Temple-Colour and UAV123, with AUC score reported in the figure legend.

TABLE II

THE OP RESULTS OF TRACKERS USING HAND-CRAFTED FEATURES ON OTB2013, OTB50 AND OTB100. (TOP THREE RESULTS ARE SHOWN IN RED, BLUE AND BROWN)

		KCF	CSRDCF	Staple	STAPLE_CA	SRDCF	BACF	ECO	C-COT	LADCF
Mean OP (%)	OTB2013	60.8	74.4	73.8	77.6	76.0	84.0	82.4	78.9	85.0
	OTB50	47.7	66.4	66.5	68.1	66.1	70.9	73.4	72.3	77.5
	OTB100	54.4	70.5	70.2	73.0	71.1	77.6	78.0	75.7	81.3
Mean FPS (%) on CPU	OTB100	92.7	4.6	23.8	20.1	2.7	16.3	15.1	1.8	18.2

TABLE III

THE OP RESULTS OF TRACKERS USING DEEP FEATURES/STRUCTURES ON OTB2013, OTB50 AND OTB100. (TOP THREE RESULTS ARE SHOWN IN RED, BLUE AND BROWN)

		CFNet*	SiamFC*	ACFN*	CREST*	MCPF*	ECO*	C-COT*	LADCF*
Mean OP (%)	OTB2013	78.3	77.9	75.0	86.0	85.8	88.7	83.7	90.7
	OTB50	68.8	68.0	63.2	68.8	69.0	81.0	80.9	82.5
	OTB100	73.6	73.0	69.2	77.6	78.0	84.9	82.3	86.7
Mean FPS (%) on CPU	OTB100	1.4	-	-	-	0.5	1.1	0.2	1.3
Mean FPS (%) on GPU		8.7	12.6	13.8	10.1	3.2	8.5	2.2	10.8

and Nvidia GeForce GTX 960M GPU. The MATLAB code is available at https://github.com/XU-TIANYANG/LADCF.git.

B. Experimental Setup

Datasets: We evaluate the performance of our tracking method on five benchmark datasets: OTB2013 [3], OTB50 [5], OTB100 [5], Temple-Colour [14], and UAV123 [15]. OTB2013, OTB50 and OTB100 are widely used datasets that respectively contain 51, 50 and 100 annotated video sequences with 11 sequence attributes. Temple-Colour is composed of 128 colour video sequences and UAV123 consists of 123 challenging sequences.

Evaluation metrics: We follow the One Pass Evaluation (OPE) protocol [3] to evaluate the performance of different trackers. The precision and success plots are reported based on centre location error and bounding box overlap. The Area Under Curve (AUC), Overlap Precision (OP, percentage of overlap ratios exceeding 0.5) and Distance Precision (DP, percentage of location errors within 20 pixels) are the criteria used in the evaluation. The speed of a tracking algorithm is measured in Frames Per Second (FPS).

State-of-the-art competitors: We compare our proposed tracking method with 13 state-of-the-art trackers, including Staple [40], SRDCF [11], KCF [62], CSRDCF [12], STAPLE_CA [73], BACF [13], C-COT [56], ECO [57], ACFN* [74], CREST* [34], SiamFC* [32], CFNet* [33], MCPF* [75], C-COT* and ECO*. Specifically, trackers followed by * are equipped with deep features/structures. For a fair comparison, we use the original publicly available codes from the authors for the evaluation.

C. Results and Analysis

Quantitative results of the overall tracking performance: We report the experimental results using the precision and success plots obtained on OTB100 in Fig. 3 and Fig. 4, with DP and AUC scores in the figure legend. Compared with the hand-crafted feature based trackers in Fig. 3, our LADCF performs best by achieving DP of 86.4% and AUC score of 66.4%. In addition, compared with deep features/structures based trackers in Fig. 4, our LADCF* outperforms the other trackers in terms of AUC metric with a score of 69.6%. Our LADCF* achieves DP of 90.6%, which ranks the second best, with only 0.4% fall behind ECO*. The main difference between our LADCF* and ECO* is that ECO* employs continuous convolution operators and is equipped with mixed deep layers (Conv-1 and Conv-5), that benefit the accuracy.

Table II and Table III show the OP results of the trackers on OTB2013, OTB50 and OTB100. Among the trackers with hand-crafted features, LADCF achieves the best results with absolute gains of 1%, 4.1% and 3.3% respectively over

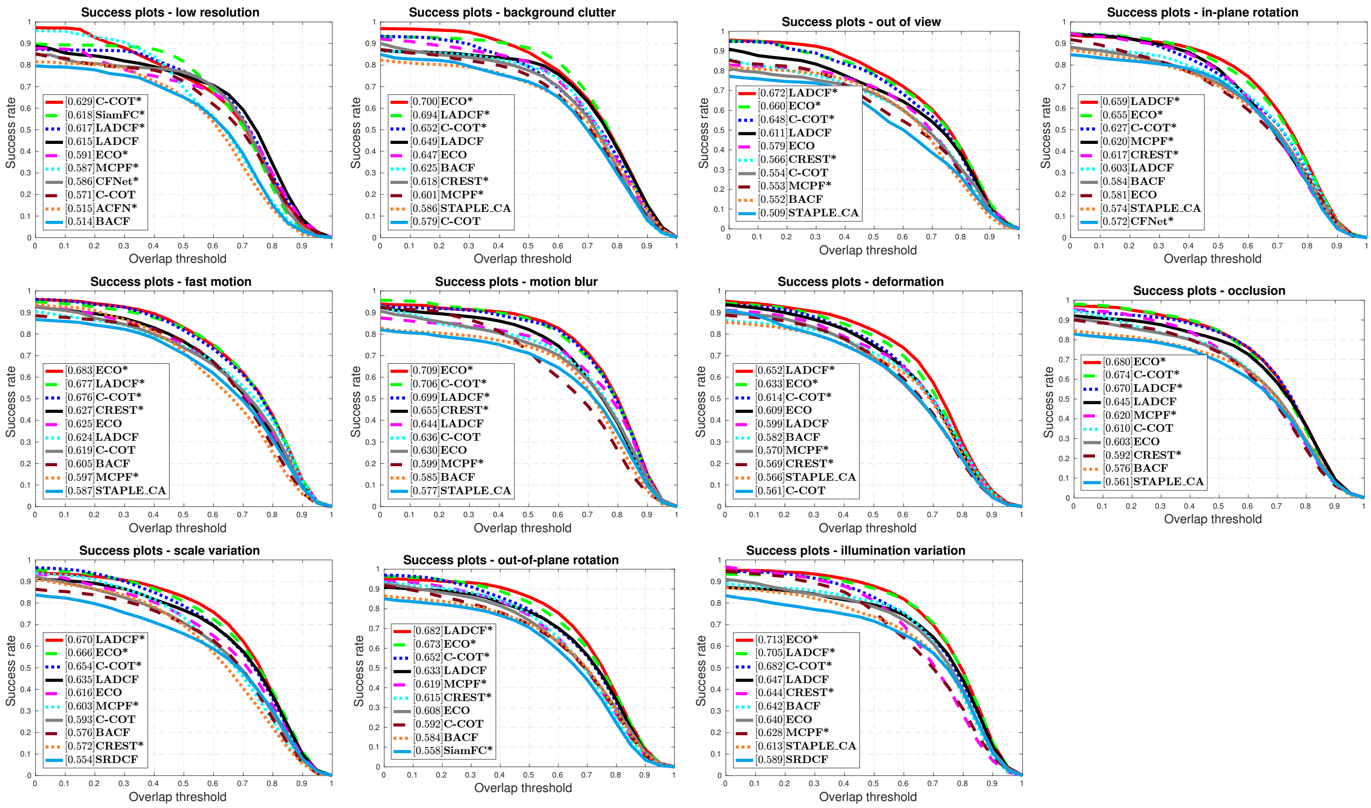

Fig. 6. Success plots based on tracking results on OTB100 in 11 sequence attributes, *i.e.* low resolution, background clutter, out of view, in-plane rotation, fast motion, motion blur, deformation, occlusion, scale variation, out-of-plane rotation and illumination variation. For clarity, only the top 10 trackers are presented for each attributes.

the second best one on the three datasets. The comparison with deep features/structures based trackers also supports the superiority of LADCF* over the state-of-the-art approaches. LADCF* performs better than ECO* by 2%, 1.5% and 1.8% on the listed datasets.

We also present the success plots on OTB2013, OTB50, Temple-Colour and UAV123 in Fig. 5 (We focus on the success plots here as overlaps are more important for evaluating the tracking performance). Our LADCF is the best, compared with other hand-crafted trackers on these four datasets. It outperforms recent trackers, *i.e.* CSRDCF (by 6.8% $\sim$ 9.1%), STAPLE_CA (by 4.6% $\sim$ 7.5%), C-COT (by 2.2% $\sim$ 4.7%), BACF (by 1.8% $\sim$ 6.9%) and ECO (by 0.5% $\sim$ 3.8%). In addition, LADCF* is better than ECO* and C-COT* except on UAV123. Specifically, LADCF* achieves an AUC score of 60.6% on Temple-Colour, which is better than ECO* and C-COT* by 2.9% and 3.3%, respectively. We find our consistent feature selection model is particularly useful on Temple-Colour, because all the sequences in Temple-Colour are of colour format, that is more suitable for Colour-Names and CNN features. However, ECO* leads with 1% above LADCF* on UAV123. A plausible reason is that the average number of frames per sequence in UAV123 is 915. This is much higher than OTB2013 (578), OTB50 (591) and Temple-Colour (429). Our consistent feature selection model only stores the filter model $\boldsymbol{\theta}_{\text{model}}$ from the previous frame, while ECO* is more sophisticated to deal with long-term tracking by collecting clusters of historical training samples during tracking.

Quantitative tracking performance results on sequence attributes: We provide the tracking results parameterised by 11 attributes on OTB100 in Fig. 6. Our LADCF* outperforms other trackers in 5 attributes, *i.e.* out of view, in-plane rotation, deformation, scale variation and out-of-plane rotation. Our consistent feature selection embedded appearance model enables adaptive spatial layout recognition, focusing on the relevant target and background regions with shared motion properties to create a robust complementary tracking pattern. The results of LADCF* in the other 6 attributes are among the top 3, demonstrating the effectiveness and robustness of our method. In addition, the superiority of LADCF is more obvious as it achieves the best performance in 9 attributes, compared to the other hand-crafted feature trackers. In particular, the performance of LADCF exhibits significant gains (4.4%, 3.2% and 3.5% as compared with the second best one in the attributes of low resolution, out of view and occlusion.

Qualitative tracking performance results: Fig. 7 shows the qualitative results of the state-of-the-art methods, *i.e.* BACF, STAPLE_CA, SRDCF, CFNet*, SiamFC*, C-COT*, ECO*, CREST*, MCPF* as well as our LADCF and LADCF*, on some challenging video sequences. The difficulties are posed by rapid changes in the appearance of targets. Our LADCF and LADCF* perform well on these challenges as we employ consistent embedded feature selection to identify the pertinent spatial layout. Sequences with deformations (*Bolt*,

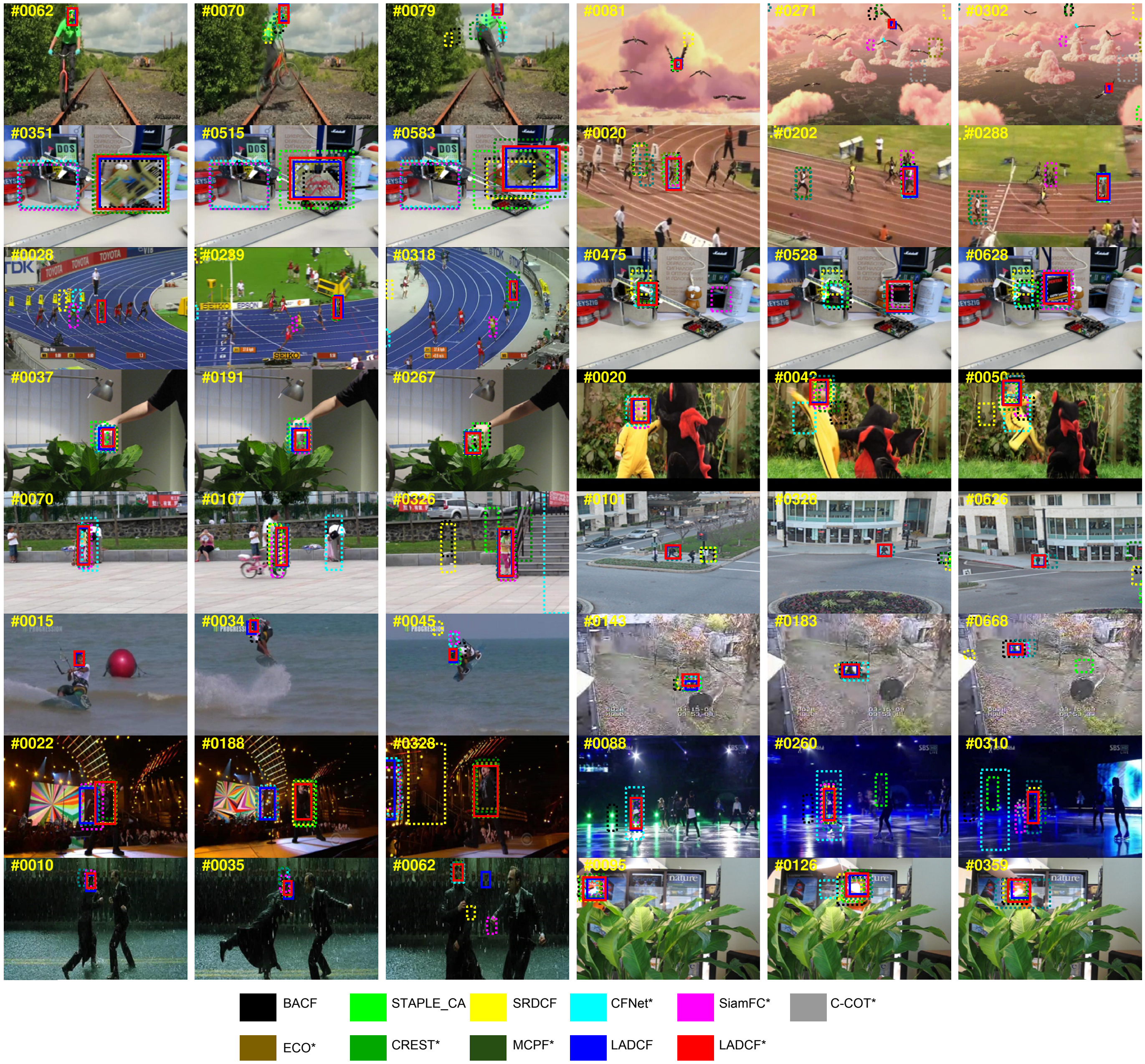

Fig. 7. Illustration of qualitative tracking results on challenging sequences (Left column top to down: *Biker*, *Board*, *Bolt*, *Coke*, *Girl2*, *Kitesurf*, *Singer2* and *Matrix*. Right column top to down: *Bird1*, *Bolt2*, *Box*, *Dragonbaby*, *Human3*, *Panda*, *Skating1* and *Tiger2*). The colour bounding boxes are the corresponding results of BACF, STAPLE_CA, SRDCF, CFNet*, SiamFC*, C-COT*, ECO*, CREST*, MCPF*, LADCF and LADCF*, respectively.

Dragonbaby) and out of view (*Biker*, *Bird1*) can be successfully tracked by our methods without any failures. Videos with occlusions (*Board*, *Girl2*, *Human3*, *Tiger2*) also benefit from our strategy of employing temporal consistency. Specifically, LADCF and LADCF* are expert in solving in-plane and out-of-plane rotations (*Coke*, *Dragonbaby*, *Skating1*), because the proposed adaptive spatial regularisation approach provides a novel solution to fusing the appearance information from the central region and surroundings.

D. *Self Analysis*

In this part, we provide a deep analysis to each component of our proposed LADCF method, *i.e.* feature configurations, temporal consistency and feature selection.

First, we employ 7 feature configurations to test our model using AUC metric on OTB100. As shown in Table IV, LADCF achieves 2.1% improvement by combining Colour-Names with HOG. The middle convolutional layers (Conv-3 and Conv-4) significantly improve the performance, compared with low (Conv-1 and Conv-2) and high layers (Conv-5).

Second, an analysis of the sensitivity to the learning rate α and feature selection ratio r is presented. As shown in

TABLE IV
TRACKING PERFORMANCE ON OTB100 WITH DIFFERENT FEATURE CONFIGURATIONS.

Features		AUC score
Hand-crafted	HOG	64.30%
	HOG+CN	66.44%
Hand-crafted+CNN	HOG+CN+Conv-1	66.72%
	HOG+CN+Conv-2	66.99%
	HOG+CN+Conv-3	69.65%
	HOG+CN+Conv-4	69.72%
	HOG+CN+Conv-5	68.15%

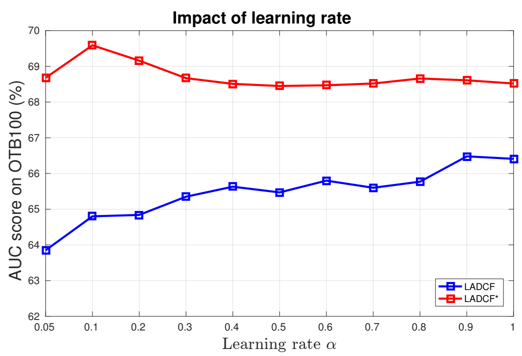

Fig. 8. The experimental results obtained by our temporal consistency preserving spatial feature selection enhanced appearance model on OTB100 for different learning rates.

Fig. 8, the tracking results vary smoothly with respect to the learning rate α, demonstrating that our method achieves stable performance with the proposed temporal consistency by forcing the learned filters to be instantiated in a low-dimensional manifold space to preserve diversity and generalisation. In addition, we analyse the impact of the feature selection ratio r in Fig. 9. The dash-dotted lines (LADCF-FS, LADCF*-FS) denote the corresponding results without feature selection. It is worth mentioning that hand-crafted and deep features achieve impressive improvements with the selection ratios ranging from $2\% \sim 20\%$ and $3\% \sim 40\%$ respectively. It supports the conclusion that tracking performance can be improved by using the proposed feature selection embedded filter learning scheme.

VI. CONCLUSION

We proposed an effective temporal consistency preserving spatial feature selection embedded approach to realise real-time visual tracking with outstanding performance. By reformulating the appearance learning model with embedded feature selection and imposing temporal consistency, we achieve adaptive discriminative filter learning on a low dimensional manifold with enhanced interpretability. Both hand-crafted and deep features are considered in our multi-channel feature representations. The extensive experimental results on tracking benchmark datasets demonstrate the effectiveness and robustness of our method, compared with state-of-the-art trackers.

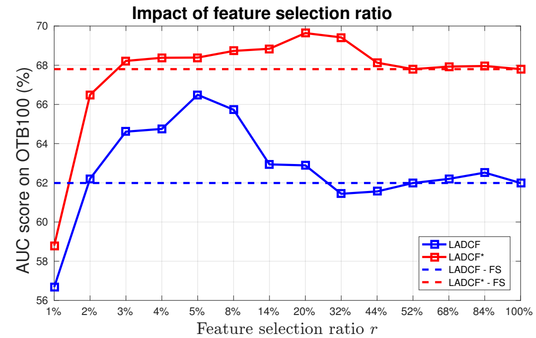

Fig. 9. The experimental results obtained by our temporal consistency preserving spatial feature selection enhanced appearance model on OTB100 for different feature selection ratios.

ACKNOWLEDGMENT

This work was supported in part by the EPSRC Programme Grant (FACER2VM) EP/N007743/1, EPSRC/dstl/MURI project EP/R018456/1, the National Natural Science Foundation of China (61373055, 61672265) and the NVIDIA GPU Grant Program.

REFERENCES

[1] A. Yilmaz, “Object tracking: A survey,” *ACM Computing Surveys*, vol. 38, no. 4, p. 13, 2006.
[2] H. Yang, L. Shao, F. Zheng, L. Wang, and Z. Song, “Recent advances and trends in visual tracking: A review,” *Neurocomputing*, vol. 74, no. 18, pp. 3823–3831, 2011.
[3] Y. Wu, J. Lim, and M. H. Yang, “Online object tracking: A benchmark,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 2411–2418.
[4] A. W. Smeulders, D. M. Chu, R. Cucchiara, S. Calderara, A. Dehghan, and M. Shah, “Visual tracking: An experimental survey,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 7, pp. 1442–1468, 2014.
[5] Y. Wu, J. Lim, and M. H. Yang, “Object tracking benchmark,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 9, pp. 1834–1848, 2015.
[6] A. Li, M. Lin, Y. Wu, M. H. Yang, and S. Yan, “Nus-pro: A new visual tracking challenge,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 2, pp. 335–349, 2016.
[7] M. Kristan, R. Pflugfelder, J. Matas, A. Leonardis, M. Felsberg, L. Cehovin, G. Fernandez, T. Vojir, G. Hager, and G. Nebehay, “The visual object tracking vot2015 challenge results,” in *IEEE International Conference on Computer Vision Workshop*, 2015, pp. 564–586.
[8] M. Kristan, A. Leonardis, J. Matas, M. Felsberg, R. Pflugfelder, L. ehovin, T. Vojr, G. Hger, A. Lukei, and G. Fernndez, “The visual object tracking vot2016 challenge results,” in *Computer Vision ECCV 2016 Workshops*, vol. 8926, 2016, pp. 191–217.
[9] J. Henriques, o F, 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.
[10] H. K. Galoogahi, T. Sim, and S. Lucey, “Correlation filters with limited boundaries,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 4630–4638.
[11] M. Danelljan, G. Hager, F. S. Khan, and M. Felsberg, “Learning spatially regularized correlation filters for visual tracking,” in *IEEE International Conference on Computer Vision*, 2015, pp. 4310–4318.
[12] A. Lukezic, T. Vojir, L. C. Zajc, J. Matas, and M. Kristan, “Discriminative correlation filter with channel and spatial reliability,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4847–4856.
[13] H. Kiani Galoogahi, A. Fagg, and S. Lucey, “Learning background-aware correlation filters for visual tracking,” in *IEEE International Conference on Computer Vision*, 2017.

[14] P. Liang, E. Blasch, and H. Ling, "Encoding color information for visual tracking: Algorithms and benchmark," *IEEE Transactions on Image Processing*, vol. 24, no. 12, pp. 5630–5644, 2015.

[15] 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.

[16] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *International Joint Conference on Artificial Intelligence*, 1981, pp. 674–679.

[17] S. Baker and I. Matthews, "Lucas-kanade 20 years on: A unifying framework," *International Journal of Computer Vision*, vol. 56, no. 3, pp. 221–255, 2004.

[18] D. Comaniciu, V. Ramesh, and P. Meer, "Real-time tracking of non-rigid objects using mean shift," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2000, pp. 142–149.

[19] A. Adam, E. Rivlin, and I. Shimshoni, "Robust fragments-based tracking using the integral histogram," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2006, pp. 798–805.

[20] D. A. Ross, J. Lim, R. S. Lin, and M. H. Yang, "Incremental learning for robust visual tracking," *International Journal of Computer Vision*, vol. 77, pp. 125–141, 2008.

[21] X. Mei and H. Ling, "Robust visual tracking using l1 minimization," in *IEEE International Conference on Computer Vision*, 2009, pp. 1436–1443.

[22] H. Ji, H. Ling, Y. Wu, and C. Bao, "Real time robust l1 tracker using accelerated proximal gradient approach," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 1830–1837.

[23] X. Jia, "Visual tracking via adaptive structural local sparse appearance model," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 1822–1829.

[24] K. Zhang, L. Zhang, and M. H. Yang, "Real-time compressive tracking," in *European Conference on Computer Vision*, 2012, pp. 864–877.

[25] T. Zhang, B. Ghanem, S. Liu, and N. Ahuja, "Robust visual tracking via structured multi-task sparse learning," *International Journal of Computer Vision*, vol. 101, no. 2, pp. 367–383, 2013.

[26] H. Grabner, M. Grabner, and H. Bischof, "Real-time tracking via on-line boosting," in *British Machine Vision Conference 2006, Edinburgh, Uk, September*, 2006, pp. 47–56.

[27] B. Babenko, M. H. Yang, and S. Belongie, "Robust object tracking with online multiple instance learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 8, pp. 1619–1632, 2011.

[28] K. Z, M. K, and M. J, "Tracking-learning-detection." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 7, pp. 1409–1422, 2012.

[29] S. Hare, A. Saffari, and P. H. S. Torr, "Struck: Structured output tracking with kernels," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 10, pp. 2096–2109, 2016.

[30] 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.

[31] R. Tao, E. Gavves, and A. W. Smeulders, "Siamese instance search for tracking," in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2016, pp. 1420–1429.

[32] Bertinetto, Luca and Valmadre, Jack and Henriques, Joao F and Vedaldi, Andrea and Torr, Philip HS, "Fully-convolutional siamese networks for object tracking," in *European conference on computer vision*. Springer, 2016, pp. 850–865.

[33] J. Valmadre, L. Bertinetto, J. Henriques, A. Vedaldi, and P. H. Torr, "End-to-end representation learning for correlation filter based tracking," in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2017, pp. 5000–5008.

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

[35] K. Briechle and U. D. Hanebeck, "Template matching using fast normalized cross correlation," in *Proceedings of SPIE*, vol. 4387, 2001, pp. 95–102.

[36] D. S. Bolme, J. R. Beveridge, B. A. Draper, and Y. M. Lui, "Visual object tracking using adaptive correlation filters," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2010, pp. 2544–2550.

[37] K. Zhang, L. Zhang, Q. Liu, D. Zhang, and M. H. Yang, "Fast visual tracking via dense spatio-temporal context learning," in *European Conference on Computer Vision*, 2014, pp. 127–141.

[38] J. V. D. 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. 1512–23, 2009.

[39] M. Danelljan, F. S. Khan, M. Felsberg, and J. V. D. Weijer, "Adaptive color attributes for real-time visual tracking," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1090–1097.

[40] L. Bertinetto, J. Valmadre, S. Golodetz, O. Miksik, and P. H. S. Torr, "Staple: Complementary learners for real-time tracking," in *IEEE Conference on Computer Vision and Pattern Recognition*, vol. 38, 2016, pp. 1401–1409.

[41] Y. Li and J. Zhu, "A scale adaptive kernel correlation filter tracker with feature integration," in *European Conference on Computer Vision Workshops*. Springer, 2014, pp. 254–265.

[42] M. Danelljan, G. Häger, F. Khan, and M. Felsberg, "Accurate scale estimation for robust visual tracking," in *British Machine Vision Conference*. BMVA Press, 2014, pp. 1–5.

[43] M. Danelljan, G. Häger, F. S. Khan, and M. Felsberg, "Discriminative scale space tracking," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 8, pp. 1561–1575, 2017.

[44] C. Ma, X. Yang, C. Zhang, and M. H. Yang, "Long-term correlation tracking," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 5388–5396.

[45] M. Tang and J. Feng, "Multi-kernel correlation filter for visual tracking," in *IEEE International Conference on Computer Vision*, 2015, pp. 3038–3046.

[46] S. Liu, T. Zhang, X. Cao, and C. Xu, "Structural correlation filter for robust visual tracking," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4312–4320.

[47] W. Zuo, X. Wu, L. Lin, L. Zhang, and M.-H. Yang, "Learning support correlation filters for visual tracking," *arXiv preprint arXiv:1601.06032*, 2016.

[48] T. Zhang, A. Bibi, and B. Ghanem, "In defense of sparse tracking: Circulant sparse tracker," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3880–3888.

[49] T. Liu, G. Wang, and Q. Yang, "Real-time part-based visual tracking via adaptive correlation filters," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 4902–4912.

[50] A. Bibi, M. Mueller, and B. Ghanem, "Target response adaptation for correlation filter tracking," in *European Conference on Computer Vision*, 2016, pp. 419–433.

[51] Y. Sui, Z. Zhang, G. Wang, Y. Tang, and L. Zhang, "Real-time visual tracking: Promoting the robustness of correlation filter learning," in *European Conference on Computer Vision*, 2016, pp. 662–678.

[52] Z. Hong, Z. Chen, C. Wang, X. Mei, D. Prokhorov, and D. Tao, "Multi-store tracker (muster): A cognitive psychology inspired approach to object tracking," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 749–758.

[53] K. Chen and W. Tao, "Convolutional regression for visual tracking," *IEEE Transactions on Image Processing*, vol. 27, no. 7, pp. 3611–3620, 2018.

[54] F. Liu, C. Gong, X. Huang, T. Zhou, J. Yang, and D. Tao, "Robust visual tracking revisited: From correlation filter to template matching," *IEEE Transactions on Image Processing*, vol. 27, no. 6, pp. 2777–2790, 2018.

[55] M. Danelljan, G. Hger, F. S. Khan, and M. Felsberg, "Adaptive decontamination of the training set: A unified formulation for discriminative visual tracking," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1430–1438.

[56] D. Martin, R. Andreas, K. Fahad, and F. Michael, "Beyond correlation filters: Learning continuous convolution operators for visual tracking," in *European Conference on Computer Vision*, 2016, pp. 472–488.

[57] M. Danelljan, G. Bhat, F. S. Khan, and M. Felsberg, "Eco: Efficient convolution operators for tracking," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6931–6939.

[58] H. Dadkhahi and M. F. Duarte, "Masking strategies for image manifolds," *IEEE Transactions on Image Processing*, vol. 25, no. 9, pp. 4314–4328, 2016.

[59] M. Barni, F. Bartolini, and A. Piva, "Improved wavelet-based watermarking through pixel-wise masking," *IEEE Transactions on Image Processing*, vol. 10, no. 5, pp. 783–91, 2001.

[60] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning: with Applications in R*. Springer New York, 2013.

[61] J. Gui, Z. Sun, S. Ji, D. Tao, and T. Tan, "Feature selection based on structured sparsity: A comprehensive study," *IEEE transactions on neural networks and learning systems*, vol. 28, no. 7, pp. 1490–1507, 2017.

[62] J. F. Henriques, C. Rui, 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.

[63] T. Zhang, S. Liu, N. Ahuja, M.-H. Yang, and B. Ghanem, "Robust visual tracking via consistent low-rank sparse learning," *International Journal of Computer Vision*, vol. 111, no. 2, pp. 171–190, 2015.

[64] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 68, no. 1, pp. 49–67, 2006.

[65] F. Nie, H. Huang, X. Cai, and C. H. Ding, "Efficient and robust feature selection via joint $\ell_{2,1}$-norms minimization," in *Advances in neural information processing systems*, 2010, pp. 1813–1821.

[66] R. J. Tibshirani, "Regression shrinkage and selection via the lasso. j r stat soc b," *Journal of the Royal Statistical Society*, vol. 58, pp. 267–288, 1996.

[67] F. Bach, R. Jenatton, J. Mairal, and G. Obozinski, "Structured sparsity through convex optimization," *Statistical Science*, pp. 450–468, 2012.

[68] 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.

[69] Y. Yang, H. T. Shen, Z. Ma, Z. Huang, and X. Zhou, "l2, 1-norm regularized discriminative feature selection for unsupervised learning," in *IJCAI proceedings-international joint conference on artificial intelligence*, vol. 22, no. 1, 2011, p. 1589.

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

[71] E. Gundogdu and A. A. Alatan, "Good features to correlate for visual tracking," *IEEE Transactions on Image Processing*, vol. 27, no. 5, pp. 2526–2540, 2018.

[72] A. Vedaldi and K. Lenc, "Matconvnet: Convolutional neural networks for matlab," in *Proceedings of the 23rd ACM international conference on Multimedia*. ACM, 2015, pp. 689–692.

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

[74] J. Choi, H. J. Chang, S. Yun, T. Fischer, Y. Demiris, J. Y. Choi *et al.*, "Attentional correlation filter network for adaptive visual tracking," in *IEEE Conference on Computer Vision and Pattern Recognition*, vol. 2, 2017.

[75] T. Zhang, C. Xu, and M.-H. Yang, "Multi-task correlation particle filter for robust object tracking," in *IEEE Conference on Computer Vision and Pattern Recognition*, vol. 1, no. 2, 2017, p. 3.