Joint Group Feature Selection and Discriminative Filter Learning for Robust Visual Object Tracking

Tianyang Xu[1,2] Zhen-Hua Feng[2] Xiao-Jun Wu[1*] Josef Kittler[2]
[1]School of Internet of Things Engineering, Jiangnan University, Wuxi, China
[2] Centre for Vision, Speech and Signal Processing (CVSSP), University of Surrey, Guildford, UK
tianyang_xu@163.com, z.feng@surrey.ac.uk, wu_xiaojun@jiangnan.edu.cn, j.kittler@surrey.ac.uk

Abstract

We propose a new Group Feature Selection method for Discriminative Correlation Filters (GFS-DCF) based visual object tracking. The key innovation of the proposed method is to perform group feature selection across both channel and spatial dimensions, thus to pinpoint the structural relevance of multi-channel features to the filtering system. In contrast to the widely used spatial regularisation or feature selection methods, to the best of our knowledge, this is the first time that channel selection has been advocated for DCF-based tracking. We demonstrate that our GFS-DCF method is able to significantly improve the performance of a DCF tracker equipped with deep neural network features. In addition, our GFS-DCF enables joint feature selection and filter learning, achieving enhanced discrimination and interpretability of the learned filters.

To further improve the performance, we adaptively integrate historical information by constraining filters to be smooth across temporal frames, using an efficient low-rank approximation. By design, specific temporal-spatial-channel configurations are dynamically learned in the tracking process, highlighting the relevant features, and alleviating the performance degrading impact of less discriminative representations and reducing information redundancy. The experimental results obtained on OTB2013, OTB2015, VOT2017, VOT2018 and TrackingNet demonstrate the merits of our GFS-DCF and its superiority over the state-of-the-art trackers. The code is publicly available at `https://github.com/XU-TIANYANG/GFS-DCF`.

1. Introduction

To consistently and accurately track an arbitrary object in video sequences is a very challenging task. The difficulties are posed by a wide spectrum of appearance variations of an object in unconstrained scenarios. Among existing tracking algorithms, Discriminative Correlation Filters (DCF-) based trackers [27] have exhibited promising results in recent benchmarks [81, 82, 53, 43] and competitions such as the Visual Object Tracking (VOT) challenges [36, 32, 33, 34].

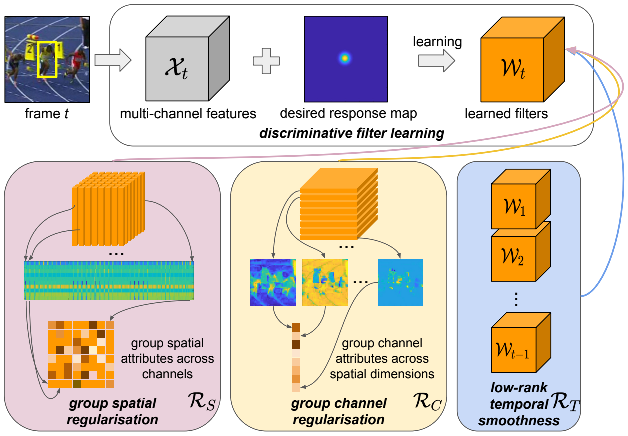

frame t
$\mathcal{X}_t$
multi-channel features
desired response map
learning
$\mathcal{W}_t$
learned filters
discriminative filter learning
group spatial attributes across channels
group spatial regularisation
$\mathcal{R}_S$
group channel attributes across spatial dimensions
group channel regularisation
$\mathcal{R}_C$
$\mathcal{W}_1$
$\mathcal{W}_2$
$\mathcal{W}_{t-1}$
low-rank temporal smoothness
$\mathcal{R}_T$

Figure 1. In contrast to the classical DCF paradigm, our GFS-DCF performs channel and spatial group feature selection for the learning of correlation filters. Group sparsity is enforced in the channel and spatial dimensions to highlight relevant features with enhanced discrimination and interpretability. Additionally, a low-rank temporal smoothness constraint is employed across temporal frames to improve the stability of the learned filters.

The success of high-performance DCF trackers is attributed to three aspects: spatial regularisation, temporal smoothness and robust image feature representation. Regarding the first point, as natural images and videos are projections from a 3D space into a 2D plane, spatial regularisation directly improves the tracking accuracy by potentially endowing the learned filters with a specific attention mechanism, enhancing the discrimination by focusing on less ambiguous regions [15, 49, 31, 87]. Second, based on the fact that video sequences are formed by discrete image sampling of continuous dynamic scenes, reflecting the temporal smoothness of successive frames in the construction of appearance models has been shown to improve their generalisation capacity [17, 40, 13, 51]. Third, with the development of robust image feature extraction methods, *e.g.* Histogram of Oriented Gradient (HOG) [12], Colour Names (CN) [80] and Convolutional Neural Network (CNN) features [37, 70, 47, 19], the performance of DCF-based trackers has been greatly improved [6, 34, 33]. It is indisputable that recent advances in DCF-based tracking owe to a great extent to the use of robust deep CNN features.

Despite the rapid progress in visual tracking by equipping the trackers with robust image features, the structural relevance of multi-channel features to the filtering system has not been adequately investigated. In particular, due to the limited number of training samples available for visual tracking, DCF-based trackers usually use a deep network pre-trained on other computer vision tasks, such as VGG [64] or ResNet [24] trained on ImageNet [63]. In such a case, the extracted deep feature channels (maps) for an arbitrary object, which may exceed thousands, may not be compact. They may include irrelevant as well as redundant descriptors and their presence may degrade the target detection performance. However, the tension between discrimination, information relevance, information redundancy and high-dimensional feature representations has not been systematically studied in the existing DCF paradigm. We argue it is absolutely crucial to perform dimensionality reduction along the channel dimension to suppress irrelevant features as well as redundancy for deep neural network features.

To redress the above oversight, we propose a new Group Feature Selection method for DCF-based visual object tracking, namely GFS-DCF. To be more specific, we reduce the information redundancy and irrelevance of high-dimensional multi-channel features by performing group feature selection across both spatial and channel dimensions, resulting in compact target representations. It should be highlighted that our GFS-DCF differs significantly from existing DCF-based trackers, in which only spatial regularisation or selection is used. Additionally, as supervised frame-to-frame data fitting may create excessive variability in prediction, we constrain a learned predictor (filter) to be smooth across the time dimension (frames).

Fig. 1 depicts the basic learning scheme of the proposed GFS-DCF method. Given the predicted location of an object in the tth frame, we first extract multi-channel features. Then the extracted features and desired response map are used to learn the correlation filters for the prediction of the target in the next frame. In the filter learning stage, the combination of channel-spatial group feature selection and low-rank constraints adaptively identifies a specific temporal-spatial-channel configuration for robust discriminative filter learning. As a result, relevant features are highlighted to improve discrimination and decrease redundancy. The main contributions of our GFS-DCF method include:

- A new group feature selection method for multi-channel image representations, reducing the dimensionality across both spatial and channel dimensions. To the best of our knowledge, this is the first work that considers feature compression along both spatial and channel dimensions. According to our experiments, the proposed group channel feature selection method improves the performance of a DCF-based tracker significantly when using deep CNN features.
- A temporal smoothness regularisation term used to obtain highly correlated filters among successive frames. To this end, we use an efficient low-rank approximation, forcing the learned filters to lie in a low-dimensional manifold with consistent temporal-spatial-channel configurations.
- A comprehensive evaluation of GFS-DCF on a number of well-known benchmarks, including OTB2013 [81], OTB2015 [82], VOT2017 [33], VOT2018 [34], and TrackingNet [55]. The results demonstrate the merits of GFS-DCF, as well as its superiority over the state-of-the-art trackers.

2. Related Work

Existing visual object tracking approaches include template matching [48], statistical learning [2], particle filters [1], subspace learning [62], discriminative correlation filters [27], deep neural networks [61] and Siamese networks [91, 77, 39]. In this section, we focus on DCF-based approaches due to their outstanding performance as evidenced by recent tracking competitions such as VOT [32, 34]. For the other visual tracking approaches, readers are referred to comprehensive reviews [65, 82, 32, 38] .

One of the seminal works in the development of DCF is MOSSE [7], which formulates the tracking task as discriminative filter learning [8] rather than template matching [10]. The concept of circulant matrix [21] is introduced to DCF by CSK [26] with a padded search window, which generates more background samples for the learning stage. Additionally, spatial-temporal context [86] and kernel tricks [27] are used to improve the learning formulation with the consideration of local appearance and nonlinear metric, respectively. The DCF paradigm has further been extended by exploiting scale detection [41, 14, 16], structural patch analysis [42, 46, 45], multi-clue fusion [71, 50, 28, 4, 72], sparse representation [88, 90], support vector machine [75, 92], enhanced sampling mechanisms [89, 54] and end-to-end deep neural networks [73, 67].

Despite the great success of DCF in visual object tracking, it is still a very challenging task to achieve high-performance tracking for an arbitrary object in unconstrained scenarios. The main obstacles include: spatial boundary effect, limited feature representation capacity and temporal filter degeneration.

To alleviate the boundary effect problem caused by the circulant structure, SRDCF [15] stimulates the interest in spatial regularisation [17, 51, 13, 40], which allocates more energy for the central region of a filter using a predefined spatial weighting function. A similar idea has been pursued by means of pruning the training samples or learned filters with a predefined mask [20, 49, 31, 40]. Different from those approaches, to achieve spatial regularisation, LSART forces the output to focus on a specific region of a target [69]. A common characterisation of the above approaches is that they are all based on a fixed spatial regularisation pattern, for example, a predefined mask or weighting function. To achieve adaptive spatial regularisation, LADCF [83] embeds dynamic spatial feature selection in the filter learning stage. Thanks to this innovation it has achieved the best results in the public VOT2018 dataset [34]. The above spatial regularisation methods decrease the ambiguity emanating from the background and enable a relatively large search window for tracking. Nevertheless, these approaches only consider information compression along the spatial dimension. In contrast, our GFS-DCF method performs group feature selection along both the channel and spatial dimensions, resulting in more compact object appearance descriptions.

Second, as feature representation is the most essential factor from the point of view of high-performance visual tracking [76], combinations of hand-crafted and deep features have widely been used in DCF-based trackers [18, 27, 6]. However, the structural relevance of multi-channel features in the filter learning system has not been considered. The redundancy and interference from the high-dimensional representations impede the effectiveness of learning dense filters. To unify the process of information selection across the spatial and channel dimensions, our GFS-DCF performs group feature selection and discriminative filter learning jointly.

Last, to mitigate temporal filter degeneration, historical clues are reflected in SRDCFdecon [17] and C-COT [51], with enhanced robustness and temporal smoothness, by gathering multiple previous frames in the filter learning stage. To alleviate the computational burden, ECO [13] manages the inherent computational complexity by clustering historical frames and employing projection matrix for multi-channel features. Our GFS-DCF, on the other hand, is robust to temporal appearance variations by constraining the learned filters to be smooth across frames using an efficient low-rank approximation. Consequently, the relevant spatial-channel features are consistently highlighted in a dynamic low-dimensional subspace.

3. DCF-based Visual Object Tracking

Given the initial location of an object in a video, the aim of visual object tracking is to localise the object in the successive video frames. Assume we have the estimated location of the object in the tth frame. To localise the object in the $t+1$th frame, DCF [27] learns multi-channel filters, $\mathcal{W}_t \in \mathbb{R}^{N\times N\times C}$, using a pair of training samples $\{\mathcal{X}_t, \mathbf{Y}\}$, where $\mathcal{X}_t \in \mathbb{R}^{N\times N\times C}$ is a tensor consisting of C-channel features extracted from the tth frame and $\mathbf{Y} \in \mathbb{R}^{N\times N}$ is the desired response map identifying the object location. To obtain $\mathcal{W}_t$, DCF formulates the objective as a regularised least square problem:

$$\widetilde{\mathcal{W}_t} = \arg\min_{\mathcal{W}_t} \left\| \sum_{k=1}^{C} \mathbf{W}_t^k \circledast \mathbf{X}_t^k - \mathbf{Y} \right\|_F^2 + \mathcal{R}\left(\mathcal{W}_t\right), \quad (1)$$

where $\circledast$ is the circular convolution operator [26], $\mathbf{X}_t^k \in \mathbb{R}^{N\times N}$ is the k-th channel feature representation, $\mathbf{W}_t^k \in \mathbb{R}^{N\times N}$ is the corresponding discriminative filter and $\mathcal{R}\left(\mathcal{W}_t\right) = \lambda \sum_{k=1}^{C} \|\mathbf{W}_t^k\|_F^2$ is a regularisation term. A closed-form solution to the above optimisation task can efficiently be obtained in the frequency domain [27].

In the tracking stage, the filters learned from the first frame are directly used to localise the object in the second frame. For the other frames, the filers are updated as:

$$\mathcal{W}_t = \alpha\widetilde{\mathcal{W}_t} + (1-\alpha)\,\mathcal{W}_{t-1}, \quad (2)$$

where $\alpha \in [0, 1]$ is a pre-defined updating rate. Given a search window in the $(t+1)$st frame, we first extract multi-channel features, $\mathcal{X}_{t+1}$. Then the learned correlation filters, $\mathcal{W}_t$, from the tth frame are used to estimate the response map in the frequency domain efficiently:

$$\hat{\mathbf{R}} = \sum_{k=1}^{C} \hat{\mathbf{X}}_{t+1}^k \odot \hat{\mathbf{W}}_t^k, \quad (3)$$

where $\hat{\cdot}$ denotes Discrete Fourier Transform (DFT) and $\odot$ denotes element-wise multiplication. The element with the maximal value in the original response map, obtained by inverse DFT, corresponds to the predicted target location.

4. Group Feature Selection for DCF

4.1. GFS-DCF

In DCF-based visual object tracking, multi-channel features are extracted from a large search window, in which only a small region is of interest. In such a case, multi-channel image features are usually redundant and may bring uncertainty in the prediction stage. To address this issue, spatial feature selection or regularisation has been widely used in existing DCF-based trackers, such as the use of fixed spatial masks [31, 49, 40, 13]. More recently, a learning-based adaptive mask [83] has been proposed to inject spatial regularisation to DCF-based visual tracking, achieving the best performance on the VOT2018 public dataset [34]. However, investigations aiming at reducing the information redundancy and noise across feature channels, especially as it applies to hundreds or thousands of deep CNN feature maps, are missing from the existing literature. To close this gap, in this paper, we advocate a new feature selection method, namely Group Feature Selection (GFS), for DCF-based visual object tracking.

In contrast to previous studies, our GFS-DCF incorporates group feature selection, in both spatial and channel dimensions, in the original DCF optimisation task. Additionally, a low-rank regularisation term is used to achieve temporal smoothness of the learned filters during the tracking process. We assume that the learning of correlation filters is conducted for the tth frame and omit the subscript '$_t$' for simplicity. The objective function of our GFS-DCF is:

$$\begin{aligned} \widetilde{\mathcal{W}} = \arg\min_{\mathcal{W}} & \left\| \sum_{k=1}^{C} \mathbf{W}^k \circledast \mathbf{X}^k - \mathbf{Y} \right\|_F^2 \\ & + \lambda_1 \mathcal{R}_S(\mathcal{W}) + \lambda_2 \mathcal{R}_C(\mathcal{W}) + \lambda_3 \mathcal{R}_T(\mathcal{W}), \end{aligned} \tag{4}$$

where $\mathcal{R}_S(\mathcal{W})$ is the spatial group regularisation term for spatial feature selection, $\mathcal{R}_C(\mathcal{W})$ is the group regularisation term for channel selection, $\mathcal{R}_T(\mathcal{W})$ is the temporal regularisation term and each λ_i is a balancing parameter. These regularisation terms are introduced in detail in the remaining part of this section and a solution of the above optimisation task is given in Section 4.4.

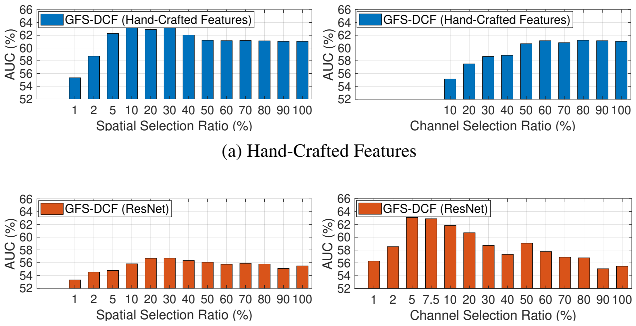
4.2. Group Spatial-Channel Regularisation

Grouping is introduced into the model with the aim of exploiting prior knowledge that is scientifically meaningful [30]. Considering the current nature of feature representations that are invariably multi-channel, and the spatial coherence of a tracked object, the grouping information is employed in $\mathcal{R}_S$ and $\mathcal{R}_C$ to achieve spatial-channel selection by allocating individual variables into specific groups with certain visual meaning (spatial location and channel attributes). This strategy has been demonstrated to be effective in visual data science [57, 3, 85, 29, 22, 78, 79].

To perform group feature selection for the spatial domain, we define the spatial regularisation term as:

$$\mathcal{R}_S(\mathcal{W}) = \sum_{i=1}^{N} \sum_{j=1}^{N} \left\| \mathbf{w}_{ij:} \right\|_2 , \tag{5}$$

in which we use ℓ_2 norm to obtain the grouping attribute of each spatial location, calculated across all the feature channels. To be more specific, we concatenate all the elements at the ith location of the first order and the jth location of the second order of the multi-channel feature tensor, $\mathcal{W} \in \mathbb{R}^{N \times N \times C}$, into a vector $\mathbf{w}_{ij:} = [w_{ij1}, ..., w_{ijC}]^\top$, as illustrated in Fig. 1. The grouping attribute is obtained by the ℓ_2 norm and then the implicit ℓ_1 norm of all the spatial grouping attributes is used to regularise the optimisation of correlation filters. This naturally injects sparsity into the spatial domain by grouping all the elements across channels. Such structured spatial sparsity enables robust group feature selection that reflects the joint contribution of features in the spatial domain.



(a) Hand-Crafted Features

(b) Deep CNN Features

Figure 2. A comparison of spatial and channel group feature selection on OTB2015 using either (a) hand-crafted or (b) deep CNN features, parameterised by selection ratio.

In our preliminary experiments, we found that the proposed group feature selection in spatial domain is able to improve the performance of a DCF tracker when using hand-crafted features. However, the improvement is minor when we tried to impose spatial feature selection into deep CNN features. We argue that the main reason is that an element in deep CNN feature maps stands for higher-level concepts thus performing spatial feature selection on such features cannot achieve fine-grained selection of the target region from the background. For example we use the feature maps at the 'res4x' layer of ResNet50 [24] in the proposed method. Each deep CNN feature map has the resolution of 13 × 13, in which each pixel corresponds to a 16 × 16 region in the original input image. To perform spatial selection on such a small resolution feature map cannot achieve very accurate spatial feature selection results. But deep CNN features usually have many channels, with the cosequence of injecting information redundancy. To address this issue, we propose channel selection by defining a group regularisation term in the channel dimension:

$$\mathcal{R}_C(\mathcal{W}) = \sum_{k=1}^{C} \left\| \mathbf{W}^k \right\|_F , \tag{6}$$

where we use the Frobenius norm to obtain the grouping attributes for feature channels $\{\mathbf{W}^k\}_{k=1}^{C}$. Note again that implicitly the constraint in (6) is a sparsity inducing ℓ_1 norm.

In practice, to perform spatial/channel feature selection, we use the measures in Equ. (5) and Equ. (6). Specifically, we first calculate the group attributes in spatial/channel domain and then eliminate the features across channel/spatial dimensions corresponding to a pre-defined proportion with the lowest grouping attributes. This selection strategy has been commonly used in many previous studies [84, 59, 66]. Additionally, the proposed feature selection method is applied to each individual feature type separately.

To evaluate the effectiveness of the proposed spatial and channel group feature selection methods, we compare the proposed GFS-DCF with the classical DCF formulation on the OTB2015 dataset. The results are shown in Fig. 2. It should be noted that the bar at the selection ratio of 100% stands for the original DCF tracker without feature selection, using either hand-crafted features or deep CNN features. We use Colour Names, Intensity Channels and HOG for hand-crafted features, and ResNet50 for deep features. Detailed experimental settings are introduced in Section 5.1. As reported in the figure, for hand-crafted features, impressive improvements are achieved with the spatial selection ratios ranging from $5\% \sim 40\%$. But channel selection cannot improve the performance for hand-crafted features. The only merit of performing channel selection on hand-crafted features is that we can maintain the performance when compressing the features to 60% in size. For deep features, the use of spatial feature selection only improves the performance marginally. But, deep features benefit significantly from the channel selection regularisation, with the AUC increasing from 55.49% to 63.07% even when we only use 5% of the original channels. These results demonstrate that deep features are highly redundant across channels, and exhibit undesirable interference. The evaluation validates the proposed spatial-channel group feature selection strategy.

As such, the proposed method offers a scope for dimensionality reduction by the proposed group spatial-channel regularisation, leading to performance boosting. While hand-crafted features are extracted in a fixed manner with relatively high resolutions compared to deep features, different attributes are considered for different channels, with more redundancy and ambiguity in the spatial dimension. The results support the conclusion that the tracking performance can be improved by using the proposed group-feature-selection-embedded filter learning scheme.

4.3. Temporal Smoothness

Despite the success of feature selection in many computer vision and pattern recognition tasks, it suffers from the instability of solutions, especially in the presence of information redundancy [52]. To mitigate this problem and take appearance variation into consideration [62], we improve the robustness of learned correlation filters by injecting temporal smoothness. Specifically, a low-rank constraint is enforced on the estimates across video frames, so that the temporal coherence in the filter design is promoted. We define the constraint as minimising:

$$rank\left(\mathbb{W}_t\right) - rank\left(\mathbb{W}_{t-1}\right), \tag{7}$$

where $\mathbb{W}_t = [\mathbf{vec}(\mathcal{W}_1), ..., \mathbf{vec}(\mathcal{W}_t)] \in \mathbb{R}^{N^2C \times t}$ is a matrix, with each column storing the vectorised correlation filters, $\mathcal{W}$.

Here, the constraint (7) imposes a low-rank property across frames because it impacts on the selection process since the second frame. However, it is inefficient to calculate $rank\left(\mathbb{W}_t\right)$, especially in long-term videos with many frames. Therefore, we use its sufficient condition as a substitute:

$$d\left(\mathcal{W}_t - \mathcal{U}_{t-1}\right), \tag{8}$$

where $\mathcal{U}_{t-1} = \sum_{k=1}^{t-1} \mathcal{W}_k/(t-1)$ is the mean over all the previous learned filters and d is a distance metric. The brief proof of sufficiency is provided as follows.

Proof: Given $\mathbb{W}_{t-1}$ and $\mathcal{U}_{t-1}$, the mean vector of $\mathbb{W}_t$ is influenced by $\mathcal{W}_t$. We denote $\check{\mathbf{w}}_t = \mathbf{vec}\left(\sqrt{\frac{t-1}{t}}\left(\mathcal{W}_t - \mathcal{U}_{t-1}\right)\right)$. Expressing $\mathbb{W}_{t-1}$ in terms of its SVD as $\mathbb{W}_{t-1} = \mathbf{U}_{t-1}\mathbf{\Sigma}_{t-1}\mathbf{W}_{t-1}^\top$, we have,

$$\mathbb{W}_t = \left[\mathbf{U}_{t-1}, \mathbf{vec}\left(\mathcal{W}_t\right)_\perp\right] \mathbf{R} \begin{bmatrix} \mathbf{W}_{t-1}^\top & 0 \\ 0 & 1 \end{bmatrix}, \tag{9}$$

and

$$\mathbf{R} = \begin{bmatrix} \mathbf{\Sigma}_{t-1} & \mathbf{\Sigma}_{t-1}^\top \mathbf{vec}\left(\mathcal{W}_t\right) \\ 0 & \check{\mathbf{w}}_t^\top \left(\mathbf{I} - \mathbf{U}_{t-1}\mathbf{U}_{t-1}^\top\right) \check{\mathbf{w}}_t \end{bmatrix}, \tag{10}$$

where $\mathbf{I} \in \mathbb{R}^{N^2Ct \times N^2Ct}$ is the identity matrix and $_\perp$ performs orthogonalisation of the vector, $\mathbf{vec}\left(\mathcal{W}_t\right)$, to the matrix, $\mathbf{U}_{t-1}$. If $\mathcal{W}_t = \mathcal{U}_{t-1}$, then $\mathbf{\Sigma}_{t-1}$ dominates the eigenvalues of $\mathbf{R}$. Consequently, $\mathbf{R}$ shares the same eigenvalues as $\mathbb{W}_t$. □

Therefore, we propose to adaptively enforce the temporal low-rank property with the regularisation term:

$$\mathcal{R}_{\mathrm{T}}\left(\mathcal{W}\right) = \lambda_3 \sum_{k=1}^{C} \left\|\mathbf{W}_t^k - \mathbf{W}_{t-1}^k\right\|_F^2. \tag{11}$$

4.4. Solution

Due to the convexity of the proposed formulation, we apply the augmented Lagrange method [44] to optimise Equ. (4). Concretely, we introduce slack variable $\mathcal{W}' = \mathcal{W}$ and construct the following Lagrange function:

$$\begin{aligned} \mathcal{L} = & \left\|\sum_{k=1}^{C} \mathbf{W}_t^k \circledast \mathbf{X}_t^k - \mathbf{Y}\right\|_F^2 + \lambda_1 \sum_{k=1}^{C} \left\|\mathbf{W}_t'^k\right\|_F \\ & + \lambda_2 \sum_{i=1}^{N}\sum_{j=1}^{N} \left\|\mathbf{w}_{ij_t}'\right\|_2 + \lambda_3 \sum_{k=1}^{C} \left\|\mathbf{W}_t^k - \mathbf{W}_{t-1}^k\right\|_F^2 \\ & + \frac{\mu}{2} \sum_{k=1}^{C} \left\|\mathbf{W}_t^k - \mathbf{W}_t'^k + \frac{\mathbf{\Gamma}^k}{\mu}\right\|_F, \end{aligned} \tag{12}$$

where Γ is the Lagrange multiplier sharing the same size as $\mathcal{X}$, $\mathbf{\Gamma}^k$ is its k-th channel, and μ is the corresponding penalty. Then the Alternating Direction Method of Multipliers [9] is employed to perform iterative optimisation with guaranteed convergence as follows [60]:

$$\hat{\mathbf{w}}_{ij_t} = \left(\mathbf{I} - \frac{\hat{\mathbf{x}}_{ij_t}\hat{\mathbf{x}}_{ij_t}^H}{\left(\lambda_3 + \mu/2\right) N^2 + \hat{\mathbf{x}}_{ij_t}^H \hat{\mathbf{x}}_{ij_t}}\right)\mathbf{q}, \tag{13a}$$

$$w'^k_{ij_t} = \max\left(0, 1 - \frac{\lambda_1}{\mu\left\|\mathbf{P}^k\right\|_F} - \frac{\lambda_2}{\mu\left\|\mathbf{p}_{ij}\right\|_2}\right) p^k_{ij}, \tag{13b}$$

$$\Gamma = \Gamma + \mu\left(\mathcal{W}_t - \mathcal{W}'_t\right), \tag{13c}$$

where $\mathbf{q} = (\hat{\mathbf{x}}_{ij_t}\hat{y}_{ij}/N^2 + \mu\hat{\mathbf{w}}'_{ij_t} - \mu\hat{\gamma}_{ij} + \lambda_3\hat{\mathbf{w}}_{ij_{t-1}})/(\lambda_3 + \mu)$ and $p^k_{ij} = w^k_{ij} + \gamma^k_{ij}/\mu$.

5. Evaluation

5.1. Implementation and Evaluation Settings

We implement our GFS-DCF using MATLAB 2018a. The speed of GFS-DCF is 8 frames per second (fps) on a platform with one Intel Xeon E5-2637 v3 CPU and NVIDIA GeForce GTX TITAN X GPU. We set $\lambda_1 = 10$ and $\lambda_2 = 1$ for group feature selection. In order to guarantee a fixed number of the selected channels and spatial units, we set up the channel selection ratio r_c and spatial selection ratio r_s to truncate the remaining channels and spatial units. We extract hand-crafted features using Colour Names (CN), HOG, Intensity Channels (IC), and deep CNN features using ResNet-50 [24, 74]. For hand-crafted features, we set the parameters as $r_c = 90\%$, $r_s = 10\%$, $\lambda_3 = 16$ and $\alpha = 0.6$. For deep features, we set the parameters as $r_c = 7.5\%$, $r_s = 90\%$, $\lambda_3 = 12$ and $\alpha = 0.05$.

We evaluated the proposed method on several well-known benchmarks, including OTB2013/OTB2015 [81, 82], VOT2017/VOT2018 [33, 34] and TrackingNet Test dataset [55], and compared it with a number of state-of-the-art trackers, such as VITAL [68], MetaT [58], ECO [13], MCPF [89], CREST [67], BACF [31], CFNet [73], CACF [54], ACFN [11], CSRDCF [49], C-COT [51], Staple [4], SiamFC [5], SRDCF [15], KCF [27], SAMF [41], DSST [16] and other advanced trackers in VOT challenges, *i.e.*, CFCF [23], CFWCR [25], LSART [69], UPDT [6], SiamRPN [91], MFT [34] and LADCF [83].

To measure the tracking performance, we follow the corresponding protocols [82, 32, 35]. We use precision plot and success plot [81] for OTB2013 and OTB2015. Four numerical values, *i.e.* centre location error (CLE), distance precision (DP), overlap precision (OP) and area under curve (AUC), are further employed to measure the performance. For VOT2017 and VOT2018, we employ expected average overlap (EAO), accuracy value and robustness to evaluate the performance [32]. For TrackingNet, we adopt success score, precision score and normalised precision to analyse the results [55].

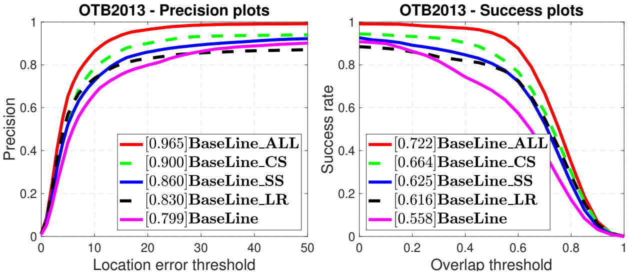


Figure 3. A comparison of different regularisation terms in GFS-DCF, evaluated on OTB2013. The precision plots with **DP** and the success plots with **AUC** in the legends are presented.

Table 1. A comparison of different methods on different videos of OTB2015, in terms of the rank of the matrix formed by stacking all the vectorised filters of all the frames in a video. (The best three results are highlighted by red, blue and brown.)

Video [#frames]	KCF	CACF	ECO	C-COT	**GFS-DCF**
Deer [71]	71	14	**4**	**3**	**2**
Basketball [725]	526	134	**23**	**10**	**9**
Boy [602]	274	63	**19**	**8**	**4**
David3 [252]	252	53	**8**	**3**	**6**
Girl [500]	267	57	**18**	**8**	**5**
Suv [945]	701	49	**16**	**4**	**6**
Skater [160]	160	38	**19**	**3**	**5**
Woman [597]	384	111	**15**	**6**	**7**

5.2. Ablation Study

We first evaluate the effect of each innovative component in GFS-DCF, including the spatial selection term $\mathcal{R}_S$ (SS), channel selection term $\mathcal{R}_C$ (CS) and low-rank temporal smooth term $\mathcal{R}_T$ (LR). The baseline is the original DCF tracker equipped with the same features (both hand-crafted and deep features) and updating rate as our GFS-DCF. We construct 5 trackers, *i.e.*, BaseLine, BaseLine_SS, BaseLine_CS, BaseLine_LR and BaseLine_ALL, to analyse internal effectiveness. The results evaluated on OTB2013 are reported in Fig. 3.

According to the figure, the proposed channel selection, spatial selection and low-rank smoothness terms improve the performance of the classical DCF (BaseLine). Compared with the classical DCF, the grouping channel/spatial selection terms $\mathcal{R}_C$/$\mathcal{R}_S$ (BaseLine_CS/BaseLine_SS) significantly improve the performance in terms of DP and AUC by 10.1%/6.1% and 10.6%/6.7%. The results are consistent with Fig. 2, demonstrating the redundancy and noise in the multi-channel representations and the advantage of performing group feature selection to achieve parsimony. On the other hand, the low-rank temporal smoothness term $\mathcal{R}_T$

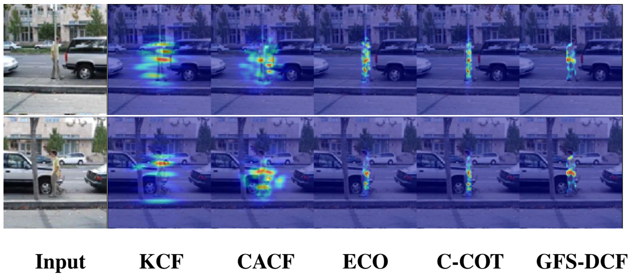


Figure 4. Visualisation of filters using *David3* in OTB2015. We visualise the corresponding filters in frame #50 (the 1st row) and #200 (the 2nd row). To better visualise the sparsity, we present the heat-maps of the obtained filters by gathering the energy across all the channels.

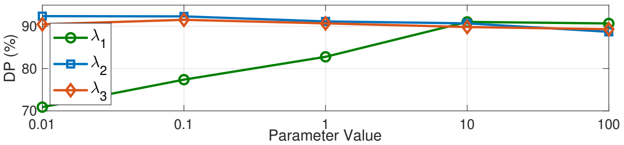


Figure 5. Impact of λ_1, λ_2 and λ_3, evaluated on OTB2015.

(BaseLine_LR) also leads to improvement in the tracking performance. Intuitively explained, a low-rank constraint across temporal frames enables the learned filters to become more invariant to appearance variations. To verify the practical low-rank property, we further collect the filters of each frame, concatenate them together, and calculate the rank. To guarantee the quality of the collected filters, we only consider some simple sequences where all the involved trackers can successfully track the target across the entire frames, *i.e.* the filters are effective in distinguishing the target from surroundings. The results are presented in Table 1, which show that our simplified regularisation term, Equ. (11), can achieve the low-rank property by only considering the filter model. Note, C-COT and ECO also share the low-rank property, achieved by taking into account historical appearance in the learning stage, but at the expense of increased complexity and storage.

We further visualise the filters of 5 different trackers in Fig. 4. Note that ECO and C-COT achieve sparsity by spatial regularisation, with more energy concentrating in the centre region. In contrast, our GFS-DCF realises sparsity without a pre-defined mask or weighting. The filters are adaptively shrunk to specific groups (channels/spatial units). Therefore, our GFS-DCF may shrink the elements even within the centre region.

In addition, we perform the sensitivity analysis of λ_1, λ_2 and λ_3. As shown in Fig. 5, our GFS-DCF achieves stable performance with $\lambda_1, \lambda_2 \in [0.01, 100]$ and $\lambda_3 \in [10, 100]$. Though we have to set 7 parameters, the selection ratios are most essential, as shown in Fig. 2. We employ threshold pruning operators to fix the ratio of selected spatial units and channels, enabling robustness against regularisation parameters.

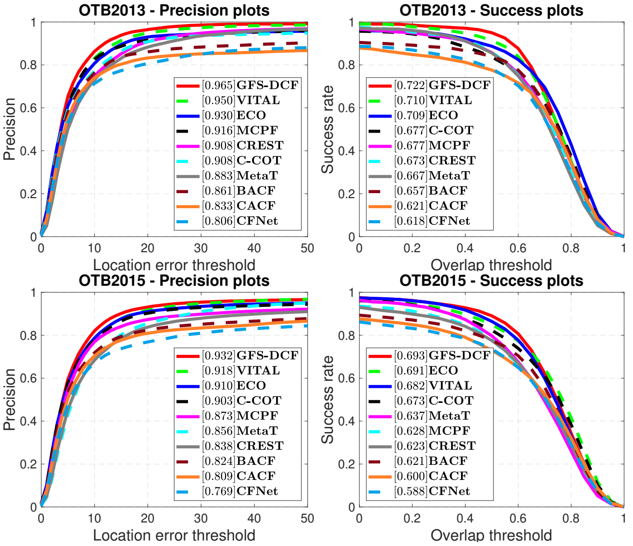


Figure 6. The experimental results on OTB2013, and OTB2015. The precision plots with **DP** reported in the figure legend (*first column*) and the success plots with **AUC** reported in the figure legend (*second column*) are presented.

Table 2. Tracking results with different features on OTB2015.

Feature	Method	OP	DP
HOG	BACF	77.6%	82.4%
	CSRDCF	70.5%	79.4%
	SRDCF	71.1%	76.7%
	LADCF	**78.5%**	83.1%
	GFS-DCF	78.2%	**85.2%**
HOG+CN	ECO	78.0%	85.1%
	C-COT	75.7%	84.1%
	GFS-DCF	**81.5%**	**86.3%**
HOG+CN+VGG-M	ECO	84.9%	91.0%
	C-COT	82.3%	90.3%
	GFS-DCF	**85.5%**	**91.2%**

Finally, the combination of all the components (BaseLine_ALL) becomes our GFS-DCF tracker (Fig. 3), which achieves the best performance compared with individual components. The results demonstrate the effectiveness of the proposed grouping and low-rank formulations.

5.3. Comparison with the State-of-the-art

OTB We report the precision and success plots for OTB2013 and OTB2015 in Fig. 6. Overall, our GFS-DCF outperforms all the other state-of-the-art trackers in terms of DP and AUC. Compared with the second best tracker, GFS-DCF achieves the improvements by $1.5\%/1.2\%$ and $1.4\%/0.2\%$ (in DP/AUC) on OTB2013 and OTB2015, respectively.

To achieve a fair comparison of mathematical formula-

Table 3. The **OP** and **CLE** results on OTB2013, TB50 and OTB2015. (The best three results are highlighted by red, blue and brown.)

		KCF	SAMF	DSST	SRDCF	SiamFC	Staple	C-COT	CSRDCF	ACFN
OP/CLE (%/pixels)	**OTB2013**	60.8/36.3	69.6/29.0	59.7/39.2	76.0/36.8	77.9/29.7	73.8/31.4	83.7/15.6	74.4/31.9	75.0/18.7
	TB50	47.7/54.3	59.3/40.5	45.9/59.5	66.1/42.7	68.0/36.8	66.5/32.3	80.9/**12.3**	66.4/30.3	63.2/32.1
	OTB2015	54.4/45.1	64.6/34.6	53.0/49.1	71.1/39.7	73.0/33.2	70.2/31.8	82.3/**14.0**	70.5/31.1	69.2/25.3
SPEED (fps)		**82.7**	11.5	15.6	2.7	12.6	**23.8**	2.2	4.6	13.8
		CACF	CFNet	BACF	CREST	MCPF	ECO	MetaT	VITAL	**GFS-DCF**
OP/CLE (%/pixels)	**OTB2013**	77.6/29.8	78.3/35.2	84.0/26.2	86.0/**10.2**	85.8/11.2	**88.7**/16.2	85.6/11.5	**91.4/7.4**	**95.0/5.92**
	TB50	68.1/36.3	68.8/36.7	70.9/30.3	68.8/32.6	69.9/30.9	**81.0**/13.2	73.7/17.0	**81.3/12.5**	**82.8/12.4**
	OTB2015	73.0/33.1	73.6/36.0	77.6/28.2	77.6/21.2	78.0/20.9	**84.9**/14.8	79.8/14.2	**86.5/9.9**	**89.0/10.3**
SPEED (fps)		**18.1**	8.7	16.3	10.1	0.5	12.5	0.8	1.3	7.8

Table 4. Tracking results on VOT2017/VOT2018. (The best three results are highlighted by red, blue and brown.)

	ECO [13]	CFCF [23]	CFWCR [25]	LSART [69]	UPDT [6]	SiamRPN [91]	MFT [34]	LADCF [83]	GFS-DCF
EAO	0.280	0.286	0.303	0.323	0.378	0.383	**0.385**	**0.389**	**0.397**
Accuracy	0.483	0.509	0.484	0.493	**0.536**	**0.586**	0.505	0.503	**0.511**
Robustness	0.276	0.281	0.267	0.218	0.184	0.276	**0.140**	**0.159**	**0.143**

tions, we also compared our method with the state-of-the-art trackers using the same features on OTB2015. As shown in Table 2, our GFS-DCF performs better than almost all the other approaches regardless of the features used, demonstrating the advantage of the proposed method.

We also present the detailed OP, CLE and speed (fps) of all the involved trackers on OTB2013, TB50 and OTB2015 in Table 3. On OTB2013, our GFS-DCF tracker achieves the OP of 95.0% and CLE of $5.92\ pixels$. Compared with the recent VITAL and MetaT trackers based on end-to-end deep neural networks, our performance gain is $3.6\%/1.48\ pixel$ and $8.4\%/5.58\ pixels$ in terms of OP and CLE, respectively. On TB50, GFS-DCF performs better than C-COT (by 1.9%) in terms of OP but with a lower CLE (by $0.1\ pixel$). In addition, on OTB2015, our tracker outperforms many recent trackers, *i.e.* CSRDCF (by $18.5\%/20.8\ pixels$), CACF (by $16.0\%/22.8\ pixels$), C-COT (by $6.7\%/3.7\ pixels$), BACF (by $11.4\%/17.9\ pixels$) and ECO (by $4.1\%/4.5\ pixels$) in terms of OP/CLE.

VOT Table 4 presents the results obtained on VOT2017/VOT2018 dataset [34]. Our method achieves the best EAO score, 0.397, outperforming recent advanced trackers, *e.g.*, LADCF, UPDT and SiamRPN. In addition, the reported Accuracy (0.511) and Robustness (0.143) results of GFS-DCF are also within the top three, demonstrating the effectiveness of the proposed group selection framework.

TrackingNet We also report the results generated by the TrackingNet [55] evaluation server (511 test sequences) in Table 5. Our GFS-DCF achieves 71.97% in normalised precision, demonstrating its advantages as compared with the other state-of-the-art methods.

Table 5. Evaluation on the TrackingNet test set.

Method	Success	Precision	Normalised Precision
CACF [54]	53.59%	46.72%	60.84%
ECO [13]	56.13%	48.86%	62.14%
MDNet [56]	$\mathbf{61.35}\%$	55.53%	71.00%
GFS-DCF	60.90%	$\mathbf{56.57\%}$	$\mathbf{71.79\%}$

In conclusion, the proposed GFS-DCF tracking method achieves advanced performance, as compared to the state-of-the-art trackers, with favourable speed.

6. Conclusion

We proposed an effective appearance model with outstanding performance by learning spatial-channel group-sparse discriminative correlation filters, constrained by low-rank approximation across successive frames. By reformulating the appearance learning model so as to incorporate group-sparse regularisation and a temporal smoothness constraint, we achieved adaptive temporal-spatial-channel filter learning on a low dimensional manifold with enhanced interpretability of the learned model. The extensive experimental results on visual object tracking benchmarks demonstrate the effectiveness and robustness of our method, compared with the state-of-the-art trackers. The diversity of hand-crafted and deep features in terms of spatial and channel dimensions is examined to support the conclusion that different selection strategies should be performed on different feature categories.

Acknowledgment This work was supported in part by the National Natural Science Foundation of China (61672265, U1836218, 61876072), the 111 Project of Ministry of Education of China (B12018), the EPSRC Programme Grant (FACER2VM) EP/N007743/1, EPSRC/dstl/MURI project EP/R018456/1 and the NVIDIA GPU Grant Program.

References

[1] M Sanjeev Arulampalam, Simon Maskell, Neil Gordon, and Tim Clapp. A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking. *IEEE TSP*, 50(2):174–188, 2002.

[2] Shai Avidan. Support vector tracking. *IEEE TPAMI*, 26(8):1064–1072, 2004.

[3] Francis Bach, Rodolphe Jenatton, Julien Mairal, Guillaume Obozinski, et al. Structured sparsity through convex optimization. *Statistical Science*, 27(4):450–468, 2012.

[4] Luca Bertinetto, Jack Valmadre, Stuart Golodetz, et al. Staple: Complementary learners for real-time tracking. In *CVPR*, volume 38, pages 1401–1409, 2016.

[5] Bertinetto, Luca and Valmadre, Jack and Henriques, Joao F and Vedaldi, Andrea and Torr, Philip HS. Fully-convolutional siamese networks for object tracking. In *ECCV*, pages 850–865, 2016.

[6] Goutam Bhat, Joakim Johnander, Martin Danelljan, Fahad Shahbaz Khan, and Michael Felsberg. Unveiling the power of deep tracking. *arXiv preprint arXiv:1804.06833*, 2018.

[7] David S Bolme, J Ross Beveridge, Bruce A Draper, and Yui Man Lui. Visual object tracking using adaptive correlation filters. In *CVPR*, pages 2544–2550, 2010.

[8] David S Bolme, Bruce A Draper, and J Ross Beveridge. Average of synthetic exact filters. In *CVPR*, pages 2105–2112, 2009.

[9] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122, 2011.

[10] Kai Briechle and Uwe D Hanebeck. Template matching using fast normalized cross correlation. In *Optical Pattern Recognition XII*, volume 4387, pages 95–103. International Society for Optics and Photonics, 2001.

[11] Jongwon Choi, Hyung Jin Chang, Sangdoo Yun, Tobias Fischer, et al. Attentional correlation filter network for adaptive visual tracking. In *CVPR*, pages 4807–4816, 2017.

[12] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *CVPR*, pages 886–893, 2005.

[13] Martin Danelljan, Goutam Bhat, Fahad Shahbaz Khan, and Michael Felsberg. ECO: Efficient convolution operators for tracking. In *CVPR*, pages 6931–6939, 2017.

[14] Martin Danelljan, Gustav Häger, Fahad Khan, and Michael Felsberg. Accurate scale estimation for robust visual tracking. In *BMVC*, pages 1–5, 2014.

[15] Martin Danelljan, Gustav Hager, Fahad Shahbaz Khan, and Michael Felsberg. Learning spatially regularized correlation filters for visual tracking. In *ICCV*, pages 4310–4318, 2015.

[16] Martin Danelljan, Gustav Häger, Fahad Shahbaz Khan, and Michael Felsberg. Discriminative scale space tracking. *IEEE TPAMI*, 39(8):1561–1575, 2017.

[17] Martin Danelljan, Gustav Hager, Fahad Shahbaz Khan, and Michael Felsberg. Adaptive decontamination of the training set: A unified formulation for discriminative visual tracking. In *CVPR*, pages 1430–1438, 2016.

[18] Martin Danelljan, Fahad Shahbaz Khan, Michael Felsberg, and Joost Van De Weijer. Adaptive color attributes for real-time visual tracking. In *CVPR*, pages 1090–1097, 2014.

[19] Xingping Dong, Jianbing Shen, Wenguan Wang, Yu Liu, Ling Shao, and Fatih Porikli. Hyperparameter optimization for tracking with continuous deep q-learning. In *CVPR*, pages 518–527, 2018.

[20] Hamed Kiani Galoogahi, Terence Sim, and Simon Lucey. Correlation filters with limited boundaries. In *CVPR*, pages 4630–4638, 2015.

[21] Robert M Gray. Toeplitz and circulant matrices : a review. *Foundations and Trends in Communications and Information Theory*, 2(3):155–239, 2006.

[22] Jie Gui, Zhenan Sun, Shuiwang Ji, Dacheng Tao, and Tieniu Tan. Feature selection based on structured sparsity: A comprehensive study. *IEEE TNNLS*, 28(7):1490–1507, 2017.

[23] Erhan Gundogdu and A Aydın Alatan. Good features to correlate for visual tracking. *IEEE TIP*, 27(5):2526–2540, 2018.

[24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[25] Zhiqun He, Yingruo Fan, Junfei Zhuang, Yuan Dong, and HongLiang Bai. Correlation filters with weighted convolution responses. In *ICCV*, pages 1992–2000, 2017.

[26] Joao F Henriques, Rui Caseiro, Pedro Martins, and Jorge Batista. Exploiting the circulant structure of tracking-by-detection with kernels. In *ECCV*, pages 702–715, 2012.

[27] Joao F. Henriques, Caseiro Rui, Pedro Martins, and Jorge Batista. High-speed tracking with kernelized correlation filters. *IEEE TPAMI*, 37(3):583–596, 2015.

[28] Zhibin Hong, Zhe Chen, Chaohui Wang, Xue Mei, Danil Prokhorov, and Dacheng Tao. Multi-store tracker (muster): A cognitive psychology inspired approach to object tracking. In *CVPR*, pages 749–758, 2015.

[29] Yaohua Hu, Chong Li, Kaiwen Meng, Jing Qin, and Xiaoqi Yang. Group sparse optimization via lp, q regularization. *JMLR*, 18(1):960–1011, 2017.

[30] Jian Huang, Patrick Breheny, and Shuangge Ma. A selective review of group selection in high-dimensional models. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 27(4), 2012.

[31] Hamed Kiani Galoogahi, Ashton Fagg, and Simon Lucey. Learning background-aware correlation filters for visual tracking. In *ICCV*, pages 1135–1143, 2017.

[32] Matej Kristan, Aleš Leonardis, Jiri Matas, and et al. The visual object tracking vot2016 challenge results. Springer, Oct 2016.

[33] Matej Kristan, Ales Leonardis, Jiri Matas, and et al. The Visual Object Tracking VOT2017 Challenge Results, 2017.

[34] Matej Kristan, Ales Leonardis, Jiri Matas, and et al. The sixth Visual Object Tracking VOT2018 challenge results. In *ECCV Workshops*, 2018.

[35] Matej Kristan, Jiri Matas, Ales Leonardis, et al. A novel performance evaluation methodology for single-target trackers. *IEEE TPAMI*, 38(11):2137–2155, 2016.

[36] Matej Kristan, Roman Pflugfelder, Jiri Matas, et al. The visual object tracking vot2015 challenge results. In *ICCVW*, pages 564–586, 2015.

[37] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, pages 1097–1105, 2012.

[38] Annan Li, Min Lin, Yi Wu, Ming Hsuan Yang, and Shuicheng Yan. Nus-pro: A new visual tracking challenge. *IEEE TPAMI*, 38(2):335–349, 2016.

[39] Bo Li, Junjie Yan, Wei Wu, Zheng Zhu, and Xiaolin Hu. High performance visual tracking with siamese region proposal network. In *CVPR*, pages 8971–8980, 2018.

[40] Feng Li, Cheng Tian, Wangmeng Zuo, Lei Zhang, and Ming-Hsuan Yang. Learning spatial-temporal regularized correlation filters for visual tracking. *arXiv preprint arXiv:1803.08679*, 2018.

[41] Yang Li and Jianke Zhu. A scale adaptive kernel correlation filter tracker with feature integration. In *ECCVW*, pages 254–265, 2014.

[42] Yang Li, Jianke Zhu, and Steven CH Hoi. Reliable patch trackers: Robust visual tracking by exploiting reliable patches. In *CVPR*, pages 353–361, 2015.

[43] Pengpeng Liang, Erik Blasch, and Haibin Ling. Encoding color information for visual tracking: Algorithms and benchmark. *IEEE TIP*, 24(12):5630–5644, 2015.

[44] Zhouchen Lin, Minming Chen, and Yi Ma. The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices. *arXiv preprint arXiv:1009.5055*, 2010.

[45] Si Liu, Tianzhu Zhang, Xiaochun Cao, and Changsheng Xu. Structural correlation filter for robust visual tracking. In *CVPR*, pages 4312–4320, 2016.

[46] Ting Liu, Gang Wang, and Qingxiong Yang. Real-time part-based visual tracking via adaptive correlation filters. In *CVPR*, pages 4902–4912, 2015.

[47] Xiankai Lu, Chao Ma, Bingbing Ni, Xiaokang Yang, Ian Reid, and Ming-Hsuan Yang. Deep regression tracking with shrinkage loss. In *ECCV*, pages 353–369, 2018.

[48] Bruce D. Lucas and Takeo Kanade. An iterative image registration technique with an application to stereo vision. In *IJCAI*, pages 674–679, 1981.

[49] Alan Lukezic, Tomas Vojir, Luka Cehovin Zajc, Jiri Matas, and Matej Kristan. Discriminative correlation filter with channel and spatial reliability. In *CVPR*, pages 4847–4856, 2017.

[50] Chao Ma, Xiaokang Yang, Chongyang Zhang, and Ming Hsuan Yang. Long-term correlation tracking. In *CVPR*, pages 5388–5396, 2015.

[51] Danelljan Martin, Robinson Andreas, Khan Fahad, and Felsberg Michael. Beyond correlation filters: Learning continuous convolution operators for visual tracking. In *ECCV*, pages 472–488, 2016.

[52] Nicolai Meinshausen and Peter Bühlmann. Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):417–473, 2010.

[53] Matthias Mueller, Neil Smith, and Bernard Ghanem. A benchmark and simulator for uav tracking. In *ECCV*, pages 445–461, 2016.

[54] Matthias Mueller, Neil Smith, and Bernard Ghanem. Context-aware correlation filter tracking. In *CVPR*, pages 1396–1404, 2017.

[55] Matthias Muller, Adel Bibi, Silvio Giancola, Salman Alsubaihi, and Bernard Ghanem. Trackingnet: A large-scale dataset and benchmark for object tracking in the wild. In *ECCV*, pages 300–317, 2018.

[56] Hyeonseob Nam and Bohyung Han. Learning multi-domain convolutional neural networks for visual tracking. In *CVPR*, pages 4293–4302, 2016.

[57] Feiping Nie, Heng Huang, Xiao Cai, and Chris H Ding. Efficient and robust feature selection via joint $\ell_{2,1}$-norms minimization. In *NIPS*, pages 1813–1821, 2010.

[58] Eunbyung Park and Alexander C Berg. Meta-tracker: Fast and robust online adaptation for visual object trackers. *arXiv preprint arXiv:1801.03049*, 2018.

[59] Xi Peng, Zhang Yi, and Huajin Tang. Robust subspace clustering via thresholding ridge regression. In *AAAI*, pages 3827–3833, 2015.

[60] Kaare Brandt Petersen, Michael Syskind Pedersen, et al. The matrix cookbook. *Technical University of Denmark*, 7(15):510, 2008.

[61] Yuankai Qi, Shengping Zhang, Lei Qin, Hongxun Yao, Qingming Huang, Jongwoo Lim, and Ming-Hsuan Yang. Hedged deep tracking. In *CVPR*, pages 4303–4311, 2016.

[62] David A Ross, Jongwoo Lim, Ruei-Sung Lin, and Ming-Hsuan Yang. Incremental learning for robust visual tracking. *IJCV*, 77(1-3):125–141, 2008.

[63] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, 115(3):211–252, 2015.

[64] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[65] Arnold WM Smeulders, Dung M Chu, Rita Cucchiara, Simone Calderara, Afshin Dehghan, and Mubarak Shah. Visual tracking: An experimental survey. *IEEE TPAMI*, 36(7):1442–1468, 2014.

[66] Xiaoning Song, Zhen-Hua Feng, Guosheng Hu, Josef Kittler, William Christmas, and Xiao-Jun Wu. Dictionary integration using 3d morphable face models for pose-invariant collaborative-representation-based classification. *IEEE TIFS*, 13(11):2734–2745, 2018.

[67] Yibing Song, Chao Ma, Lijun Gong, Jiawei Zhang, Rynson Lau, and Ming-Hsuan Yang. Crest: Convolutional residual learning for visual tracking. In *ICCV*, pages 2555–2564, 2017.

[68] Yibing Song, Chao Ma, Xiaohe Wu, Lijun Gong, Linchao Bao, Wangmeng Zuo, Chunhua Shen, Rynson Lau, and Ming-Hsuan Yang. Vital: Visual tracking via adversarial learning. *arXiv preprint arXiv:1804.04273*, 2018.

[69] Chong Sun, Dong Wang, Huchuan Lu, and Ming-Hsuan Yang. Learning spatial-aware regressions for visual tracking. In *CVPR*, pages 8962–8970, 2018.

[70] Christian Szegedy, Wei Liu, Yangqing Jia, et al. Going deeper with convolutions. In *CVPR*, pages 1–9, 2015.

[71] Ming Tang and Jiayi Feng. Multi-kernel correlation filter for visual tracking. In *ICCV*, pages 3038–3046, 2015.

[72] Ming Tang, Bin Yu, Fan Zhang, and Jinqiao Wang. High-speed tracking with multi-kernel correlation filters. In *CVPR*, pages 4874–4883, 2018.

[73] Jack Valmadre, Luca Bertinetto, João Henriques, Andrea Vedaldi, and Philip HS Torr. End-to-end representation learning for correlation filter based tracking. In *CVPR*, pages 5000–5008, 2017.

[74] Andrea Vedaldi and Karel Lenc. Matconvnet: Convolutional neural networks for matlab. In *ACM International Conference on Multimedia*, pages 689–692, 2015.

[75] Mengmeng Wang, Yong Liu, and Zeyi Huang. Large margin object tracking with circulant feature maps. In *CVPR*, pages 21–26, 2017.

[76] Naiyan Wang, Jianping Shi, Dit-Yan Yeung, and Jiaya Jia. Understanding and diagnosing visual tracking systems. In *ICCV*, pages 3101–3109, 2015.

[77] Qiang Wang, Zhu Teng, Junliang Xing, Jin Gao, Weiming Hu, and Stephen Maybank. Learning attentions: residual attentional siamese network for high performance online visual tracking. In *CVPR*, pages 4854–4863, 2018.

[78] Wenguan Wang and Jianbing Shen. Deep visual attention prediction. *IEEE TIP*, 27(5):2368–2378, 2017.

[79] Wenguan Wang, Jianbing Shen, Fatih Porikli, and Ruigang Yang. Semi-supervised video object segmentation with super-trajectories. *IEEE TPAMI*, 41(4):985–998, 2018.

[80] Joost Van De Weijer, Cordelia Schmid, Jakob Verbeek, and Diane Larlus. Learning color names for real-world applications. *IEEE TIP*, 18(7):1512–23, 2009.

[81] Yi Wu, Jongwoo Lim, and Ming Hsuan Yang. Online object tracking: A benchmark. In *CVPR*, pages 2411–2418, 2013.

[82] Yi Wu, Jongwoo Lim, and Ming-Hsuan Yang. Object tracking benchmark. *IEEE TPAMI*, 37(9):1834–1848, 2015.

[83] Tianyang Xu, Zhen-Hua Feng, Xiao-Jun Wu, and Josef Kittler. Learning adaptive discriminative correlation filters via temporal consistency preserving spatial feature selection for robust visual tracking. *IEEE TIP*, 2019.

[84] Yong Xu, David Zhang, Jian Yang, and Jing-Yu Yang. A two-phase test sample sparse representation method for use with face recognition. *IEEE TCSVT*, 21(9):1255–1262, 2011.

[85] Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006.

[86] Kaihua Zhang, Lei Zhang, Qingshan Liu, David Zhang, and Ming Hsuan Yang. Fast visual tracking via dense spatio-temporal context learning. In *ECCV*, pages 127–141, 2014.

[87] Mengdan Zhang, Qiang Wang, Junliang Xing, Jin Gao, Peixi Peng, Weiming Hu, and Steve Maybank. Visual tracking via spatially aligned correlation filters network. In *ECCV*, pages 469–485, 2018.

[88] Tianzhu Zhang, Adel Bibi, and Bernard Ghanem. In defense of sparse tracking: Circulant sparse tracker. In *CVPR*, pages 3880–3888, 2016.

[89] Tianzhu Zhang, Changsheng Xu, and Ming-Hsuan Yang. Multi-task correlation particle filter for robust object tracking. In *CVPR*, pages 4335–4343, 2017.

[90] Tianzhu Zhang, Changsheng Xu, and Ming-Hsuan Yang. Robust structural sparse tracking. *IEEE TPAMI*, 41(2):473–486, 2019.

[91] Zheng Zhu, Qiang Wang, Bo Li, Wei Wu, Junjie Yan, and Weiming Hu. Distractor-aware siamese networks for visual object tracking. In *ECCV*, pages 103–119, 2018.

[92] Wangmeng Zuo, Xiaohe Wu, Liang Lin, Lei Zhang, and Ming-Hsuan Yang. Learning support correlation filters for visual tracking. *arXiv preprint arXiv:1601.06032*, 2016.