Xu et al. *Visual Intelligence* (2023) 1:4
https://doi.org/10.1007/s44267-023-00002-1

Visual Intelligence

RESEARCH Open Access

Learning spatio-temporal discriminative model for affine subspace based visual object tracking

Tianyang Xu[1], Xue-Feng Zhu[1] and Xiao-Jun Wu[1*]

Abstract

Discriminative correlation filters (DCF) with powerful feature descriptors have proven to be very effective for advanced visual object tracking approaches. However, due to the fixed capacity in achieving discriminative learning, existing DCF trackers perform the filter training on a single template extracted by convolutional neural networks (CNN) or hand-crafted descriptors. Such single template learning cannot provide powerful discriminative filters with guaranteed validity under appearance variation. To pinpoint the structural relevance of spatio-temporal appearance to the filtering system, we propose a new tracking algorithm that incorporates the construction of the Grassmannian manifold learning in the DCF formulation. Our method constructs the model appearance within an online updated affine subspace. It enables joint discriminative learning in the origin and basis of the subspace, achieving enhanced discrimination and interpretability of the learned filters. In addition, to improve tracking efficiency, we adaptively integrate online incremental learning to update the obtained manifold. To this end, specific spatio-temporal appearance patterns are dynamically learned during tracking, highlighting relevant variations and alleviating the performance degrading impact of less discriminative representations from a single template. The experimental results obtained on several well-known datasets, i.e., OTB2013, OTB2015, UAV123, and VOT2018, demonstrate the merits of the proposed method and its superiority over the state-of-the-art trackers.

Keywords: Visual object tracking, Discriminative model, Affine subspace, Grassmannian manifold, Online tracking

1 Introduction

Visual object tracking is one of the most fundamental topics in pattern recognition and computer vision, which plays crucial roles in a wide range of visual intelligent systems, e.g., medical image analysis, human-computer interaction, transportation intelligence, and robotics. To consistently and accurately track an arbitrary object in unconstrained scenarios is very challenging due to deformable shape, changing aspect, and textural variations of the target. Considering existing advanced tracking algorithms, discriminative correlation filter (DCF-) based trackers [1] have exhibited promising performance in various benchmarks [2–4] and competitions such as the Visual Object Tracking (VOT) challenges [5, 6] and VisDrone [7, 8]. In general, the advantages of DCF include its spatial appearance model exploiting the circulant matrix structure [9] and efficient optimization in the frequency domain [10]. More recent innovations focus on scale detection [11], joint regularization [12], continuous domain mapping [13], multi-response fusion [14], etc.

The success of current advanced DCF trackers can be attributed to two main factors: spatial regularization and temporal fusion. Regarding the spatial regularization, as images and videos rectify the 2D planes from the view of a camera, the proposal of spatial regularization enables a direct improvement of the tracking performance by potentially endowing the learned classifiers with a specific attention mechanism, enhancing the model's discrim-

*Correspondence: wu_xiaojun@jiangnan.edu.cn
[1]School of Artificial Intelligence and Computer Science, Jiangnan University, Wuxi, 214122, China

ination by focusing on less ambiguous and background regions [15–17]. Considering the temporal fusion techniques, advanced DCF trackers highlight the online appearance clues by gathering more historical target information or constructing temporally consistent constraints on the discriminative learning stage [18–20]. To this end, the above-mentioned spatio-temporal model methodologies have received continuous attention in the visual tracking community, especially for the powerful deep learning representations developed in recent years [21–23].

However, from the geometry viewpoint, the current DCF paradigm extracts the discriminative information from independent training templates (points), without unifying the spatio-temporal appearance jointly. Specifically, the multi-channel feature representations from different frames obtained by a pre-trained convolutional neural network (CNN) are simply inputs to the DCF learning stage with a moving average. Therefore, the model capacity against appearance variation can only be guaranteed within a limited ℓ_2-norm ball around the training templates, impeding the generalization of the learned filters, as illustrated in Fig. 1. Motivated by this observation, we argue the necessity of constructing the appearance from independent points to spatio-temporal affine subspace. The relationships among multiple historical frames can be jointly considered in the affine subspace, effectively extending the model capacity. In our design, during the online tracking process, all the previous frames are collected to construct the affine subspace, consisting of an origin and a linear subspace. To mitigate the increased calculation complexity in obtaining the linear subspace when a large number of frames are involved, we employ the incremental learning technique to update the origin and the linear subspace online, resulting in efficient affine subspace learning and updating.

In addition to constructing the affine subspace to reflect the spatio-temporal appearance, we also propose to endow the DCF model with parsimony and consistency constraints. In principle, with the development of robust visual features, e.g., the Haar descriptor, histogram of oriented gradient (HOG), and convolutional blocks of deep architectures (AlexNet, VGGNet, ResNet) [22–24], the volume of the feature representations has witnessed a continuous swell. Accordingly, these high dimensional feature maps provide improved discriminative information to achieve better tracking performance in distinguishing the target from its corresponding surroundings. However, there exists inevitable redundancy and noise in these feature maps. Therefore, to highlight the relevance between deep feature representations and the discriminative learning task, we propose to regularize the learned filters to be sparse. In addition, temporal smoothness is also emphasized in the DCF learning objective to achieve consistency in filter training, improving the stability of the tracking model.

To combine the DCF learning paradigm in the constructed affine subspace, we use the origin and the basis of the subspace to train one main filter and multiple auxiliary filters. In principle, the filter learning process corresponding to the origin is similar to that in the standard DCF paradigm, where a moving average template is employed to train the classifier in the current frame. The novelty of our affine subspace DCF (ASDCF) learning approach emphasizes the design of learning auxiliary filters corresponding to the basis of the subspace. Specifically, after obtaining the basis of the current K dimensional subspace, we propose to train K separate auxiliary filters corresponding to the K basis representation. To this end, each auxiliary filter is associated with specific appearance variation, improving the capacity of the proposed learning model. The proposed ASDCF injects the spatio-temporal information represented by the deep features to the online updated affine subspace, unifying the spatial visual features and the changing temporal variations, with improved discrimination and interpretability compared with the standard DCF framework.

The Simaese-based trackers have recently achieved remarkable performance by learning to map the target template and instance into an appearance variation that has preserved feature space through an end-to-end network. However, the Siamese-based trackers conduct tracking by relying on a fixed template, and the appearance capacity is not modeled, resulting in performance that is especially dependent on the invariance of extracted features. In contrast, the constructed affine space of this work enhances the ability to model target appearance, increasing the tolerance of the learned model to spatio-temporal appearance variation of an object. Therefore, by combining the proposed affine space construction and updating with DCF learning, more accurate and stable target tracking results can be realized.

The main contributions of the proposed ASDCF tracking approach include the following:

1) A new affine subspace construction technique in online visual tracking to unify the spatial and temporal discriminative information, with an efficient incremental learning method to update the affine subspace during tracking.

2) An effective DCF learning objective imposing sparsity and temporal smoothness regularization for the filters.

3) A comprehensive evaluation of ASDCF on several well-known public available benchmarking datasets, including OTB2013 [2], OTB2015 [3], UAV123 [4], and VOT2018 [6]. The results support the advantage of the proposed ASDCF, with superior tracking performance compared with the state-of-the-art trackers.

The rest of this paper is organized as follows. In Sect. 2, we briefly review relevant tracking approaches for constructing spatio-temporal appearance models, especially

the development of the DCF framework. The proposed affine subspace construction is presented in Sect. 3. The details of the proposed ASDCF method are introduced in Sect. 4, accompanied by an efficient optimization scheme. The implementation details and experimental results are reported in Sect. 5, with ablation studies and comparative analysis. Conclusions are presented in Sect. 6.

2 Related work

Existing visual object tracking approaches include generative learning and discriminative learning, e.g., image matching [25], statistical theory [26], particle filtering framework [27], subspace learning methodology [28], discriminative correlation filters [1], and deep neural networks [29]. In this section, we focus on introducing the development of the above-mentioned tracking approaches that are pertinent to our ASDCF. Continuously improved tracking performance has been evidenced by recent tracking benchmarking datasets and competitions such as VOT [6]. Readers are recommended to refer to recent surveys [3, 30–32] for detailed and comprehensive reviews of the visual tracking approaches.

2.1 Generative learning framework

Generative learning frameworks aim at learning the intrinsic target state distribution to represent the target appearance, based on which similarity metric or reconstruction error can be employed to calculate the final probabilities for the candidates in the next frame. Typical generative learning models in the early visual tracking research stage include optical flow [25] and mean-shift [33]. The basic assumptions behind these two methodologies are consistent brightness and limited appearance variations. Although these two methods provide complete mathematical derivations to model the visual tracking task, their rigid constraints cannot satisfy the real-world scenarios, resulting in poor tracking performance when processing challenging videos. To enhance the tracking robustness, the particle filtering system is applied to visual tracking [27, 34] to estimate the posterior distribution of the target via Bayes's theorem and sampling techniques. Specifically, the conditional distribution is approximated via the similarity between the current samples and the model distribution, providing nonlinear inference for the tracking scope. It should be noted that improved performance can be achieved with the increasing number of involved particles while sacrificing the model efficiency. Due to the convenience that the particle filtering system is an external predicting framework, it has been widely studied and extended to fuse with other generative methods, e.g., sparse subspace representations and low-rank representations [35–37]. In principle, the subspace-based tracking paradigm has received wide attention since the proposal of the incremental subspace learning scheme [28], which assumes that the target can be linearly represented by its corresponding eigenvectors. Sparse trackers assume the target to be sparsely represented by an over-complete dictionary. Accordingly, the representation coefficients and reconstruction errors are used to gauge the quality of candidates. Furthermore, low-rank constraints have been proposed to increase the relevance of particles by suppressing spurious information [37].

The advantage of generative learning framework focuses on its exploration of enlarging the tracking model capacity via carefully designed appearance representation and inference systems. However, generative tracking methods suffer from the limitation of neglecting the background appearance, resulting in less discriminative performance.

2.2 Discriminative learning framework

In addition to generative learning methods, various classification methods, such as support vector machine [26], multiple instance boosting [38], and linear regression [10] have been employed in constructing learning models in a discriminative manner, exploring the discriminative information between the target region and its surroundings. Discriminative learning approaches construct a tracking task as a classification or regression problem, aiming at directly inferring the output of a sampling candidate by estimating the conditional distribution of labels for the given inputs. Therefore, the optimal sampling candidate with the maximal response is selected as the final tracking result. However, a common limitation of the above discriminative trackers is that the initialization of the learning model is performed in the initial frame with insufficient appearance information, without guaranteed tracking robustness for the following frames. More recently, Siamese networks [29, 39–41] have been successfully applied in visual tracking. Taking the advantage of large annotated tracking datasets, deep architectures and powerful graphical processing units, Siamese networks achieve efficient visual tracking by performing efficient template matching in the learned feature embedding space.

Compared with basic generative learning approaches, discriminative methods developed a comparatively more robust modeling paradigm that extracts and analyzes appearance from both foreground and background, achieving better tracking performance.

2.3 Discriminative correlation filter

DCF belongs to the discriminative learning paradigm, and we provide detailed instructions of its development in this subsection as it is the baseline of our proposed ASDCF. The seminal work of the DCF framework is minimum output sum of squared error (MOSSE) [42], which formulates the tracking task as discriminative filter learning [43] rather than template matching [44], achieving improved

tracking efficiency. Based on this modeling technique, the concept of circulant matrix [9] is introduced to DCF by CSK [10] with an enlarged search window, enabling the generation of more negative training samples in the discriminative filter learning stage. To further explore the potential of the DCF framework, spatial-temporal context information [45] and kernel modeling technique [1] are leveraged to improve the learning formulation by involving local appearance and nonlinear metrics, respectively. In recent years, the DCF paradigm has further been extended by exploiting scale detection [46, 47], structural patch analysis [48, 49], multi-clue fusion [14, 50, 51], sparse representation [36, 52, 53], support vector machine [54, 55], enhanced sampling mechanisms [56, 57] and end-to-end deep neural networks [29, 40, 58].

Despite the outstanding performance of the DCF framework in visual object tracking, it is still a very challenging task to achieve high-performance tracking for a spatio-temporal changing arbitrary object, especially in unconstrained scenarios. The main obstacles include spatial bounding effect and temporal inconsistency. To alleviate the boundary effect problem caused by the circulant structure, SRDCF [15] proposes introducing spatial regularization in the DCF formulation, which allocates more filter energy for the central region and less energy for the surroundings using a pre-defined spatial smooth weighting function. A similar technique has been pursued by pruning the training samples or learned filters with pre-defined binary mask [16, 59–62]. To achieve adaptive spatial regularization, LADCF [63] embeds dynamic spatial feature selection in the filter learning stage, activating the supportive spatial regions not only from the foreground but also from the background. Similarly, A^{3}DCF [64] proposes an adaptive attribute-aware mechanism to learn channel-wise masks to enhance discriminative elements of feature maps while suppressing irrelevant features. ADTrack [65] adopts image pre-treatment to achieve mask generation for discriminative filter learning. The above spatial regularization approaches decrease the ambiguity emanating from the background and enable a relatively enlarged search window for DCF tracking. However, these approaches only consider information redundancy and unbalance along the spatial dimension. On the other hand, to mitigate temporal filter inconsistency, historical appearance information is rearranged in SRDCFdecon [18] and C-COT [13], with enhanced robustness and temporal stability, by gathering multiple previous frames in the filter learning stage. In addition, to alleviate the computational burden caused by involving a large number of historical samples, ECO [20] decreases the inherent computational complexity by clustering historical frames in a generative sample space and employing projection matrix to reduce the channel numbers for the feature representations.

To advance the DCF modeling space, we introduce the affine subspace to enlarge the representative power for the potential appearance variations from a geometric viewpoint. Therefore, performing discriminative modeling in the affine subspace can unify the spatial and temporal discriminative information, enhancing the DCF capacity for challenging video sequences.

3 Affine subspace generation

To accommodate appearance variations for spatio-temporal changing objects, we propose to employ affine subspace to represent both static and dynamic information. An affine subspace can be formulated as:

$$\mathcal{A} = \left\{ \boldsymbol{x} \in \mathbb{R}^D : \boldsymbol{x} = \boldsymbol{\mu} + \boldsymbol{U}\boldsymbol{z} \right\}, \tag{1}$$

where $\boldsymbol{\mu} \in \mathbb{R}^D$ denotes the origin of the affine subspace, and $\boldsymbol{U}$ ($\boldsymbol{U} = [\boldsymbol{u}_1, \boldsymbol{u}_2 \ldots]$) represents the corresponding basis of the subspace, as depicted in Fig. 1. Based on the formulation in Eq. (1), we can gather the historical appearance with an updated affine subspace, realizing extended representation capacity compared with a single template. Specifically, the origin $\boldsymbol{\mu}$ reflects the weighted average static appearance from all the previous frames, while the basis $\boldsymbol{U}$ constructs the detailed variations during the tracking process. Here, the basis is obtained by calculating the dominant K eigenvectors of the subspace based on singular value decomposition (SVD).

It should be noted that the affine subspace has to be updated once a new frame is available, resulting in an increasing burden for the SVD calculation. Therefore, the computational burden would explode if hundreds of high-dimensional representations were involved in the affine subspace construction. To mitigate this issue, we propose to introduce an incremental learning technique to achieve efficient updates for the origin $\boldsymbol{\mu}$ and the basis $\boldsymbol{U}$. Given a data matrix $\boldsymbol{A} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n] \in \mathbb{R}^{D\times n}$, where each column $\boldsymbol{x}_i$ denotes the appearance gathered from the ith frame.

Update the origin: Suppose we have already obtained the mean vector of $\boldsymbol{A}$ as $\boldsymbol{\mu}_A$. When the appearance representations from new m frames are available, denoted as $\boldsymbol{B} = [\boldsymbol{x}_{n+1}, \boldsymbol{x}_{n+2}, \ldots, \boldsymbol{x}_{n+m}] \in \mathbb{R}^{D\times m}$, the aim is to incrementally calculate the mean vector of the new data matrix $[\boldsymbol{A}\ \boldsymbol{B}]$. Denoting the mean vector of $\boldsymbol{B}$ as $\boldsymbol{\mu}_B$, the updated origin for the affine subspace expanded by $[\boldsymbol{A}\ \boldsymbol{B}]$ can be calculated as:

$$\boldsymbol{\mu} = \frac{n}{m+n}\boldsymbol{\mu}_A + \frac{m}{m+n}\boldsymbol{\mu}_B. \tag{2}$$

Update the basis: Suppose we have already obtained the SVD of $\boldsymbol{A}$ as $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$. When the appearance representations from the new m frames are available, denoted as $\boldsymbol{B} = [\boldsymbol{x}_{n+1}, \boldsymbol{x}_{n+2}, \ldots, \boldsymbol{x}_{n+m}] \in \mathbb{R}^{D\times m}$, the aim is to incrementally calculate the SVD results of the new data matrix $[\boldsymbol{A}\ \boldsymbol{B}]$

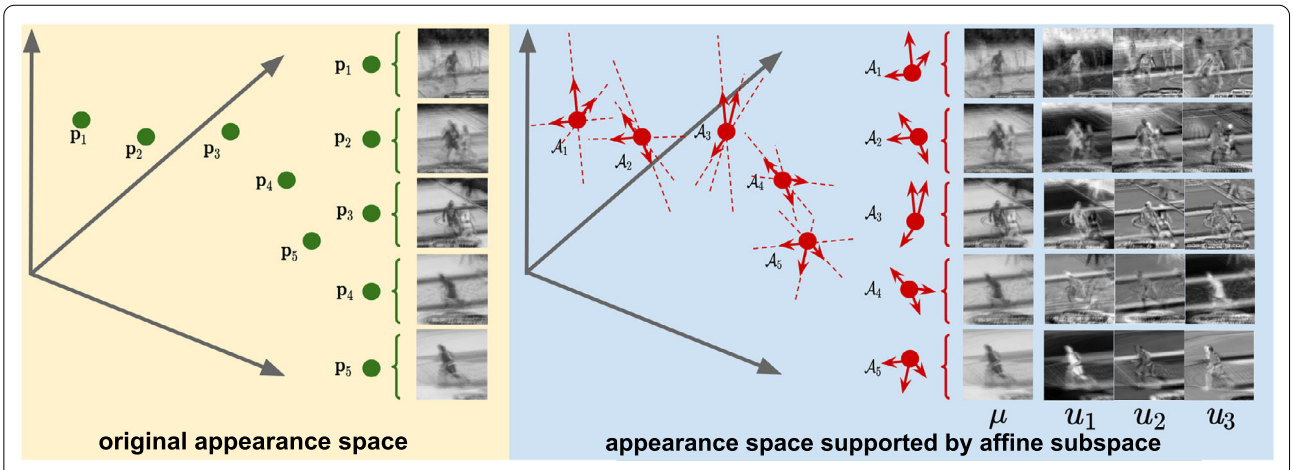

Figure 1 Illustration of the proposed appearance learning space that is supported by affine subspace. *left*: The appearance of a single template point is used to train the filter in each frame node. *right*: The appearance generated by the affine subspace is used to train one main filter (from origin $\boldsymbol{\mu}$) and K auxiliary filters (from basis $\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_K$) to provide improved generalization ($K = 3$)

as $[\boldsymbol{A}\ \boldsymbol{B}] = \boldsymbol{U}'\boldsymbol{\Sigma}'\boldsymbol{V}'^{\top}$. Denoting the component of $\boldsymbol{B}$ that is orthogonal to $\boldsymbol{U}$ as $\tilde{\boldsymbol{B}}$, such that the SVD of $[\boldsymbol{A}\ \boldsymbol{B}]$ can be partitioned as follows:

$$\begin{bmatrix}\boldsymbol{A} & \boldsymbol{B}\end{bmatrix} = \begin{bmatrix}\boldsymbol{U} & \tilde{\boldsymbol{B}}\end{bmatrix} \boldsymbol{R} \begin{bmatrix}\boldsymbol{V}^{\top} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I}\end{bmatrix}, \tag{3}$$

where $\boldsymbol{R} = \begin{bmatrix}\boldsymbol{\Sigma} & \boldsymbol{U}^{\top}\boldsymbol{B} \\ \boldsymbol{0} & \tilde{\boldsymbol{B}}^{\top}\boldsymbol{B}\end{bmatrix}$. To balance the tracking efficiency and effectiveness, we retain the first K eigenvectors in $\boldsymbol{U}$. Considering the size of the new frames, m, the SVD of $\boldsymbol{R}$ can be calculated in constant time regardless of the number of frames in $\boldsymbol{A}$, $\boldsymbol{R} = \tilde{\boldsymbol{U}}\tilde{\boldsymbol{\Sigma}}\tilde{\boldsymbol{V}}^{\top}$. Therefore, Eq. (3) can be formulated as:

$$\begin{bmatrix}\boldsymbol{A} & \boldsymbol{B}\end{bmatrix} = \begin{bmatrix}\boldsymbol{U} & \tilde{\boldsymbol{B}}\end{bmatrix} \tilde{\boldsymbol{U}}\tilde{\boldsymbol{\Sigma}}\tilde{\boldsymbol{V}}^{\top} \begin{bmatrix}\boldsymbol{V}^{\top} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I}\end{bmatrix}. \tag{4}$$

Based on Eq. (4), the final eigenvectors $\boldsymbol{U}'$ can be obtained as $\boldsymbol{U}' = [\boldsymbol{U}\ \tilde{\boldsymbol{B}}]\tilde{\boldsymbol{U}}$. The corresponding eigenvalues $\boldsymbol{\Sigma}' = \tilde{\boldsymbol{\Sigma}}$. After obtaining $\boldsymbol{U}'$, we retain the first K eigenvectors in $\boldsymbol{U}'$ to represent the basis of the updated affine subspace.

4 Approach

4.1 Basic discriminative correlation filter

Given the location and scale of a target at frame t, visual object tracking aims at predicting the location of the target in the next frame. In the learning stage, we aim to train a discriminative filter that obtains high-value responses around the target center and low-value responses for the background. DCF is formulated to learn a filter that distinguishes the target from the near background. In general, a padded search window centered around the target location from frame t is extracted with corresponding feature representation $\boldsymbol{x} = [x_1, x_2, \ldots, x_n]^{\top} \in \mathbb{R}^D$. The circulant matrix can be generated as [9]:

$$\boldsymbol{X} = \begin{bmatrix} x_1 & x_2 & x_3 & \ldots & x_D \\ x_D & x_1 & x_2 & \ldots & x_{D-1} \\ x_{D-1} & x_D & x_1 & \ldots & x_{D-2} \\ \vdots & & \ddots & & \vdots \\ x_2 & x_3 & x_4 & \ldots & x_1 \end{bmatrix}, \tag{5}$$

where each row in $\boldsymbol{X}$ can be considered an augmented sample, and therefore the DCF formulation employs $\boldsymbol{X}$ as the training data matrix [10]. Given labeled training sample pairs $\{\boldsymbol{X}, \boldsymbol{y}\}$, the learning stage of DCF is formulated as a ridge regression problem:

$$\begin{aligned}\boldsymbol{w} &= \arg\min_{\boldsymbol{w}} \|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|^2 + \lambda\|\boldsymbol{w}\|^2 \\ &= \arg\min_{\boldsymbol{w}} \|\boldsymbol{x} * \boldsymbol{w} - \boldsymbol{y}\|^2 + \lambda\|\boldsymbol{w}\|^2,\end{aligned} \tag{6}$$

where λ is the balancing parameter for the regularization term, and $*$ denotes the cross correlation operator. According to the time-frequency convolution theorem, a closed-form solution in the frequency domain can be obtained as:

$$\hat{\boldsymbol{w}} = \frac{\hat{\boldsymbol{x}} \odot \hat{\boldsymbol{y}}^*}{\hat{\boldsymbol{x}} \odot \hat{\boldsymbol{x}}^* + \lambda \boldsymbol{1}}, \tag{7}$$

where $\odot$ denotes the element-wise multiplication, $\boldsymbol{1}$ is an all-ones vector sharing the same size with $\hat{\boldsymbol{x}}$, $\hat{\cdot}$ denotes discrete fourier transform (DFT) representation and $\cdot^*$ represents the complex conjugate.

4.2 Sparse discriminative correlation filter

Though promising tracking results have been achieved by the basic DCF formulations, the impact of the redundancy

and noise in the high dimensional feature representations is not well addressed, especially in the feature maps extracted from deep CNN architectures, e.g., AlexNet, VGGNet, and ResNet. To this end, with the aim of highlighting the relevance between deep feature representations and the discriminative learning task, we propose to regularize the learned filters to be sparse. In addition, temporal smoothness is also emphasized in the proposed DCF learning objective to achieve consistency in filter training, improving the stability of the tracking model. In principle, the filter learning objective is formulated as follows:

$$\begin{aligned}\boldsymbol{w} = \arg\min_{\boldsymbol{w}} \|\boldsymbol{x} * \boldsymbol{w} - \boldsymbol{y}\|^2 + \lambda_1 \|\boldsymbol{w}\|_1 \\ + \lambda_2 \|\boldsymbol{w} - \boldsymbol{w}_{t-1}\|^2,\end{aligned} \tag{8}$$

where $\|\cdot\|_1$ denotes the ℓ_1-norm, λ_1 and λ_2 are the corresponding balancing parameters for the sparse regularization and temporal smoothness terms, respectively. Based on the formulation in Eq. (8), the model parsimony can be achieved for high-dimensional feature representations with temporally enforced stability.

4.3 Affine subspace discriminative correlation filters

In the implementation, multi-channel feature maps from CNN are used to enhance the representation power, and we transform the objective in Eq. (8) from the single-channel to multi-channel formulation as follows:

$$\begin{aligned}\boldsymbol{w} = \arg\min_{\boldsymbol{w}} \left\| \sum_{c=1}^{C} \boldsymbol{x}^c * \boldsymbol{w}^c - \boldsymbol{y} \right\|^2 + \lambda_1 \sum_{c=1}^{C} \left\| \boldsymbol{w}^c \right\|_1 \\ + \lambda_2 \sum_{c=1}^{C} \left\| \boldsymbol{w}^c - \boldsymbol{w}_{t-1}^c \right\|^2.\end{aligned} \tag{9}$$

In general, the origin contains the global static appearance of the target, while each eigenvector in the basis focuses on specific variation during the past tracking frames. To perform DCF learning in the affine subspace, we propose to learn the discriminative filters for the origin and the basis separately. Specifically, in frame t, the current affine subspace $\mathcal{A}_t$ can be represented by the origin $\boldsymbol{\mu}$ and the K eigenvectors in $\boldsymbol{U}$, $\{\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_K\}$. We consider the $K+1$ vectors, $\{\boldsymbol{\mu}_t, \boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_K\}$, as our training data to train one main filter $\boldsymbol{w}_{\boldsymbol{\mu}}$ and K auxiliary filters $\{\boldsymbol{w}_{\boldsymbol{u1}}, \boldsymbol{w}_{\boldsymbol{u2}}, \ldots, \boldsymbol{w}_{\boldsymbol{uK}}\}$ based on Eq. (9). As presented in Sect. 3, the proposed affine space can represent both static and dynamic information of the target and is updated by incremental learning. In this way, the spatio-temporal information of target appearance variations in tracking can be well modeled. Through further sparse DCF learning framework, this enhanced representation of target appearance served for tracking model learning which leads to better tracking performance, especially in challenging situations.

4.4 Optimization

According to the convexity of the proposed formulation in Eq. (9), we employ the augmented Lagrange method to optimize the problem. Here, we introduce a slack variable $\boldsymbol{w}' = \boldsymbol{w}$ for the estimate. The Lagrange function can be expressed as follows:

$$\begin{aligned}\mathcal{L} = \left\| \sum_{c=1}^{C} \boldsymbol{x}^c * \boldsymbol{w}^c - \boldsymbol{y} \right\|^2 + \lambda_1 \sum_{c=1}^{C} \left\| \boldsymbol{w}'^c \right\|_1 \\ + \lambda_2 \sum_{c=1}^{C} \left\| \boldsymbol{w}^c - \boldsymbol{w}_{t-1}^c \right\|^2 \\ + \frac{\nu}{2} \sum_{c=1}^{C} \left\| \boldsymbol{w}^c - \boldsymbol{w}'^c + \frac{\boldsymbol{\gamma}^c}{\nu} \right\|^2,\end{aligned} \tag{10}$$

where $\boldsymbol{\gamma}$ is the Lagrange multiplier with the same size as $\boldsymbol{x}$, and ν is the corresponding penalty parameter for the slack variable $\boldsymbol{w}'$. We exploit the alternating direction method of multipliers [66] approach to iteratively optimize the following sub-problems:

$$\begin{cases} \boldsymbol{w} = \arg\min_{\boldsymbol{w}} \mathcal{L}(\boldsymbol{w}, \boldsymbol{w}', \boldsymbol{\gamma}, \nu), \\ \boldsymbol{w}' = \arg\min_{\boldsymbol{w}'} \mathcal{L}(\boldsymbol{w}, \boldsymbol{w}', \boldsymbol{\gamma}, \nu), \\ \boldsymbol{\gamma} = \arg\min_{\boldsymbol{\gamma}} \mathcal{L}(\boldsymbol{w}, \boldsymbol{w}', \boldsymbol{\gamma}, \nu). \end{cases} \tag{11}$$

4.4.1 Optimizing $\boldsymbol{w}$

To optimize $\boldsymbol{w}$, we exploit the circulant structure [1] and Parseval's theorem to transfer the sub-problems from the original spatial domain to the frequency domain:

$$\begin{aligned}\min \left\| \sum_{c=1}^{C} \hat{\boldsymbol{x}}^c \odot \hat{\boldsymbol{w}}^c - \hat{\boldsymbol{y}} \right\|^2 + \lambda_2 \sum_{c=1}^{C} \left\| \hat{\boldsymbol{w}}^c - \hat{\boldsymbol{w}}_{t-1}^k \right\|^2 \\ + \frac{\nu}{2} \sum_{c=1}^{C} \left\| \hat{\boldsymbol{w}}^c - \hat{\boldsymbol{w}}'^c + \frac{\hat{\boldsymbol{\gamma}}^c}{\nu} \right\|^2.\end{aligned} \tag{12}$$

A closed-form solution for the above sub-problem can be obtained as [67]:

$$\hat{\boldsymbol{w}}_i = \left(\boldsymbol{I} - \frac{\hat{\boldsymbol{x}}_i \hat{\boldsymbol{x}}_i^{\top}}{\lambda_2 + \nu/2 + \hat{\boldsymbol{x}}_i^{\top} \hat{\boldsymbol{x}}_i} \right) \boldsymbol{g}, \tag{13}$$

where $\boldsymbol{g} = (\hat{\boldsymbol{x}}_i \hat{y}_i + \nu \hat{\boldsymbol{w}}_i' - \nu \hat{\boldsymbol{\gamma}}_i + \lambda_2 \hat{\boldsymbol{w}}_{t-1\,i})/(\lambda_2 + \nu)$, the vectors $\hat{\boldsymbol{w}}_i$ ($\hat{\boldsymbol{w}}_i = [\hat{w}_i^1, \hat{w}_i^2, \ldots, \hat{w}_i^C] \in \mathbb{C}^C$), $\hat{\boldsymbol{x}}_i$, and $\hat{\boldsymbol{w}}_{t-1\,i}$ denote the i-th units of $\hat{\boldsymbol{w}}$, $\hat{\boldsymbol{x}}$ and $\hat{\boldsymbol{w}_{t-1}}$, respectively, across all C channels, and $i \in \{1, 2, \ldots, D\}$.

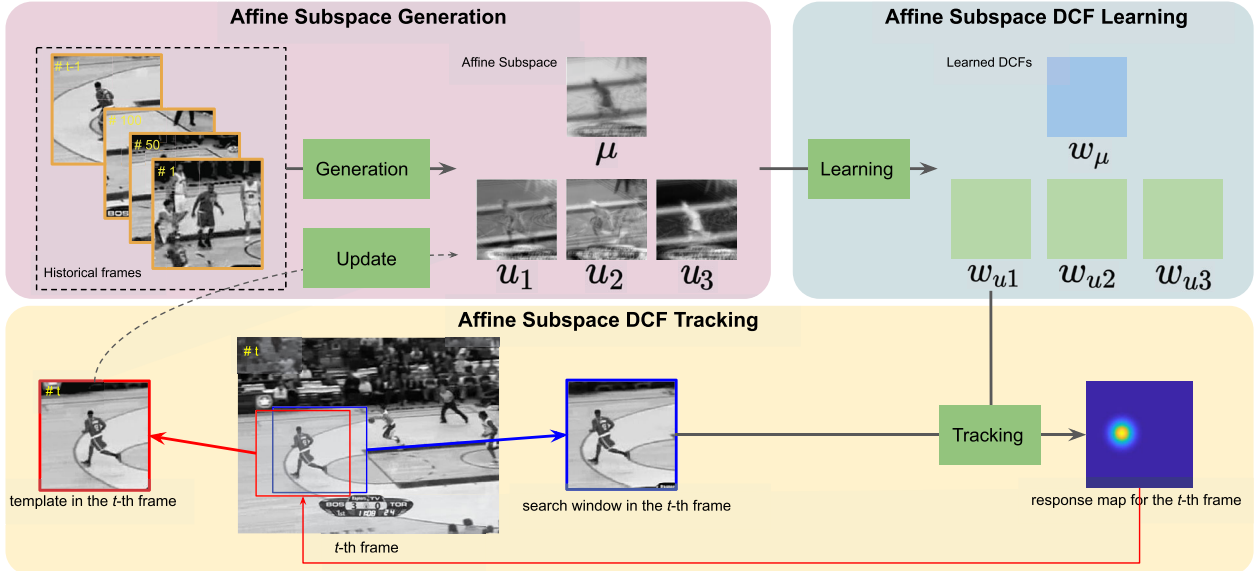

Figure 2 Overview of the proposed ASDCF. Affine subspace generation and online updating are introduced in Sect. 3. Details of the affine sparse DCF formulation and optimization are presented in Sect. 4.3 and Sect. 4.4. We discuss the online tracking stage of our ASDCF in Sect. 4.5 ($K = 3$ in the illustration)

4.4.2 *Optimizing* $\boldsymbol{w}'$

To optimize $\boldsymbol{w}'$, we need to minimize the following sub-problem:

$$\min \lambda_1 \sum_{c=1}^{C} \left\| \boldsymbol{w}'^c \right\|_1 + \frac{\nu}{2} \sum_{c=1}^{C} \left\| \boldsymbol{w}^c - \boldsymbol{w}'^c + \frac{\boldsymbol{\gamma}^c}{\nu} \right\|^2. \tag{14}$$

The soft-threshold shrinkage operator is used here to form a closed-form solution for each element $w_i'^c$ in the vector $\boldsymbol{w}'$ separately:

$$w_i'^c = \text{sign}(p) \max\left(0, |p| - \frac{\lambda_1}{\nu}\right), \tag{15}$$

where $p = w_i^c + \frac{\gamma_i^c}{\nu}$, with w_i^c and γ_i^c being the values corresponding to the elements at the i-th spatial unit and c-th channel in $\boldsymbol{w}$ and $\boldsymbol{\gamma}$, respectively.

4.4.3 *Optimizing multiplier* $\boldsymbol{\gamma}$ *and penalty* ν

The multiplier $\boldsymbol{\gamma}$ and the penalty ν are updated at the end of each iteration as:

$$\begin{cases} \boldsymbol{\gamma} = \boldsymbol{\gamma} + \nu(\boldsymbol{w} - \boldsymbol{w}'), \\ \nu = \min(\rho\nu, \nu_{\max}), \end{cases} \tag{16}$$

where ρ is the parameter that controls the strictness of the penalty and $\nu_{\max}$ is the corresponding upper threshold.

4.5 ASDCF algorithm

We summarize our ASDCF in detail in two stages, i.e, tracking and learning.

4.5.1 *Tracking stage*

As shown in Fig. 2, given a new image in frame t and the predicted target state of frame $t-1$ (target center p_{t-1}, the target width, w_{t-1}, and height h_{t-1}), we extract a search window $\{\boldsymbol{I}\}$ centered around p_{t-1}. The search window patch is of $n' \times n'$ pixels. We re-size the patch to the $n \times n$ basic search window size. n' is determined by the target size $w_{t-1} \times h_{t-1}$ and the padding parameter, ϱ as: $n' = (1+\varrho)\sqrt{w_{t-1} \times h_{t-1}}$. Then we extract multi-channel features of the search window as $\boldsymbol{x} \in \mathbb{R}^{D \times C}$. Given the filter model obtained from the previous frame, one main filter $\boldsymbol{w_\mu}$ and K auxiliary filters $\{\boldsymbol{w_{u1}}, \boldsymbol{w_{u2}}, \ldots, \boldsymbol{w_{uK}}\}$, the response map $\boldsymbol{y}$ can efficiently be calculated in the frequency domain as:

$$\hat{\boldsymbol{y}} = \sum_{c=1}^{C} \hat{\boldsymbol{x}}^c \odot \hat{\boldsymbol{w}}_{\boldsymbol{\mu}}^c + \lambda_3 \sum_{k=1}^{K} \sum_{c=1}^{C} \left(\hat{\boldsymbol{x}}^c - \hat{\boldsymbol{\mu}}^c\right) \odot \hat{\boldsymbol{w}}_{\boldsymbol{uk}}^c, \tag{17}$$

where λ_3 is a balancing parameter. The new position corresponds to the maximal value in the response maps $\boldsymbol{y}$.

4.5.2 *Learning stage*

To balance the accuracy and efficiency, our tracker performs filter training every 5 frames. In the filter learning stage, we first extract the 5 feature representations,

$\{\boldsymbol{x}_{t-4}, \boldsymbol{x}_{t-3}, \ldots, \boldsymbol{x}_t\}$ of the target appearance from frame $t-4$ to frame t based on the tracking results. Then the affine subspace $\mathcal{A}$ is updated according to Sect. 3. After obtaining $\mathcal{A}$, the main filter $\boldsymbol{w}_\mu$ and K auxiliary filters $\{\boldsymbol{w}_{u1}, \boldsymbol{w}_{u2}, \ldots, \boldsymbol{w}_{uK}\}$ are trained according to Eq. (10)-Eq. (16).

5 Evaluation

5.1 Implementation

To evaluate the performance of the proposed ASDCF, we implement the tracking algorithm in the MATLAB platform on an Intel i7 2.20 GHz CPU with an Nvidia GTX 1050Ti GPU. The detailed settings for the parameters used in Sect. 4.5 are as follows. The number of auxiliary filters $K = 3$, corresponding to the number of eigenvectors we use to represent the subspace. We set the basic window size $n \times n = 240 \times 240$ pixels, the padding parameter $\varrho = 4$. We equip the proposed ASDCF with both hand-crafted and deep CNN features. The hand-crafted set includes HOG and color names (CN) features, with 4 pixel cell size, $\lambda_1 = 10^{-5}$, $\lambda_2 = 30$, and $\lambda_3 = 0.3$. Specifically, the HOG (31 channels) and CN (10 channels) features are concatenated along the channel dimension to obtain the final hand-crafted feature representation $\boldsymbol{x} \in \mathbb{R}^{3600\times 41}$. We use ResNet-50 (the output of layer 3) to extract deep feature representations using the MatConvNet toolbox[1] [68]. The regularization parameters $\lambda_1 = 10^{-6}$, $\lambda_2 = 5$, and $\lambda_3 = 0.2$. The dimensionality of the ResNet-50 feature representation is $\boldsymbol{x} \in \mathbb{R}^{225\times 1024}$.

5.2 Evaluation metrics

We perform an experimental evaluation on 4 challenging benchmarks: OTB2013 [2], OTB2015 [3], UAV123 [4], and VOT2018 [6]. For OTB2013, OTB2015, and UAV123, we employ precision plots and success plots to measure the tracking performance [2]. The precision plot indicates the proportion of frames with the distance between the tracking results and the ground truth less than a certain number of pixels. The distance precision (DP) is defined by the corresponding value when the precision threshold is 20 pixels. Center location error (CLE) measures the mean distance between the centers of the tracking results and the ground truth values. The success plot describes the percentage of successful frames with a threshold ranging from 0 to 1. The target in a frame is considered successfully tracked if the overlap of the two bounding boxes exceeds a given threshold. The overlap precision (OP) is defined by the corresponding value when the overlap threshold is 0.5. The area under the curve (AUC) of the success plot quantifies the result in terms of overlap evaluation. For VOT2018, we use the expected average overlap (EAO), accuracy and robustness metrics for performance evaluation [69].

[1] http://www.vlfeat.org/matconvnet/

We compare our method against recent state-of-the-art tracking approaches, including A^{3}DCF [64], KYS [70], ASRCF [71], VITAL [72], STRCF [19], ECO [20], C-COT [13], MCPF [56], MetaTracker [73], CREST [74], BACF [59], CACF [57], ACFN [75], CSRDCF [16], Staple [14], SiamFC [76], CFNet [40], SRDCF [15], DSST [47] and KCF [1]. For VOT2018, we compare our ASDCF with the top trackers in VOT2018, *i.e.*, ECO, CFCF [77], UPDT [78], SiamRPN [58], LADCF [63], ULAST [79] and FCOS_MAML [80].

5.3 Ablation studies

The proposed ASDCF aims at improving discrimination by explicitly modeling the spatio-temporal appearance in an online updated affine subspace. In addition, spatial sparsity and temporal smoothness are also fused in the DCF formulation, decreasing the redundancy and noise from the high dimensional feature representations. Therefore, the ablation studies are conducted to verify the effectiveness of performing DCF learning in the affine subspace.

The corresponding results are reported in Table 1. According to Table 1, introducing the affine subspace ($K > 0$) in the DCF framework improves the tracking performance compared with single template learning ($K = 0$). The performance witnesses a continuous improvement when increasing the number of auxiliary filters until $K = 3$. Then, slight performance degradation can be observed at $K = 4$ and $K = 5$. The above results indicate that the model capacity in the affine subspace can be enhanced before saturation, reflecting the effectiveness of the model in terms of the appearance variation in the affine subspace. In addition, the best performance is achieved with 3 auxiliary filters in the tracking system, with the improvement from 90.8% to 92.7% in terms of DP, and from 67.3% to 69.7% in terms of AUC. Ablation studies demonstrate the merits of performing DCF in the updated affine subspace, as well as the necessity of considering appearance variation with explicit modeling techniques during the online tracking system.

5.4 Comparison with state-of-the-art methods

5.4.1 Quantitative performance

First, we report the precision plots and success plots on OTB2013 and OTB2015 in Fig. 3, with the numerical DP

Table 1 Ablation performance on OTB2015 with/without affine subspace and the impact of using different numbers of auxiliary filters

Affine Subspace	Auxiliary Filters	DP	AUC
×	$K = 0$	90.8%	67.3%
√	$K = 1$	91.6%	68.0%
√	$K = 2$	92.2%	69.3%
√	$K = 3$	**92.7**%	**69.7**%
√	$K = 4$	91.9%	68.7%
√	$K = 5$	91.3%	68.1%

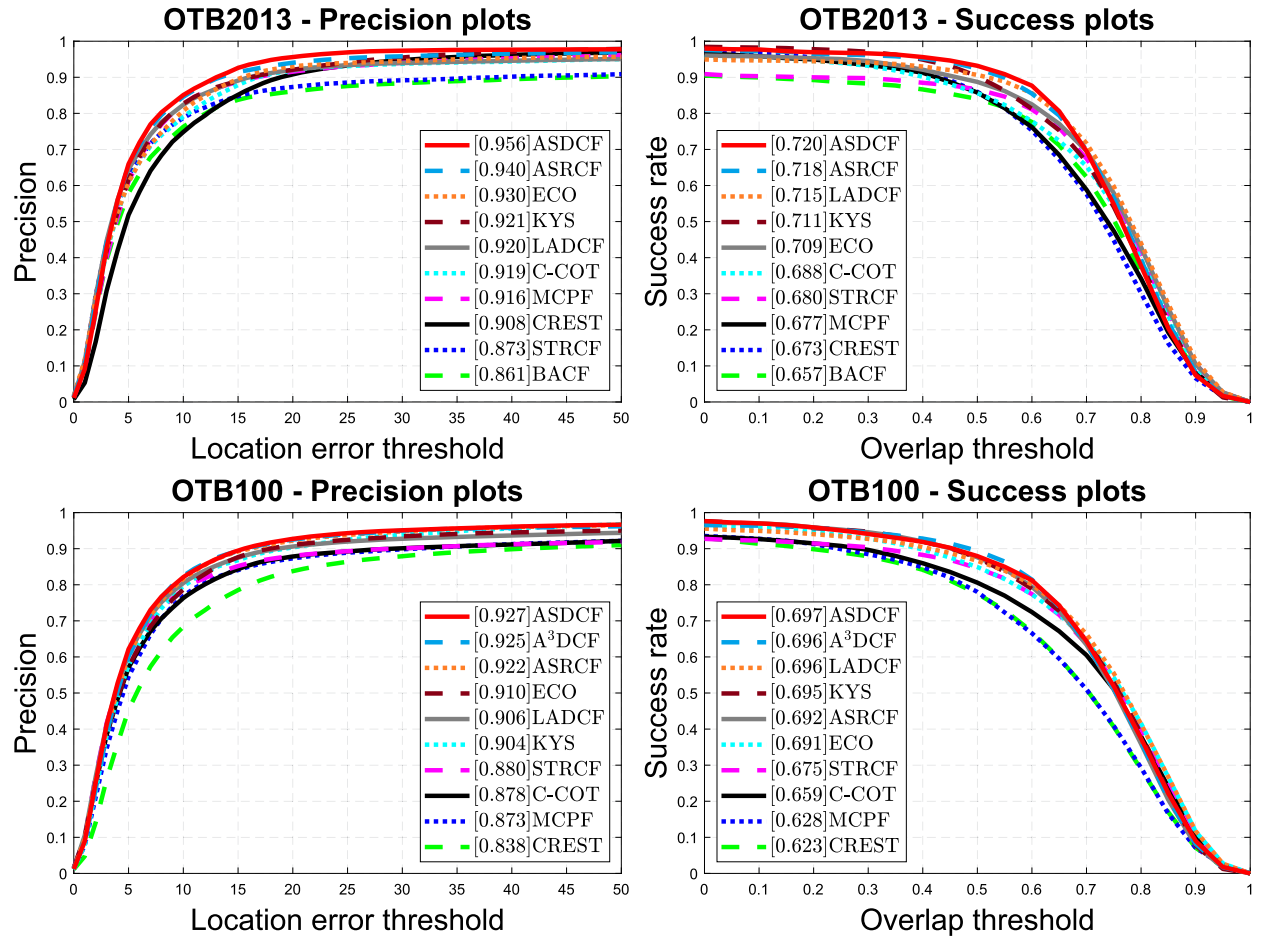

Figure 3 The experimental performance on OTB2013 and OTB2015. Precision plots (with the DP score reported in the figure legend) and the success plots (with the AUC score reported in the figure legend) are presented. Only the top ten trackers are presented for each metric

and AUC scores reported in the corresponding legends, respectively. Based on the result curves, ASDCF exhibits superior performance against the state-of-the-art trackers in both cases. On OTB2013, ASDCF achieves promising tracking results with 95.6% in DP. Compared to ECO and LADCF, which can be considered the best of a class of DCF-based trackers, our performance is better. On OTB2015, a consistent advantage of our ASDCF among the state-of-the-art methods is obtained, achieving 92.7% in terms of DP and 69.7% in terms of AUC. In addition, OP, CLE and AUC metrics on these two datasets are also reported in Table 2. Our ASDCF achieves the best OP score and AUC on both OTB2013 and OTB2015. On OTB2015, ASDCF obtains accurate and robust tracking results, with the best OP/CLE, 87.9%/9.5 pixels. We credit the performance improvement to the effective affine subspace construction, with more discriminative information retained in the filter learning stage.

We also report the precision plots and success plots on UAV123 in Fig. 4. As shown in the figure, the proposed ASDCF produces the best results in terms of both DP and AUC. ASDCF outperforms the advanced DCF trackers, i.e., ECO (by 2.0% and 0.6%), C-COT (by 5.0% and 3.1%), and LADCF (by 5.1% and 1.6%), respectively, in terms of DP and AUC. Therefore, by explicitly modeling the appearance variation during spatio-temporal changes, ASDCF exhibits adaptive context awareness with an outstanding generalization.

In addition, in Table 3, we report the tracking performance obtained on VOT2018. VOT sequences consist of diverse challenging factors, with more severe appearance variations. Our ASDCF approach performs best in the EAO metric, achieving a relative gain of 1.2% compared to the DCF approach LADCF. Compared to the deep learning based method FCOS_MAML trained offline with large-scale data, the proposed ASDCF reports a gain of 0.9% in

Table 2 Performance comparison of our ASDCF method with the state-of-the-art trackers, evaluated on OTB2013 and OTB2015 in terms of OP and CLE. The best three results are highlighted in red, blue and brown

		KCF	CSRDCF	Staple	CACF	SRDCF	BACF	SiamFC	MetaTracker
OP (%)	OTB2013	60.8	74.4	73.8	77.6	76.0	84.0	77.9	85.6
	OTB2015	54.4	70.5	70.2	73.0	71.1	77.6	73.0	79.8
CLE (pixels)	OTB2013	36.3	31.9	31.4	29.8	36.8	26.2	29.7	11.5
	OTB2015	45.1	31.1	31.8	33.1	39.7	28.2	33.2	14.2
AUC (%)	OTB2013	50.5	58.4	59.1	62.1	61.0	65.7	60.7	66.7
	OTB2015	47.3	57.4	57.7	60.0	58.7	62.1	58.2	63.7
FPS		**97.7**	7.5	**32.8**	**25.9**	6.6	26.1	25.6	16.0
		CREST	MCPF	ECO	C-COT	STRCF	VITAL	LADCF	ASDCF
OP (%)	OTB2013	86.0	85.8	88.7	83.7	86.6	**91.4**	**90.7**	**93.2**
	OTB2015	77.6	78.0	84.9	82.3	84.6	**86.5**	**86.7**	**87.9**
CLE (pixels)	OTB2013	**10.2**	11.2	16.2	15.6	21.3	**7.4**	15.7	**8.6**
	OTB2015	21.2	20.9	14.8	14.0	17.8	**9.9**	**13.8**	**9.5**
AUC (%)	OTB2013	67.3	67.7	70.9	67.7	68.0	**71.0**	**71.5**	**72.0**
	OTB2015	62.3	62.8	**69.1**	67.3	67.5	68.2	**69.6**	**69.7**
FPS		18.1	4.3	9.7	2.8	8.2	1.5	9.8	8.8

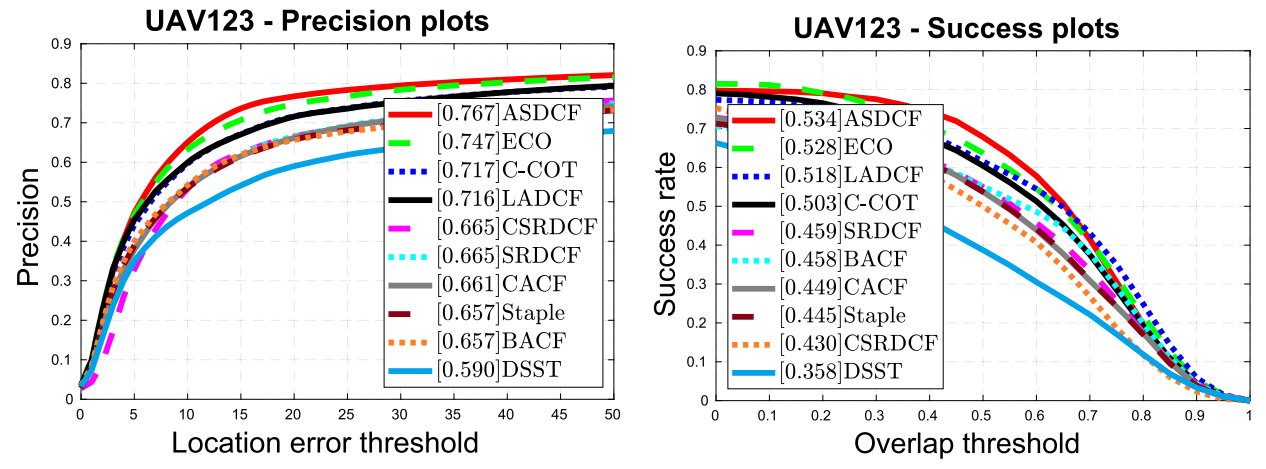

Figure 4 The experimental performance on UAV123. Precision plots (with the DP score reported in the figure legend) and the success plots (with the AUC score reported in the figure legend) are presented

Table 3 The tracking results on VOT2018. The best three results are highlighted by red, blue and brown

	ECO	CFCF	CFWCR	LSART	MFT	UPDT	SiamRPN	LADCF	ULAST	FCOS_MAML	ASDCF
EAO	0.280	0.286	0.303	0.323	0.385	0.378	0.383	**0.389**	0.355	**0.392**	**0.401**
Accuracy	0.483	0.509	0.484	0.493	0.505	0.536	**0.586**	0.503	**0.571**	**0.635**	0.523
Robustness	0.276	0.281	0.267	0.218	**0.140**	0.184	0.276	**0.159**	0.286	0.220	**0.151**

terms of EAO. For robustness, ASDCF also produces comparable results within the top 3 trackers. In principle, the proposed ASDCF realizes favorable tracking performance compared with other DCF approaches, i.e., ECO, CFWCR, UPDT, and LADCF, demonstrating the advantage of performing filter learning based on the appearance representation provided by the affine subspace.

Compared to these state-of-the-art DCF-based trackers that extract representations from independent templates, the proposed affine subspace strengthens the representation capacity for latent appearance variations. With more powerful representation, undoubtedly, the ASDCF can learn more discriminative and robust filters, leading to precise and stable tracking, even in the presence of

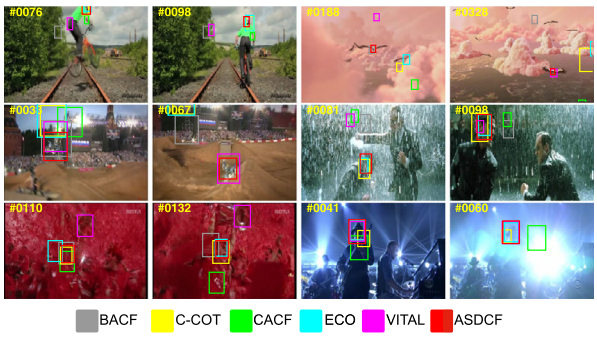

Figure 5 A qualitative comparison of our ASDCF method with the state-of-the-art trackers, including BACF [59], CACF [57], C-COT [13], ECO [20] and VITAL [72], on some challenging video sequences of the OTB2015 [3] (Left column top to bottom: *Biker*, *MotorRolling*, and *Soccer*. Right column top to bottom: *Bird1*, *Matrix*, and *Shaking*)

severe appearance variations caused by various factors. Therefore, on these challenging benchmark datasets, the proposed ASDCF outperforms the state-of-the-art DCF-based methods and some deep learning-based trackers.

5.4.2 Qualitative performance

Qualitative comparisons with tracking challenges are presented in Fig. 5, which shows the intuitive tracking results of the state-of-the-art approaches, i.e., BACF, C-COT, CACF, ECO, VITAL and the proposed ASDCF, on some challenging video sequences. The difficulties are generated by rapid changes in the appearance of both targets and the corresponding surroundings. Our ASDCF exhibits competitive performance on these challenges as it successfully identifies the pertinent spatio-temporal target patterns. Sequences with deformations (*MotorRolling*, *Matrix*) and out-of-view (*Biker*, *Bird1*) can be successfully tracked by our method without any failures. Videos with rapid motions (*Biker*, *Matrix*) also benefit from our strategy of exploring relevant deep channels to enhance discrimination. Specifically, ASDCF is an expert in solving in-plane and out-of-plane rotations (*Biker*, *MotorRolling*), because the proposed affine subspace enables adaptive appearance updating with improved model capacity compared with other DCF approaches.

6 Conclusion

In this paper, we proposed an effective appearance model with an outstanding performance by learning discriminative correlation filters in the adaptively updated affine subspace. The affine subspace enables effective spatio-temporal appearance representation, providing more discriminative clues than single template learning. A spatio-temporal regularized DCF formulation accompanied by efficient optimization also contributes to achieving accurate and robust performance in the affine subspace. The quantitative and qualitative experimental results on tracking benchmarking datasets demonstrate the consistent effectiveness of our method, compared with state-of-the-art trackers. The merits of introducing affine subspace to the DCF learning framework support the potential of exploring more effective representation spaces with spatio-temporal capacity in online visual object tracking.

Funding

This work was supported in part by the National Natural Science Foundation of China (Grant Nos. U1836218, 62106089).

Abbreviations

ASDCF, affine subspace DCF; AUC, area under the curve; CLE, center location error; CNN, convolutional neural network; DCF, discriminative correlation filter; DP, distance precision; EAO, expected average overlap; HOG, histogram of oriented gradient; MOSSE, minimum output sum of squared error; OP, overlap precision; SVD, singular value decomposition; VOT, visual object tracking.

Availability of data and materials

The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

Declarations

Competing interests

The authors declare no competing interests.

Author contributions

All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by TX, X-FZ and X-JW. The first draft of the manuscript was written by TX and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Received: 1 September 2022 Revised: 17 December 2022
Accepted: 20 February 2023 Published online: 08 May 2023

References

1. Henriques, J. F., Rui, C., Martins, P., & Batista, J. (2015). High-speed tracking with kernelized correlation filters. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *37*(3), 583–596.
2. Wu, Y., Lim, J., & Yang, M. H. (2013). Online object tracking: a benchmark. In *IEEE conference on computer vision and pattern recognition* (pp. 2411–2418). Los Alamitos: IEEE.
3. Wu, Y., Lim, J., & Yang, M.-H. (2015). Object tracking benchmark. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *37*(9), 1834–1848.
4. Mueller, M., Smith, N., & Ghanem, B. (2016). A benchmark and simulator for uav tracking. In *European conference on computer vision* (pp. 445–461). Berlin: Springer.
5. Kristan, M., Leonardis, A., Matas, J., Felsberg, M., Pflugfelder, R., Zajc, L. C., Vojir, T., Hager, G., Lukezic, A., Eldesokey, A., & Fernandez, G. (2017). The visual object tracking VOT2017 challenge results. In *2017 IEEE international conference on computer vision workshops* (pp. 1949–1972). Los Alamitos: IEEE. https://doi.org/10.1109/ICCVW.2017.230.
6. Kristan, M., Leonardis, A., Matas, J., Felsberg, M., Pfugfelder, R., Zajc, L. C., Vojir, T., Bhat, G., Lukezic, A., Eldesokey, A., Fernandez, G., et al. (2018). The sixth visual object tracking VOT2018 challenge results. In *ECCV workshops 2018* (pp. 3–53). Berlin: Springer.
7. Dawei, D., Zhu, P., Wen, L., Bian, X., Ling, H., Hu, Q., et al. (2019). VisDrone-SOT2019: the vision meets drone single object tracking challenge results. In *Proceedings of the IEEE international conference on computer vision workshops* (pp. 199–212). Los Alamitos: IEEE.
8. Fan, H., Wen, L., Du, D., Zhu, P., Hu, Q., Ling, H., et al. (2020). VisDrone-SOT2020: the vision meets drone single object tracking challenge results. In *European conference on computer vision* (pp. 728–749). Berlin: Springer.

9. Gray, R. M. (2006). Toeplitz and circulant matrices: a review. *Foundations and Trends in Communications and Information Theory*, *2*(3), 155–239.
10. Henriques, J. F., Caseiro, R., Martins, P., & Batista, J. (2012). Exploiting the circulant structure of tracking-by-detection with kernels. In *European conference on computer vision* (pp. 702–715). Berlin: Springer.
11. Danelljan, M., Khan, F. S., Felsberg, M., & Van De Weijer, J. (2014). Adaptive color attributes for real-time visual tracking. In *IEEE conference on computer vision and pattern recognition* (pp. 1090–1097). Los Alamitos: IEEE.
12. Xu, T., Feng, Z.-H., Wu, X.-J., & Kittler, J. (2019). Joint group feature selection and discriminative filter learning for robust visual object tracking. In *Proceedings of the IEEE international conference on computer vision* (pp. 7950–7960). Los Alamitos: IEEE.
13. Martin, D., Andreas, R., Fahad, K., & Michael, F. (2016). Beyond correlation filters: learning continuous convolution operators for visual tracking. In *European conference on computer vision* (pp. 472–488). Berlin: Springer.
14. Bertinetto, L., Valmadre, J., Golodetz, S., Miksik, O., & Torr, P. H. S. (2016). Staple: complementary learners for real-time tracking. In *IEEE conference on computer vision and pattern recognition* (Vol. *38*, pp. 1401–1409). Los Alamitos: IEEE.
15. Danelljan, M., Hager, G., Khan, F. S., & Felsberg, M. (2015). Learning spatially regularized correlation filters for visual tracking. In *IEEE international conference on computer vision* (pp. 4310–4318). Los Alamitos: IEEE.
16. Lukezic, A., Vojir, T., Zajc, L. C., Matas, J., & Kristan, M. (2017). Discriminative correlation filter with channel and spatial reliability. In *IEEE conference on computer vision and pattern recognition* (pp. 4847–4856). Los Alamitos: IEEE.
17. Zhang, M., Wang, Q., Xing, J., Gao, J., Peng, P., Hu, W., & Maybank, S. (2018). Visual tracking via spatially aligned correlation filters network. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 469–485). Berlin: Springer.
18. Danelljan, M., Hager, G., Khan, F. S., & Felsberg, M. (2016). Adaptive decontamination of the training set: a unified formulation for discriminative visual tracking. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1430–1438). Los Alamitos: IEEE.
19. Li, F., Tian, C., Zuo, W., Zhang, L., & Yang, M.-H. (2018). Learning spatial-temporal regularized correlation filters for visual tracking. arXiv preprint. arXiv:1803.08679.
20. Danelljan, M., Bhat, G., Khan, F. S., & Eco, M. F. (2017). Efficient convolution operators for tracking. In *IEEE conference on computer vision and pattern recognition* (pp. 6931–6939). Los Alamitos: IEEE.
21. Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. Burges, L. Bottou, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems 25 (NIPS 2012)* (pp. 1097–1105). Red Hook: Curran Associates.
22. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., & Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1–9). Los Alamitos: IEEE.
23. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778). Los Alamitos: IEEE.
24. Liu, D., Cui, W., Jin, K., Guo, Y., & Qu, H. (2018). Deeptracker: visualizing the training process of convolutional neural networks. *ACM Transactions on Intelligent Systems and Technology*, *10*(1), 1–25.
25. Lucas, B. D., & Kanade, T. (1981). An iterative image registration technique with an application to stereo vision. In *Proceedings of the 7th international joint conference on artificial intelligence (IJCAI'81)* (pp. 674–679). Los Altos: William Kaufmann.
26. Avidan, S. (2004). Support vector tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *26*(8), 1064–1072.
27. Arulampalam, M. S., Maskell, S., Gordon, N., & Clapp, T. (2002). A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking. *IEEE Transactions on Signal Processing*, *50*(2), 174–188.
28. Ross, D. A., Lim, J., Lin, R.-S., & Yang, M.-H. (2008). Incremental learning for robust visual tracking. *International Journal of Computer Vision*, *77*(1–3), 125–141.
29. Li, B., Yan, J., Wu, W., Zhu, Z., & Hu, X. (2018). High performance visual tracking with Siamese region proposal network. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 8971–8980). Los Alamitos: IEEE.
30. Li, X., Hu, W., Shen, C., Zhang, Z., Dick, A., & Van Den Hengel, A. (2013). A survey of appearance models in visual object tracking. *ACM Transactions on Intelligent Systems and Technology*, *4*(4), 1–48.
31. Li, A., Lin, M., Wu, Y., Yang, M. H., & Yan, S. (2016). Nus-pro: a new visual tracking challenge. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *38*(2), 335–349.
32. Yao, R., Lin, G., Xia, S., Zhao, J., & Zhou, Y. (2020). Video object segmentation and tracking: a survey. *ACM Transactions on Intelligent Systems and Technology*, *11*(4), 1–47.
33. Comaniciu, D., Ramesh, V., & Meer, P. (2000). Real-time tracking of non-rigid objects using mean shift. In *IEEE conference on computer vision and pattern recognition* (pp. 142–149). Los Alamitos: IEEE.
34. Hardegger, M., Roggen, D., Calatroni, A., & Tröster, G. (2016). S-smart: a unified Bayesian framework for simultaneous semantic mapping, activity recognition, and tracking. *ACM Transactions on Intelligent Systems and Technology*, *7*(3), 1–28.
35. Zhang, S., Yao, H., Sun, X., & Liu, S. (2012). Robust visual tracking using an effective appearance model based on sparse coding. *ACM Transactions on Intelligent Systems and Technology*, *3*(3), 1–18.
36. Zhang, T., Bibi, A., & Ghanem, B. (2016). In defense of sparse tracking: circulant sparse tracker. In *2016 IEEE conference on computer vision and pattern recognition* (pp. 3880–3888). Los Alamitos: IEEE.
37. Zhang, T., Liu, S., Ahuja, N., Yang, M.-H., & Ghanem, B. (2015). Robust visual tracking via consistent low-rank sparse learning. *International Journal of Computer Vision*, *111*(2), 171–190.
38. Babenko, B., Yang, M. H., & Belongie, S. (2011). Robust object tracking with online multiple instance learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *33*(8), 1619–1632.
39. Tao, R., Gavves, E., & Smeulders, A. W. M. (2016). Siamese instance search for tracking. In *IEEE conference on computer vision and pattern recognition* (pp. 1420–1429). Los Alamitos: IEEE.
40. Valmadre, J., Bertinetto, L., Henriques, J., Vedaldi, A., & Torr, P. H. S. (2017). End-to-end representation learning for correlation filter based tracking. In *IEEE conference on computer vision and pattern recognition* (pp. 5000–5008). Los Alamitos: IEEE.
41. Xu, T., Feng, Z.-H., Wu, X.-J., & Kittler, J. (2020). Afat: adaptive failure-aware tracker for robust visual object tracking. arXiv preprint. arXiv:2005.13708.
42. Bolme, D. S., Beveridge, J. R., Draper, B. A., & Lui, Y. M. (2010). Visual object tracking using adaptive correlation filters. In *2010 IEEE conference on computer vision and pattern recognition* (pp. 2544–2550). Los Alamitos: IEEE.
43. Bolme, D. S., Draper, B. A., & Beveridge, J. R. (2009). Average of synthetic exact filters. In *2009 IEEE conference on computer vision and pattern recognition* (pp. 2105–2112). Los Alamitos: IEEE.
44. Briechle, K., & Hanebeck, U. D. (2001). Template matching using fast normalized cross correlation. In *Optical pattern recognition XII* (Vol. *4387*, pp. 95–103). Bellingham: International Society for Optics and Photonics.
45. Zhang, K., Zhang, L., Liu, Q., Zhang, D., & Yang, M. H. (2014). Fast visual tracking via dense spatio-temporal context learning. In *European conference on computer vision* (pp. 127–141). Berlin: Springer.
46. Li, Y., & Zhu, J. (2014). A scale adaptive kernel correlation filter tracker with feature integration. In *European conference on computer vision workshops* (pp. 254–265). Berlin: Springer.
47. Danelljan, M., Häger, G., Khan, F. S., & Felsberg, M. (2017). Discriminative scale space tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *39*(8), 1561–1575.
48. Li, Y., Zhu, J., & Hoi, S. C. H. (2015). Reliable patch trackers: robust visual tracking by exploiting reliable patches. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 353–361). Los Alamitos: IEEE.
49. Liu, S., Zhang, T., Cao, X., & Xu, C. (2016). Structural correlation filter for robust visual tracking. In *IEEE conference on computer vision and pattern recognition* (pp. 4312–4320). Los Alamitos: IEEE.
50. Tang, M., & Feng, J. (2015). Multi-kernel correlation filter for visual tracking. In *IEEE international conference on computer vision* (pp. 3038–3046). Los Alamitos: IEEE.
51. Xu, T., Feng, Z.-H., Wu, X.-J., & Kittler, J. (2020). Learning low-rank and sparse discriminative correlation filters for coarse-to-fine visual object tracking. *IEEE Transactions on Circuits and Systems for Video Technology*, *30*(10), 3727–3739.
52. Zhang, T., Xu, C., & Yang, M.-H. (2018). Robust structural sparse tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *41*(2), 473–486.
53. Xu, T., Wu, X.-J., & Kittler, J. (2018). Non-negative subspace representation learning scheme for correlation filter based tracking. In *2018 24th international conference on pattern recognition (ICPR)* (pp. 1888–1893). Los Alamitos: IEEE.

54. Wang, M., Liu, Y., & Huang, Z. (2017). Large margin object tracking with circulant feature maps. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 21–26). Los Alamitos: IEEE.
55. Zuo, W., Wu, X., Lin, L., Zhang, L., & Yang, M.-H. (2018). Learning support correlation filters for visual tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *41*(5), 1158–1172.
56. Zhang, T., Xu, C., & Yang, M.-H. (2017). Multi-task correlation particle filter for robust object tracking. In *IEEE conference on computer vision and pattern recognition* (Vol. *1*, p. 3). Los Alamitos: IEEE.
57. Mueller, M., Smith, N., & Ghanem, B. (2017). Context-aware correlation filter tracking. In *IEEE conference on computer vision and pattern recognition* (pp. 1396–1404). Los Alamitos: IEEE.
58. Zhu, Z., Wang, Q., Li, B., Wu, W., Yan, J., & Hu, W. (2018). Distractor-aware Siamese networks for visual object tracking. In *European conference on computer vision* (pp. 103–119). Berlin: Springer.
59. Galoogahi, H. K., Fagg, A., & Lucey, S. (2017). Learning background-aware correlation filters for visual tracking. In *IEEE international conference on computer vision* (pp. 1144–1152). Los Alamitos: IEEE.
60. Xu, T., Feng, Z.-H., Wu, X.-J., & Kittler, J. (2020). An accelerated correlation filter tracker. *Pattern Recognition*, *102*, 107172.
61. Xu, L., Kim, P., Wang, M., Pan, J., Yang, X., & Gao, M. (2022). Spatio-temporal joint aberrance suppressed correlation filter for visual tracking. *Complex & Intelligent Systems*, *8*(5), 3765–3777.
62. Xu, T., Feng, Z., Wu, X.-J., & Kittler, J. (2021). Adaptive channel selection for robust visual object tracking with discriminative correlation filters. *International Journal of Computer Vision*, *129*(5), 1359–1375.
63. Xu, T., Feng, Z.-H., Wu, X.-J., & Kittler, J. (2019). Learning adaptive discriminative correlation filters via temporal consistency preserving spatial feature selection for robust visual object tracking. *IEEE Transactions on Image Processing*, *28*(11), 5596–5609.
64. Zhu, X.-F., Wu, X.-J., Xu, T., Feng, Z.-H., & Kittler, J. (2021). Robust visual object tracking via adaptive attribute-aware discriminative correlation filters. *IEEE Transactions on Multimedia*, *24*, 301–312.
65. Bowen, L., Fu, C., Ding, F., Ye, J., & Lin, F. (2021). Adtrack: target-aware dual filter learning for real-time anti-dark uav tracking. In *2021 IEEE international conference on robotics and automation (ICRA)* (pp. 496–502). Los Alamitos: IEEE.
66. Boyd, S., Parikh, N., Chu, E., Peleato, B., & Eckstein, J. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, *3*(1), 1–122.
67. Petersen, K. B., Pedersen, M. S., et al. (2008). The matrix cookbook. Technical University of Denmark, 7(15), 510.
68. Vedaldi, A., & Lenc, K. (2015). Matconvnet: convolutional neural networks for Matlab. In *Proceedings of the 23rd ACM international conference on multimedia* (pp. 689–692). New York: ACM.
69. Kristan, M., Leonardis, A., Matas, J., Felsberg, M., Pflugfelder, R., Zajc, L. Č., et al. (2016). The visual object tracking VOT2016 challenge results. In *ECCV 2016 workshops* (pp. 777–823). Berlin: Springer.
70. Bhat, G., Danelljan, M., Van Gool, L., & Timofte, R. (2020). Know your surroundings: exploiting scene information for object tracking. In *European conference on computer vision* (pp. 205–221). Berlin: Springer.
71. Dai, K., Wang, D., Lu, H., Sun, C., & Li, J. (2019). Visual tracking via adaptive spatially-regularized correlation filters. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4670–4679). Los Alamitos: IEEE.
72. Song, Y., Ma, C., Wu, X., Gong, L., Bao, L., Zuo, W., Shen, C., Lau, R., & Yang, M.-H. (2018). Vital: visual tracking via adversarial learning. arXiv preprint. arXiv:1804.04273.
73. Park, E., & Berg, A. C. (2018). Meta-tracker: fast and robust online adaptation for visual object trackers. arXiv preprint. arXiv:1801.03049.
74. Song, Y., Ma, C., Gong, L., Zhang, J., Lau, R., & Yang, M.-H. (2017). Crest: convolutional residual learning for visual tracking. In *IEEE international conference on computer vision* (pp. 2555–2564). Los Alamitos: IEEE.
75. Choi, J., Chang, H. J., Yun, S., Fischer, T., Demiris, Y., & Choi, J. Y. (2017). Attentional correlation filter network for adaptive visual tracking. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4807–4816). Los Alamitos: IEEE.
76. Bertinetto, L., Valmadre, J., Henriques, J. F., Vedaldi, A., & Torr, P. H. S. (2016). Fully-convolutional Siamese networks for object tracking. In *European conference on computer vision* (pp. 850–865). Berlin: Springer.
77. Gundogdu, E., & Alatan, A. A. (2018). Good features to correlate for visual tracking. *IEEE Transactions on Image Processing*, *27*(5), 2526–2540.
78. Bhat, G., Johnander, J., Danelljan, M., Khan, F. S., & Felsberg, M. (2018). Unveiling the power of deep tracking. arXiv preprint. arXiv:1804.06833.
79. Shen, Q., Qiao, L., Guo, J., Li, P., Li, X., Li, B., Feng, W., Gan, W., Wu, W., & Ouyang, W. (2022). Unsupervised learning of accurate Siamese tracking. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 8101–8110). Los Alamitos: IEEE.
80. Wang, G., Luo, C., Sun, X., Xiong, Z., & Zeng, W. (2020). Tracking by instance detection: a meta-learning approach. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 6288–6297). Los Alamitos: IEEE.

Publisher's Note