Bilateral Weighted Regression Ranking Model with Spatial-Temporal Correlation Filter for Visual Tracking

Hu Zhu, *Member, IEEE,* Hao Peng, Guoxia Xu, *Member, IEEE,* Lizhen Deng, *Member, IEEE,* Yueying Cheng, and Aiguo Song *Senior Member, IEEE*

***Abstract*—Many discriminative correlation filter (DCF)-based methods have successfully leveraged the guidance for solving two problems (i.e., the boundary effect and temporal filtering degradation) as a model prior to visual tracking. Regardless of the specific content of the tracking algorithms, the intuitive motivation of these methods is to control the degeneration of the updating loss of the objective function with a structural framework. While these methods rely mostly on various explicit prior regularization items, they always ignore the loss from the data fidelity term. Therefore, we propose a bilateral weighted regression ranking model with a spatial-temporal correlation filter, namely, BWRR. Here, we resort to two procedures for solving the above problems. First, BWRR introduces a bilateral constraint into the data fidelity term to control the loss of rows and columns of the filter learning data term. The weighted matrices could impose an adaptive penalty for large data loss during the learning process to avoid the tracking offset problem and model degradation problem. Second, the data of the updated weighted matrices is not directly applied to the calculation of the filter during each iteration. Instead, a new weighted product matrix is obtained by ranking and numerical transformation for updating the filter. We show that the proposed model converts the original correlation filter regression problem into a regression-with-ranking problem, thus avoiding the problem of positive and negative sample imbalance. Overall, the BWRR model is approximated as a linear equality constraint problem, which is iteratively solved by the alternating direction method of multipliers(ADMM). Qualitative and quantitative evaluations demonstrate the effectiveness and superiority of our proposed method by extensive and quantitative experiments on the OTB, VOT, and UAV datasets.**

***Index Terms*—Bilateral Weighted Regression, Spatial-Temporal, Ranking, Visual Tracking**

I. Introduction

VISUAL tracking plays an important role in computer vision, image recognition and classification. With the rapid development of research, various tracking methods have been proposed and have yielded very effective results [1], [2]. Visual tracking methods are generally divided into two categories: generative model methods and discriminative model methods. Thanks to their powerful feature learning and computing capabilities, DCF-based trackers have become the mainstream of research [3]–[6]. However, it is still difficult for DCF trackers to maintain accurate and robust tracking in unconstrained scenarios.

In DCF methods, there are two main problems that affect the visual tracking: the boundary effect and temporal filtering degradation. To solve the first issue, i.e., the boundary effect, the spatially regularized DCF (SRDCF) [7] was proposed to introduce a penalty for the background in training correlation filters. On this basis, the spatial-temporal regularized DCF (STRCF) in [8] introduces a spatial-temporal regularization to obtain a joint solution between the two major problems and achieve superior performance over the SRDCF [7]. However, in a tracker based on spatial-temporal correlation, due to the influence of noise or the background environment, the data in the image is prone to sudden changes, namely,"outliers", such that the discriminant filter cannot accurately learn from the object and instead learns from the background environment. In addition, the data fidelity term of the loss function will produce a large loss due to these "outliers", which will accelerate the degradation of the model in the model optimization and updating and ultimately affect the target tracking performance.

In [9], the checking and discarding of "outliers" are achieved by hypothesis testing, while the model refuses to carry out measurement if the "outliers" exceed the given confidence interval, which cannot fully alleviate the influence of "outlier" points. Therefore, we propose a bilateral weighted regression ranking (BWRR) model with a spatial-temporal regularization term. Inspired by the trilateral weighted sparse coding (TWSC) scheme in [10], the BWRR model embeds two weighting matrices on both sides of the data fidelity term and controls the loss of rows and columns of the data fidelity term by adjusting the parameters of the weighting matrices to improve the stability of the model. To be specific, the loss function in the classification task is susceptible to the "outlier" [11]. If the predicted value of a certain point has a large error with respect to the true value, the model tends to produce a large error. Therefore, this model uses adaptive adjustment of the weighted matrices from the data fidelity term to avoid the large loss caused by the "outliers".

If bilateral weighting is used to alleviate the adverse effects of a small amount of mutation point data, then the sparse term

This work is supported by the National Natural Science Foundation of China under Grant 62072256. *(Corresponding author: Lizhen Deng) (E-mail: alicedenglzh@gmail.com)*

Hu Zhu, Hao Peng and Yueying Cheng are with Jiangsu Province Key Lab on Image Processing and Image Communication, Nanjing University of Posts and Telecommunications, Nanjing 210003, China. Guoxia Xu is with Department of Computer Science, Norwegian University of Science and Technology, 2815 Gjovik, Norway. Lizhen Deng is with National Engineering Research Center of Communication and Network Technology, Nanjing University of Posts and Telecommunications, Nanjing, 210003, China. Aiguo Song is with the School of Instrument Science and Engineering, Southeast University, Nanjing, 210096, China.

added in BWRR improves the performance of the model by controlling the integrity of the image data. The image data are often represented by high-dimensional feature vectors in image processing and the class label of input data can be predicted by a linear classification filter [12], [13]. Our BWRR algorithm selects the most distinguishing feature subset from the entire feature set by introducing a sparse term based on the ℓ_1 norm to achieve dimensionality reduction of the feature data. In this way, the BWRR model can be regarded as a linear equality constraint problem that can simplify the data processing [10]. The selection of the channel data is realized by assigning a weight matrix. In addition to the selection control of the image data, the type of features extracted during the feature learning has a certain degree of influence on the tracking effect. To show the tracking performance of our model, we mainly use HOG features to implement the data feature of images, which have strong robustness to image geometric deformation, lighting and shadow transformation. In addition, we adopt the deep feature [14], [15] to verify the performance of BWRR. It has been proved that the tracking effect based on deep features is better than that based on HOG features.

Moreover, the sample imbalance problem over positive and negative samples is an open problem that has always existed in the one-stage target detection algorithm, which is unable to converge to a good solution for data training and updating [16]. Considering the constraint of weight matrices on channel information, the bilateral weighted matrices always appear as a constrained product over rows and columns of data. On this basis, we take the ranking of the elements and convert the values to update the product matrix inspired by [16]. In this way, the BWRR can not only mitigate the impact of "outliers" but also avoid the problem of positive and negative sample imbalance during target detection. The advantage of this is to convert the original tracking correlation filter regression problem into a regression-with-ranking problem. Although the learning adaptive discriminative correlation filters (LADCF) [17] also uses the ℓ_1-norm and ranking, the actual spatial domain of the LADCF is indeed fixed, which obviously cannot meet the ever-changing requirements of spatial characteristics. Furthermore, in the update of the LADCF, the ranking method mainly performs numerical processing on the ℓ_1-norm. Unlike the LADCF, our ranking is used to process the values of the weight matrix during the update process. Thus, the values in the weight matrix become a set of arithmetic progressions to avoid abnormally large values (these abnormally large values make the model appear to incur large loss during the updating procedure).

Through the above analysis, the least squares regression equation is used to solve the filter updating problem and the whole iterative procession is achieved by the alternating direction method of multipliers (ADMM). To fully demonstrate and analyze the superior performance of our tracking algorithm, we compare BWRR and other state-of-the-art methods based on the HOGCN feature and deep learning feature, respectively. The experimental results prove that our BWRR has excellent performance in terms of the robustness and accuracy of target tracking.

The contributions of this work are as follows:

- A bilateral weighted regression ranking (BWRR) algorithm with two weighted matrices in the data fidelity term to control the loss of rows and columns and achieve weighted constraints on multiple channels is proposed in this paper.
- A sparse term based on the ℓ_1-norm is introduced into our BWRR to select the channel data and utilize the multiple channel prior statistical knowledge. The accuracy of sparse selection is guaranteed by weight control.
- We update the bilateral weighted matrices during the optimization process and introduce the ranking method to realize the update process to better alleviate the problem of sample imbalance.
- Since the BWRR can be treated as a linear equality constraint problem, the iterative process is solved by the ADMM algorithm, and a comprehensive experiment proves the superiority of the BWRR.

II. Related Work

A. DCF-based Trackers

DCF-based trackers have recently attracted wide attention. Compared with the traditional trackers with object detection and tracking algorithms [18], the DCFs simplify the mappings with high computational efficiency and strong robustness. In the frequency domain, DCFs utilize a circular structure to solve a ridge regression problem, such as MOSSE [19], KCF [20] and Staple [21], all of which improve the reliability of visual tracking. In addition, SAMF [22] and DSST [23] were proposed to handle scale variations, and the fDSST [24] performs scale detection in the tracking stage and improves the efficiency by a joint scale and location estimation. In addition, to acquire fewer boundary effects, the BACF based on HOG features was proposed in [25]. The SRDCF [7] and STRCF tracker [8] employ spatial and spatial-temporal information, respectively, to solve the boundary effect efficiently. Subgrid tracking by learning continuous convolution operators (CCOTs) was proposed in [26]. Efficient convolution operators (ECOs) [27] were proposed to achieve a lightweight version of the CCOT with a generative sample space and dimension-reduction mechanism. Furthermore, a 3rd-order tensor was used in [28] to represent the joint features of spatial and temporal information to achieve better tracking results with incremental N-mode SVD. Moreover, supervised tensor learning-based methods [29] have been proved to perform well when using a decomposition method to overcome the tracking representation overfitting problem in the field of target tracking. In addition, some trackers [30], [31] use neural network models to process image data, which greatly improves the tracking effect of a model in a responsible environment. DCF-based tracking methods have also been exploited to support structural constraints [32], long-term memory [15], [33], support vector machines(SVMs) [34], [35], the multikernel method [36], [37], and sparse representation [38], [39]. In addition to the handcrafted features used in [7], [20], [40], the deep feature is applied in SiamFC [41], CF-Net [42] HDT [43], and HCF [15] to achieve more precise and effective object tracking performance.

B. Deep-Learning-based Trackers

In the past, tracking algorithms mainly used histogram of oriented gradient (HOG) or HOGCN features, which have strong robustness to image geometric deformations, lighting and shadow transformations. However, some experimental results [44] showed that the tracking algorithms based on low-level handcrafted features are less likely to work well in some complex scenes; therefore, several trackers combine deep feature and correlation filters into visual tracking and have achieved robust performance [27], [42], [44]. In addition, deep learning (DL) [14], [15] forms more abstract high-level representation attribute categories or features by combining low-level features to discover distributed feature representations of data. Furthermore, the top-down supervised learning in deep learning trains labeled data and fine tunes the network to improve the feature learning effect, thereby obtaining better tracking results. For example, the Siamese network was introduced in SINT [45] and SiamFC [41] to achieve more simplicity and a competitive performance. By contrast, CF-Net [42] regards the correlation filter as a differentiable layer in the deep architecture to achieve good tracking results. The hybrid neural network with high tracking performance proposed in [46] can learn in a closed-loop system to achieve second-order practical tracking, and the neural weight of the network structure strengthens the model adaptability. Since deep learning is mainly implemented using convolutional networks, the experimental part later in this article discusses the impact of deep learning on the performance of tracking on convolutional networks with different layers.

C. Convolutional Sparse Coding Model

Visual tracking has been commonly formulated within the Bayesian filtering framework. The optimal state is obtained by the maximum a posteriori (MAP) estimation over a set of N samples [13]:

$$\hat{x}_t = \arg_{x_t^i} \max p\left(z_t|x_t^i\right) p\left(x_t^i|x_{t-1}\right)$$

where x_t^i is the i-th sample at frame t. In the next section, we present a tracking algorithm within the correlation filter framework. The samples at frame t can be drawn by a Gaussian function with mean x_{t-1} and variance δ^2:

$$p\left(x_t^i|x_{t-1}\right) = G\left(x_{t-1}, \sigma^2\right) \tag{1}$$

More samples in multiple channels are used to improve the tracking robustness at the expense of increasing the computational cost. At frame t, we denote the multichannel sample set as $X = \{x^1, x^2, \cdots, x^D\}$ which is obtained by the Gaussian function using Eq. (1). The corresponding filters are denoted as $f = \{f^1, f^2, \cdots, f^D\}$, where D is the number of channels. For the Dth channel, $x^d = \{x_1, x_2, \cdots x_{M\times N}\} \in \mathcal{R}^{M\times N\times 1}$ with a feature map size of $M \times N$. y is the predefined Gaussian-shaped label at time $t-1$. The convolutional sparse coding model can be formulated as

$$\min_f \|y - x * f\|_2^2 + \lambda \|f\|_q \tag{2}$$

where $*$ indicates the convolution of x and f, λ is the penalty factor of the sparse regularization term, and $q = 0$ or 1 to enforce sparse regularization on filter f.

In our BWRR, the joint sparsity is achieved by an ℓ_1-norm calculation, i.e., $q = 1$, and this group sparsity enables robust feature selection by reflecting the joint contribution of feature maps from all channels. The sparseness of the tracked target can be obtained by solving an ℓ_1-regularized least squares optimization problem. Moreover, this formulation is different from the ℓ_1 tracking method [47], which requires solving D ℓ_1-minimization problems. By contrast, the proposed method requires solving m ℓ_1-minimization problems ($m << D$), thereby reducing the computational complexity significantly.

D. Spatial-Temporal Correlation Filter Model

Before introducing the STRCF model, we briefly introduce the DCF trackers. The classical DCF tracking method trains a classifier from an image patch. First, given a circular matrix $X = [x_1, x_2, \cdots x_D]$ in $\mathbb{R}^{M\times N\times D}$ with a Gaussian function label y trained by DCF-based trackers with a filter f which also has D channels, the goal of each DCF-based tracker is to learn a function $f(x_i; f) = f^T \cdot x_i$ to distinguish the target from the background. These trackers can utilize the fast Fourier transform (FFT) and its inverse transform F^{-1} to improve the efficiency of computation in the Fourier domain.

$$f(X; f) = f^T X = f \otimes x = F^{-1}(\hat{f} \odot \hat{x}^*)$$

Here, $\hat{x}$ is the Fourier representation of x, $\hat{x}^*$ is the complex conjugate of $\hat{x}$ in the frequency domain, $\otimes$ denotes the circular convolution operator and $\odot$ denotes the operator of elementwise multiplication.

Second, DCF trackers find the best candidate to maximize the discriminant function in the current filter based on the model parameter $\bar{f}$ from a previous estimation or prior knowledge, which is formulated as the following tracking-learning-updating framework:

$$\tilde{x}_i = \arg\max_{x_i} f(x_i; \bar{f})$$

where the candidate x_i is a feature map extracted from the image, which has a good correlation with the original image, and the result calculated in the frequency domain is significant. After obtaining the tracking feature target, the new model is trained by minimizing the loss function.

$$\tilde{f} = \arg\min_f \theta(f, \psi) + \varphi(f)$$

where $\theta(\cdot)$ is the objective and $\varphi(\cdot)$ is the regularization function. $\psi = (X, f)$ indicates that the feature sample is processed by the filter.

According to the online passive-aggressive (PA) algorithm suggested in [48] , the STRCF [8] model combines a temporal regularization and derives the bound on the cumulative of the PA algorithm, which can be expressed as $\|f - f_{t-1}\|_2^2$. The objective function of the STRCF can then be expressed as:

$$\arg\min_f \frac{1}{2}\left\|\sum_{d=1}^{D} x_t^d * f^d - y\right\|_2^2 + \frac{1}{2}\sum_{d=1}^{D}\left\|\omega \cdot f^d\right\|_2^2 + \frac{\mu}{2}\left\|f - f_{t-1}\right\|_2^2 \tag{3}$$

where x denotes the images patches, f denotes the current filter, and f_{t-1} denotes the previous filter, which are all in $\mathbb{R}^{M\times N\times D}$. In addition, μ denotes the regularization parameter. The structured space regularization term calculates the ℓ_2-norm value of each filter channel.

III. The Proposed Method for Visual Tracking

A. The Proposed Bilateral Weighted Regression Ranking Model

Motivated by the excellent success of sparse representation in vision tasks [13], we introduce two weighting matrices on the data fidelity term and sparse discriminative term into the STRCF for object tracking, unifying the entire input during feature selection adaptively. Then the loss function of visual tracking can be formulated as follows.

$$\begin{aligned}\underset{f}{\operatorname{argmin}} & \underbrace{\frac{1}{2}\left\|\sum_{d=1}^{D} W_1\cdot\left(x_t^d * f^d - y\right)\cdot W_2\right\|_2^2}_{bilateral-data-fidelity-term} + \underbrace{\frac{1}{2}\left\|\sum_{d=1}^{D} W_3\cdot f^d\right\|_1}_{sparse-term} \\ & + \underbrace{\frac{1}{2}\sum_{d=1}^{D}\left\|\omega\cdot f^d\right\|_2^2}_{spatial-term} + \underbrace{\frac{\mu}{2}\left\|f-f_{t-1}\right\|_2^2}_{temporal-term}\end{aligned} \tag{4}$$

Here, the first term of the formulation is a data fidelity term with two weighting matrices on two sides to control the loss. The second term is the sparse term, and the weighting matrix W_3 is introduced to multiply by the filter f to ensure the accuracy of channel selection. The penalty factor λ in Eq. (2) is already included in the matrix. The third item is a spatial regularazition term, which adds a spatial regularization weight matrix ω. The fourth item is a temporal regular term, which is used to indicate the correlation between the current output frame of the filter and the previous frame. It is worth noting that $w, y, x_t^d, x_{t-1}^d, f^d, f_{t-1}^d$ in $\mathbb{R}^{M*N}$ are vectors with length $M*N$, with $d\in(1,D)$, and that W_1, W_2 are diagonal matrices. Moreover, W_1 is a block diagonal matrix with a total of D blocks corresponding to D channels and each block uses the same diagonal elements to describe the image features within each channel. W_2 weights the output features of the predicted label y. Through the joint weighting of W_1 and W_2, equalization constraints of multiple channels can be achieved to reduce model degradation and achieve more robust tracking effect.

In the BWRR model, the combination of sparse representation and spatial temporal regularity reduces the interference of noise on target tracking. The overall model enhances the correlation of target tracking with different frames and improves the performance of target tracking. A schematic diagram of the model is shown in Fig. 1. The whole process of visual tracking includes the following: 1) Preprocessing the input frame image to generate multiple candidate image blocks (including target blocks and background blocks); 2) Selecting appropriate candidate target blocks as prediction targets (prediction), where there are multiple prediction targets; 3) Updating prediction targets into our model; 4) Obtaining the tracking target. Since BWRR uses a discriminative model in the tracking process, a classifier is trained to distinguish the target from the background. With consideration of the temporal regularization term of the model, the final tracking result data of the previous frame are also added to the model update for iterative operation.

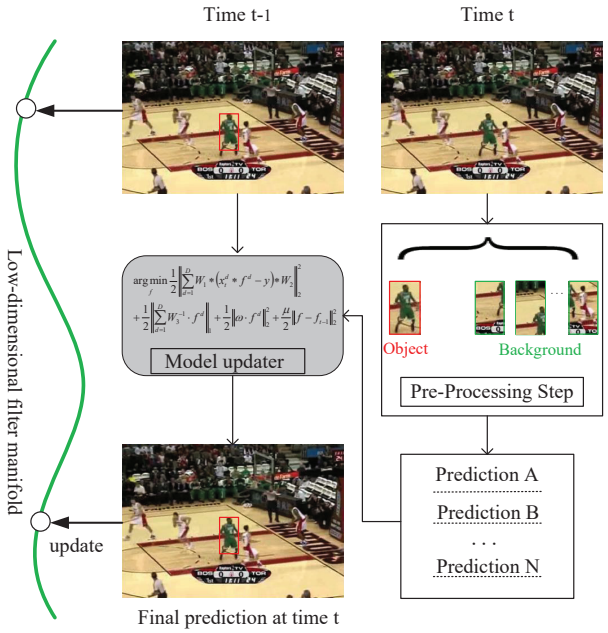


Fig. 1. Schematic diagram of visual tracking of the proposed model.

B. Model Optimization by the ADMM

The optimization of Eq. (4) is carried out by introducing the auxiliary variable g and Lagrange multiplier s. The auxiliary variable g is introduced by requiring $f^d = g^d$. Then, we obtain the Lagrangian augmentation function.

$$\begin{aligned}L(W,f,g,h) = & \frac{1}{2}\left\|\sum_{d=1}^{D} W_1\cdot\left(x_t^d * f^d - y\right)\cdot W_2\right\|_2^2 \\ & + \frac{1}{2}\left\|\sum_{d=1}^{D} W_3\cdot f^d\right\|_1 + \frac{1}{2}\sum_{d=1}^{D}\left\|\omega\cdot g^d\right\|_2^2 + \sum_{d=1}^{D}\left(f^d - g^d\right)\cdot s^d \\ & + \frac{\gamma}{2}\left\|f^d - g^d\right\|_2^2 + \frac{\mu}{2}\left\|f - f_{t-1}\right\|_2^2\end{aligned} \tag{5}$$

Let $h = \frac{1}{\gamma}s$, where γ is a step-size parameter. Then, the above formulation is converted to:

$$\begin{aligned}L(W,f,g,h) = & \left\|\sum_{d=1}^{D} W_1\cdot\left(x_t^d * f^d - y\right)\cdot W_2\right\|_2^2 + \left\|\sum_{d=1}^{D} W_3\cdot f^d\right\|_1 \\ & + \sum_{d=1}^{D}\left\|\omega\cdot g^d\right\|_2^2 + \gamma\left\|f^d - g^d + h^d\right\|_2^2 + \mu\left\|f - f_{t-1}\right\|_2^2\end{aligned} \tag{6}$$

where g^d, h^d in $\mathbb{R}^{M*N}$ have the same size as that of x^d. The Lagrangian augmentation function of the above formula

is divided into three subproblems by the ADMM algorithm.

$$\begin{cases} f^{(i+1)} = \arg\min\limits_{f} \left\| \sum\limits_{d=1}^{D} W_1 \cdot \left(x_t^d * f^d - y\right) \cdot W_2 \right\|_2^2 \\ + \left\| \sum\limits_{d=1}^{D} W_3 \cdot f^d \right\|_1 + \gamma \left\| f - g + h \right\|_2^2 + \mu \left\| f - f_{t-1} \right\|_2^2 \\ g^{(i+1)} = \arg\min\limits_{g} \sum\limits_{d=1}^{D} \left\| \omega \cdot g^d \right\|_2^2 + \gamma \left\| f - g + h \right\|_2^2 \\ h^{(i+1)} = h^{(i)} + f^{(i+1)} - g^{(i+1)} \end{cases} \tag{7}$$

We detail the solution to each subproblem for the update as follows.

Update of f:

Using Parseval's theorem, the first row of Eq. (7) can be rewritten in the Fourier domain as:

$$\begin{aligned} \underset{\hat{f}}{\operatorname{argmin}} & \left\| \sum_{d=1}^{D} W_1 \cdot \left(\hat{x}_t^d \cdot \hat{f}^d - \hat{y} \right) \cdot W_2 \right\|_2^2 + \left\| W_3 \cdot \hat{f}^d \right\|_1 \\ & + \gamma \| \hat{f} - \hat{g} + \hat{h} \|_2^2 + \mu \left\| \hat{f} - \hat{f}_{t-1} \right\|_2^2 \end{aligned} \tag{8}$$

where $\hat{f}$ denotes the discrete Fourier transform (DFT) of the filter f. Eq. (8) can be decomposed into $M * N$ subproblems, with the $j-th$ subproblem related to the $j-th$ element of f along all D channels. Let $v_j(\hat{f}) \in R^D$ denote the output of the j-th channel of the filter in D channels. Then, we obtain:

$$\begin{aligned} \underset{\hat{f}}{\operatorname{argmin}} & \tfrac{1}{2} \left\| W_1 \cdot \left(v_j\left(\hat{x}_t\right)^T \cdot v_j(\hat{f}) - \hat{y}_j \right) W_2 \right\|_2^2 + \left\| W_3 \cdot v_j(\hat{f}) \right\|_1 \\ & + \gamma \left\| v_j(\hat{f}) - v_j(\hat{g}) + v_j(\hat{h}) \right\|_2^2 + \mu \left\| v_j(\hat{f}) - v_j\left(\hat{f}_{t-1}\right) \right\|_2^2 \end{aligned} \tag{9}$$

To solve $v_j(\hat{f})$, we use the bilateral least squares regression equation, which is expressed as follows:

$$\mathrm{A}C_{k+1} + C_{k+1}B_k = E_k \tag{10}$$

where C_{k+1} is the solution required by the formulation, corresponding to the filter $f^{(i+1)}$ in our BWRR.

Since the fourth term in Eq. (9) does not conform to the formulation for the bilateral least squares regression solution, we do not include the fourth term in the calculation to make Eq. (9) satisfy the expression of Eq. (10). At the same time, we make $v_j(\hat{f}^*) = W_3 \cdot v_j(\hat{f})$. Then, Eq. (9) can be simplified to the following form:

$$\begin{aligned} \min_{\hat{f}} & \left\| W_1 \cdot \left(\hat{y}_j - v_j\left(\hat{x}_t\right)^T \cdot v_j(\hat{f}^*) \right) \cdot W_2 \right\|_2^2 \\ & + \gamma \left\| v_j(\hat{f}) - v_j(\hat{g}) + v_j(\hat{h}) \right\|_2^2 \end{aligned} \tag{11}$$

Corresponding to the expression of Eq. (10), we obtain:

$$\begin{cases} A = W_3^T \cdot v_j(\hat{x}_t)^T \cdot W_1^T \cdot W_1 \cdot v_j\left(\hat{x}_t\right) \cdot W_3 \\ B_k = \gamma \left(W_2 \cdot W_2^T \right)^{-1} \\ E_k = W_3^T \cdot v_j(\hat{x}_t)^T \cdot W_1^T \cdot W_1 \cdot \hat{y}_j \\ \quad + \left(\gamma v_j\left(\hat{g}\right) - \gamma v_j\left(\hat{h}\right) \right) \left(W_2 \cdot W_2^T \right)^{-1} \end{cases} \tag{12}$$

Substitute Eq. (12) into Eq. (10) to obtain the solution of C_{k+1} which is also the solution of filter $f^{(i+1)}$.

$$C_{k+1} = \frac{\hat{y}_j}{v_j(\hat{x}_t) \cdot W_3} + \frac{v_j(\hat{h}) - v_j(\hat{g})}{2W_2 \cdot W_2^T \cdot W_3^T \cdot v_j(\hat{x}_t)^T \cdot W_1^T \cdot W_1 \cdot v_j(\hat{x}_t) \cdot W_3} \tag{13}$$

In practice, the method of bilateral least squares regression is mainly applied to image denoising [10], and the method of target tracking is different from image denoising, which means that this method cannot be used directly. Therefore, we have improved the previous solution process. In the process of iterating $f^{(i+1)}$ with the ADMM, the weight matrix W_1, W_2 is regarded as a constant and then substituted into Eq. (9). The weight matrix W_1, W_2 is updated after an iteration is completed. According to a large number of experiments, the weight matrix W_3 is set as the identity matrix I to achieve the best effect. Then, we derive Eq. (9) to obtain the following formulation.

$$\begin{cases} v_j(\hat{f}) = V_1\left(\hat{f} | \hat{f}_{t-1}; \hat{x}_t\right) \cdot V_2\left(\hat{f} | \hat{x}_t\right) \\ V_1\left(\hat{f} | \hat{f}_{t-1}; \hat{x}_t\right) = v_j\left(\hat{x}_t\right) \cdot \hat{y}_j - \frac{1}{W_1 \cdot W_2 \cdot W_3} + \frac{2\gamma \cdot v_j(\hat{g})}{2W_1 \cdot W_2} \\ \qquad - \frac{2\gamma \cdot v_j(\hat{h})}{2W_1 \cdot W_2} + \frac{2\mu \hat{f}_{t-1}}{2W_1 W_2} \\ V_2\left(\hat{f} | \hat{x}_t\right) = \frac{W_1 \cdot W_2}{(2\gamma + 2\mu) I} \\ \qquad - \frac{W_1^T \cdot W_2^T \cdot v_j(\hat{x}_t)^T v_j(\hat{x}_t) \cdot W_2 \cdot W_1}{(2\gamma + 2\mu) I \left(W_1 \cdot v_j(\hat{x}_t)^T v_j(\hat{x}_t) \cdot W_2 + (2\gamma + 2\mu) I \right)} \end{cases} \tag{14}$$

We use the Sherman-Morrision formula to obtain:

$$v_j(\hat{f}) = \frac{W_1 \cdot W_2}{(2\gamma + 2\mu) I} - \frac{W_1^T \cdot W_2^T \cdot v_j(\hat{x}_t)^T \cdot v_j(\hat{x}_t) \cdot W_2 \cdot W_1}{(2\gamma + 2\mu) I (W_1 \cdot W_2 + (2\gamma + 2\mu) I)} \tag{15}$$

Update of W_1, W_2: When W_1 and W_2 are in the initial state, they are uniformly set to an identity matrix with the same size as that of the extracted feature matrix. When updating, since W_1 and W_2 always appear in the update formulation in the form of a product, let $W = W_1 \cdot W_2$. Update W by using the following formulation:

$$W = \exp\left(- \frac{\left\| \sum_{d=1}^{D} \left(x_t^d * f^d - y \right) \right\|_2^2}{2\tau^2} \right) \tag{16}$$

In the experiment, the parameter τ is calculated as 0.6.

In the actual experimental operation, instead of directly substituting the updated matrix W into the calculation in the next iteration, the data replacement operation is performed on W following [49].

- Step 1: Set all elements except the diagonal of the matrix to 0, leaving only the elements on the diagonal.
- Step 2: Rank the elements on the diagonal from small to large. Then each element gets an array number based on its size. We replace the corresponding element in the matrix with the permutation sequence number of each element to obtain a new diagonal matrix.
- Step 3: Convert the elements of the new diagonal matrix. Assume the value of the j-th element of the diagonal to be $W(j) = N$; then, use the following formulation to obtain a j-th new element:

$$W(j) = 1 + (N - 1) \cdot a \tag{17}$$

where a is the weight parameter. Finally, convert all the elements on the diagonal to obtain a new matrix W.

The data processing of W is actually carried out to select the features that can best maintain the data similarity of the entire feature set and give a sufficient constraint. Most feature selection algorithms evaluate the importance of each feature individually and then select them one by one. However, the existence of the imbalance problem between positive and negative samples will cause the model to learn more from negative samples, thus causing the target learning to shift. The unsupervised feature selection algorithm proposed in [49] uses a ranking method to update the discriminative W for learning feature selection. In addition, the author in [16] uses the ranking method to minimize the loss of each positive and negative sample pair and then achieves the goal of target detection. Motivated by these works, we also use a ranking method to update the matrix W to avoid the impact of sample imbalance.

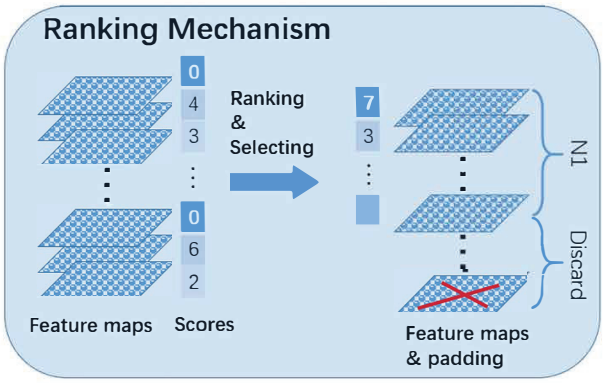


Fig. 2. The process of the ranking mechanism.

The ranking mechanism is shown in Fig. 2. As shown, we put forward the ranking mechanism to solve the problem of unbalanced samples. If a certain frame of the tracking video is extracted and the extracted feature map is $N1 \times N1 \times M$, then we have M feature maps. However, each of the M features has a different effect on the results. Therefore, we have to assign them a value first. The greater the impact on the result, the higher the importance is, and accordingly, the higher the score. Next, the M assigned feature maps are sorted and screened to select the first M with the highest score, with the redundant $(n-m)$ ones with the least impact on the result being discarded to obtain the final result.

Update of g: From the second subequation of Eq. (7), each element of g can be computed independently, and thus the closed-form solution of g can be computed by:

$$g = \left(\Sigma^T \Sigma + \gamma I\right)^{-1} (\gamma f + \gamma h) \tag{18}$$

where Σ represents the $DMN \times DMN$ diagonal matrix.

Update of the step-size parameter γ: The stepsize parameter γ is updated by Eq. (19):

$$\gamma^{(i+1)} = \min\left(\gamma^{\max}, \rho\gamma^{(i)}\right) \tag{19}$$

where γ^{max} denotes the maximum value of γ and the scale factor ρ.

Algorithm 1 Solution of the BWRR model with the ADMM algorithm

Input: y, W, μ, γ_0, ρ, K

Initialization:

$f^{(0)} = g^{(0)} = h^{(0)} = 0$; $W = 0$, $i = 0$;

Iteration:

While $(i \leq K)$ do

(1) Update $v_j(\hat{f})$ by solving Eq. (15), $j = 1, 2, \cdots D$;

(2) Update g by solving Eq. (18);

(3) Update h by solving the third subequation of Eq. (7);

(4) Update W by solving Eq. (16);

(5) Update the step-size parameter γ by solving Eq. (19);

(6) $i = i + 1$;

end while

Output:

$f^{(i+1)}$

C. Convergence Analysis and Computational Complexity

Based on the previous analysis and derivation, it can be known that the proposed BWRR model has convex properties. Moreover, since the optimization process is implemented using the ADMM algorithm, the solution for each optimization subproblem is closed. Therefore, the model guarantees convergence to global optimality, which satisfies the Eckstein-Bertsekas condition [50]. In addition, we set the number of iterations to 2. The detailed procedure is given as Algorithm 1. The convergence of Algorithm 1 can be guaranteed since the overall objective function in Eq. (4) is convex with a global optimal solution.

In each iterative calculation of subproblem f, the FFT and inverse FFT transformation are needed. Thus, the computational complexity is $\mathcal{O}(DMN\log(MN))$. Moreover, the computational complexity of subproblems W, g and h is $\mathcal{O}(DMN)$. To this end, if the number of iterations is K, the total computational complexity of the model is $\mathcal{O}(KDMN(\log(MN)+3))$. In view of this, the speed of our algorithm is not very fast, i.e., 3.7373fps.

D. Tracking Framework

The tracking framework is summarised in Algorithm 2. **Position and scale detection:** We follow fDSST [23] to achieve target position and scale detection simultaneously. The accurate scale estimation of targets is a challenging research problem in visual target tracking. Most of the most advanced methods use an exhaustive scale search to estimate the target size, but they are computationally intensive and cannot cope with major changes. Therefore, we refer to the scale adaptive tracking method of fDSST [23] and learn the appearance change caused by the change in the target scale by learning the separate discrimination correlation filter for translation and scale estimation. Then, we apply the learning scale filter at the target position to obtain an accurate estimate of the target size.

Updating and initialisation: It should be noted that in the learning stage, the multichannel input X in Eq. 7 forms the

Algorithm 2 Tracking algorithm of BWRR

1: *Input and Initialization:*
the center of the target is represented by (p_1, p_2) in the first frame image; set the scale of the search target as $m * n$, and initialize W_1, W_2, W_3 and f, g, h.
2: *Tracking:*
While (video is not over) do
(1)Extract multichannel features in the corresponding area.
(2)Perform K iterations of optimization according to Algorithm 1, and update the filter template according to Eq. 7.
(3)Update W by solving Eq. (16).
(4)Calculate and draw a new target area.
end while
3: *Output:*
The tracked video and the video tracking rate in fps.

feature representation of the padded image patch centered at (p_1, p_2) with size $m * n$. Then we calculate the filter response score f_t according to Algorithm 1 and adopt the updating strategy as the traditional DCF method:

$$f_{model} = (1 - \alpha) f_{model} + \alpha f \tag{20}$$

where α is the updating rate. More specifically, as f_{model} is not available in the learning stage for the first frame, we use a predefined mask with only the target region activated to optimise f as in BACF [25] and then initialise $f_{model} = f$ after the learning stage of the first frame.

IV. Experiments and Results

To demonstrate the superiority and effectiveness of our proposed BWRR, we compare it with several state-of-the-art trackers. To better explore the robust performance of BWRR, we conduct comparative experiments on different datasets. Our BWRR is implemented in MATLAB 2017a, and all the experiments are run on a PC equipped with an Intel i7 7700 CPU, 32 GB RAM and a single NVIDIA GTX 1070 GPU.

A. Experimental Datasets

We evaluate the performance of our BWRR and other trackers on six benchmark datasets in this section including OTB50 [51], CVPR2013 [52], OTB100 [52], Temple-Color 128 [53], UAV123 [54] and VOT2016 [55]. OTB50 contains 50 video sequences, while OTB100 contains two times as much, including 25% grayscale sequences. CVPR2013 has one more video than OTB50 and is similar to OTB50. The Temple-Color 128 dataset [53] contains all color sequences, and UAV123 [54] consists of 123 challenging sequences. VOT2016 [55] consists of 60 challenging videos. VOT datasets contain color sequences dominated by short-term data, and it is considered that tracking detection should not be separated at the same time.

To evaluate the performance of our proposed BWRR, a one-pass evaluation (OPE) is used as the evaluation index, as proposed in the OTB benchmarks. Precision plots show the accurate percentage of predicted positions and the ground-truth under different thresholds, and the success plots are measured by an average overlap, accounting for both size and position accuracy [56]. The robustness of the experimental results on OTB is judged by 11 attributes. Different from the OTB datasets, the experimental effect on the VOT datasets is reported against three metrics: Accuracy measures the average overlap ratio between the ground-truth and predicted bounding box achieved by the trackers. Robustness presents the failure rate and expected average overlap(EAO), which is used to estimate the accuracy of the estimated bounding box.

B. Comparison Methods

In this section, we mainly compare our BWRR tracker against 14 state-of-the-art trackers, including the STRCF(HOGCN) [8], ECO-HC [27], LADCF [17], BACF [25], SRDCFdecon [57], Staple+CA [58], SRDCF [7], Staple [21], SAMF+AT [59], SAMF [22], MEEM [60], DSST [23] and KCF [20] with the HOGCN feature and the STRCF(HOG), based on the OTB and CVPR2013 databases. Then, we perform a comparison experiment with the STRCF [8], LADCF [17], ECO [27], ECO-HC [27], CCOT [26] and DSST [23] on the Temple-Color 128 database. Twelve trackers are compared on the UAV123 dataset, including the STRCF [8], LADCF [17], ECO-HC [27], DSST [23], SRDCF [7], MEEM [60], MUSTER, SAMF [22], TLD [61], DSST [23], MOSSE [19] and KCF [20]. Last, we conduct experiments on the BWRR and 10 other trackers, including the STRCF [8], DSST [23], SRDCF [7], SRDCFdecon [57], MDNet-N [62], BACF [25], KCF [20], and so on, based on the VOT datasets with the HOGCN feature. In addition, BWRR underwent comparative experiments with 11 methods based on deep features, including the GFSDCF [63], ECO [27], MDNet [62], CCOT [26], ASRCF [64], HDT [43], HCF [15], DeepSTRCF [8], DeepSRDCF [65], SiamFC [41] and CF-Net [42].

C. Quantitative Analysis on Various Datasets

Results on the OTB50 and CVPR Datasets:

Since OTB50 and CVPR2013 are similar, we analyze the experimental results on OTB50 and CVPR2013 together. Fig. 3 shows the precision and success plots of our BWRR tracker and 14 other trackers with the HOGCN feature on OTB50 and CVPR2013. As can be seen in Fig. 3(a), our BWRR has the best performance in both the precision and success plots, with scores of 0.825 and 0.617 on OTB50. Compared with the STRCF, our BWRR performs better, with a gain of 1.7% and 2.8% in precision and success, respectively. From Fig. 3(b), our BWRR also achieves the best performance among the trackers on the CVPR2013 dataset. The precision score is 0.903, which is 3.3% higher than that of the STRCF(HOGCN) and the success score is 0.697, which is 3.57% higher than that of the STRCF. Compared with the results in Fig. 3, the BWRR performs better on CVPR2013 with the HOG feature.

Since our BWRR is improved by adding weight matrices and a sparse term to the STRCF, we compare the scores of BWRR and the STRCF in terms of different attributes to better reflect the superiority of the BWRR. The comparison

TABLE I
THE COMPARISON OF BWRR AND STRCF IN DIFFERENCE ATTRIBUTES ON OTB50

Attributes	Success plots			Precision plots		
	BWRR	STRCF(HOGCN)	STRCF(HOG)	BWRR	STRCF(HOGCN)	STRCF(HOG)
OPE	**0.617**	0.600	0.580	**0.825**	0.811	0.762
baceground clutter	**0.649**	0.682	0.537	**0.861**	0.843	0.674
deformation	0.556	0.533	0.519	0.769	0.764	0.701
fast motion	0.581	0.559	0.571	0.757	0.726	0.738
illumination variation	**0.616**	0.573	0.536	**0.820**	0.763	0.667
in plane rotation	**0.596**	0.576	0.534	**0.797**	0.771	0.718
low resolution	**0.576**	0.563	0.560	0.822	0.830	0.826
motion blur	**0.601**	0.552	0.576	**0.794**	0.721	0.748
occlusion	**0.602**	0.571	0.558	0.815	0.784	0.733
out of plane rotation	**0.600**	0.569	0.532	**0.804**	0.775	0.698
out of view	**0.570**	0.537	0.537	0.771	0.725	0.742
scale variation	**0.595**	0.573	0.569	0.793	0.782	0.756

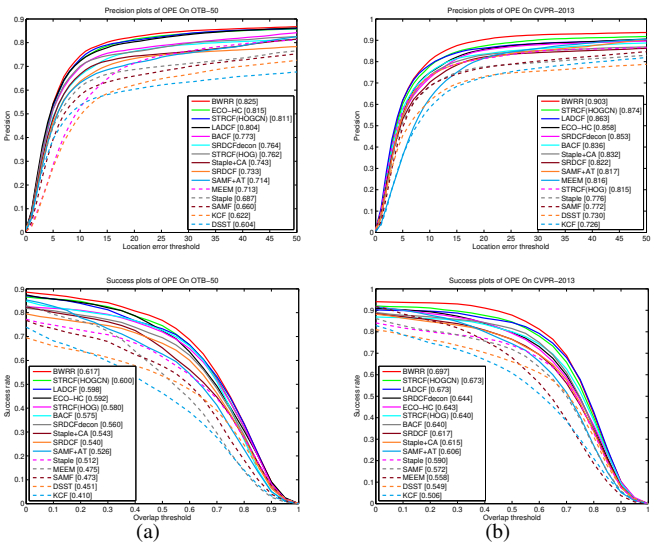


Fig. 3. Precision and success plots of all trackers with HOGCN features on (a) OTB50 dataset and (b) CVPR2013 dataset.

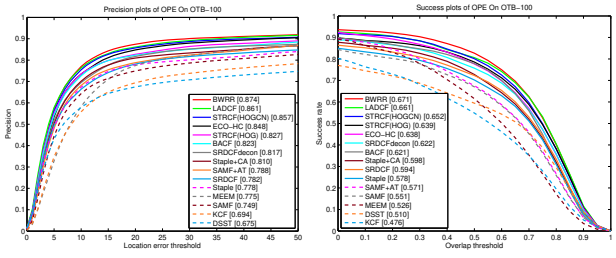


Fig. 4. Precision and success plots of the OPE on the OTB100 dataset with HOGCN features.

results are shown in Table I. The best results are marked in bold, which indicate that the BWRR performs better than the STRCF in terms of the corresponding attributes. As can be seen from Table I, the OPE scores of BWRR are much higher than those of the STRCF, and in terms of most of the attributes, the BWRR performs better than the STRCF, indicating that our BWRR does have a better tracking effect than that of the STRCF.

Results on OTB100: The results of the BWRR and the other trackers on the HOGCN feature are provided in Fig. 4. Our proposed BWRR tracker achieves a precision score of 0.874 and a success score of 0.671, both of which both are the best among all trackers. Compared with the STRCF, which takes third place based on the success plots, with a precision score of 0.857 and a success score of 0.652, our BWRR tracker shows improvements of almost 2% and 2.9%, respectively.

Similar to the results based on OTB50, the BWRR results based on the OTB100 dataset also show a very good the tracking effect on the attribute, as presented in Fig. 5. Combined with the sparse regularization term and weighted matrices, the proposed BWRR performs favorably against the state-of-the-art trackers. We selected 8 different attributes to discuss the tracking performance of BWRR in a complex environment. The results prove that our BWRR achieves the best tracking effect in different situations.

For clear visualization, we obtain the tracking results of the BWRR (red wire frame), STRCF (green wire frame) and LADCF (blue wire frame) on 3 challenging video sequences for comparison, as shown in Fig. 6. For these three video sequences, the difficulty of tracking is mainly caused by occlusion, fast movement and illumination changes. Our method successfully tracks the object every time in all 3 video sequences. However, the STRCF and LADCF both have different degrees of tracking deviation and even experience tracking failure, such as in the bird video sequences. The result shows the accuracy and robustness of BWRR for video sequences with challenging factors.

Results on the Temple-Color Dataset: We also present the results of our BWRR and other state-of-the-art trackers(i.e. CCOT [26], ECO [27], ECO-HC [27], STRCF [8] LADCF [17] and DSST [23]) on the Temple-Color dataset [53] in Fig. 7. The figure shows a comparison of the overlap success plots for all trackers. Though the performance of the BWRR is not as good as that of the ECO [27] and CCOT [26], the score of the BWRR surpasses that of its counterpart LADCF by 1.3% with the HOGCN feature.

Results on the UAV123 Dataset: We evaluate our tracker on a dataset designed for low-altitude UAV tracking. Fig. 8 shows the precision and success plots of all trackers. Among the existing methods (except ECO-HC), our BWRR achieves the best performance, with a score of 0.635 and 0.468 in

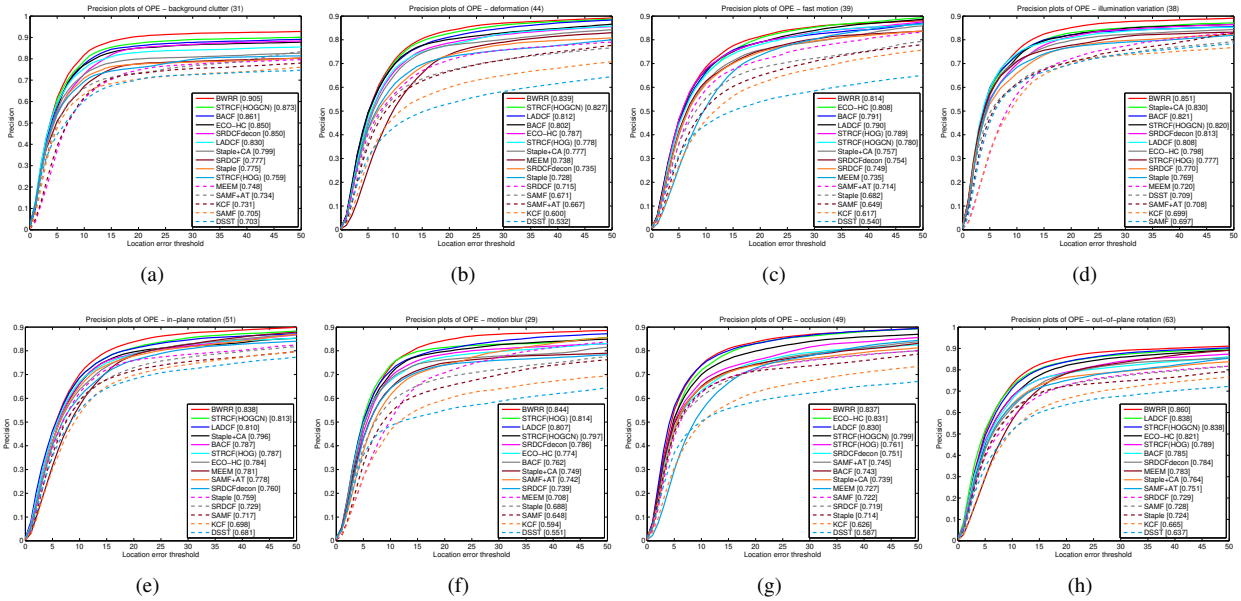


Fig. 5. The comparison of success plots on the OTB100 for the subset of challenging attributes: background clutter, deformation, fast motion, illumination variation, in plane rotation, motion blur, occlusion and out of plane rotation.

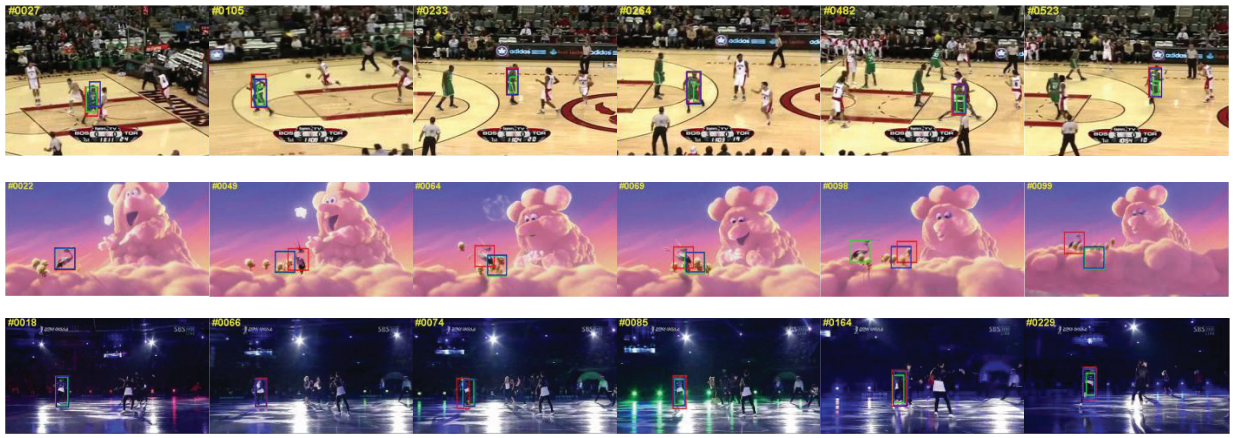


Fig. 6. The comparison of tracking for BWRR, STRCF and LADCF on 3 video sequences.

TABLE II
A COMPARISON WITH THE STATE-OF-THE-ART TRACKERS ON VOT-2016 DATASET

	KCF	DSST	STRCF	MDNet-N	BACF	SRDCF	SRDCFdecon	DPT	HCF	SHCF	BWRR
EAO↑	0.153	0.1811	0.279	0.2572	0.223	0.2471	0.267	0.235	0.231	0.267	0.289
Accuracy↑	0.412	0.537	0.53	0.5421	0.56	0.5364	0.513	0.483	0.467	0.54	0.5402
Robustness↓	2.67	2.52	1.32	1.2	1.89	1.5	1.08	0.75	1.389	1.4	1.37

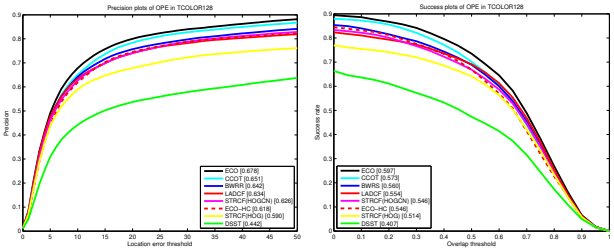


Fig. 7. Precision and success plots of OPE on the Temple-Color dataset with HOGCN features.

terms of the precision and success, respectively. Our BWRR outperforms the STRCF, with an AUC of 2.41%.

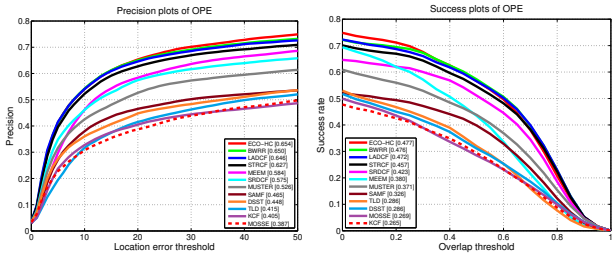
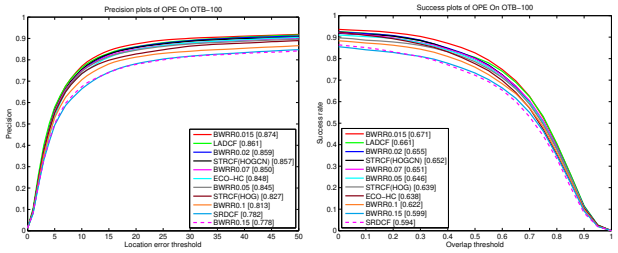


Fig. 8. Precision and success plots of the OPE on UAV123 with HOGCN features.

Results on VOT2016: The results on the VOT2016 benchmark are shown in Table II. The top three results are marked in red, green and blue respectively. We evaluate the trackers, including the BWRR, STRCF [8], DSST [23], SRDCF [7], SRDCFdecon [7], MDNet-N [62], KCF [20], and so on, with HOGCN features in terms of the accuracy, robustness and expected average overlap (EAO) to show the effect of each tracker. From Table II, the performance of our BWRR in terms of the EAO and accuracy is second and third best, respectively, among all trackers. Compared with the STRCF (HOGCN), the BWRR performs better in both accuracy and EAO, with a gain of 3.6% and 1.9%, respectively.

D. Ablation Analysis



(a) (b)

Fig. 9. The precision plots (left) and success plots (right) of the OPE for BWRR variants on the OTB100 dataset with HOGCN features.

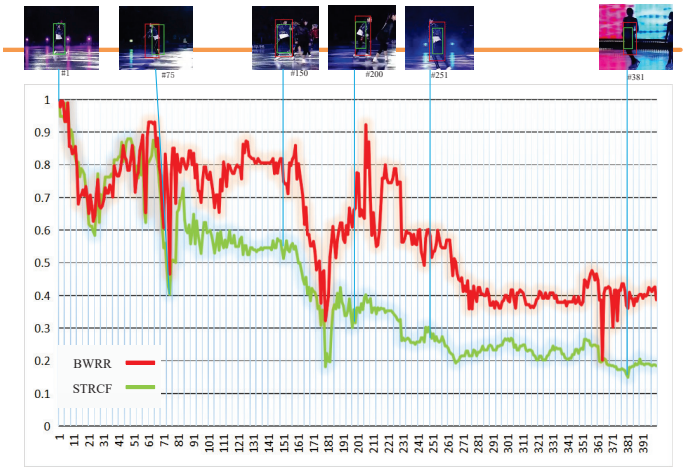


Fig. 10. The robustness analysis for BWRR, and STRCF on video sequences.

TABLE III
THE REULTS OF BWRR VARIANTS BASED ON OTB100

a	precision plot	success plot
0.015(*)	0.874	0.671
0.02(*)	0.859	0.655
0.05	0.845	0.646
0.07	0.850	0.651
0.01	0.813	0.622
0.15	0.788	0.599

Parameter Analysis: The setting of the parameters also has a great influence on the tracking effect of the experiment. The updating of matrix W in Section IV involves the parameter design. We analyze the different benefits of the different values of a in Eq. (17). We conduct experiments on OTB100, with a set to 0.15, 0.1, 0.07, 0.05, 0.02 and 0.015. As shown in Fig. 9, the results of the BWRR variants are compared with those of the ECO-HC [27], STRCF [8], SRDCF [7] and LADCF [17]. Note that the model becomes more stable and performs better when the value of a is smaller. The detailed scores of the success plots and precision plots of the trackers are exhibited in Table III based on the HOGCN feature, where $*$ indicates that the results are better than those of the STRCF. The different variants also perform better than the STRCF when the parameter a is smaller than 0.015.

Robustness Analysis: To reflect the good robustness of the BWRR model in target tracking, we compare the tracking robustness of the BWRR and STRCF frame by frame. As shown in Fig. 10, we use the overlap ratio between the estimated and ground-truth bounding boxes of each frame to reflect the robustness for each frame of the model, that is, the ordinate in the figure. Due to continuous movement, the posture of the target continues to change in the video sequence. There are also complex situations, such as lighting changes and background clutter, in the scene. As shown in Fig. 10, due to the rotation of the target person in frame 75, the overlap rates of the STRCF and BWRR fluctuate sharply, though the rate decline of the STRCF is more serious. In addition, thanks to the adjustment of the weighting matrix, the BWRR gradually recovers to a higher overlap rate in subsequent frames and remains relatively stable, a better behavior than that of the STRCF. It can be concluded that although BWRR reduces the tracking effect due to complex environments, it has the ability to recover and maintain a high tracking overlap rate. Therefore, BWRR is more robust.

E. Experimental Analysis with Deep Features

TABLE IV
THE RESULTS ON OTB100 UNDER DIFFERENT FEATURE CONFIGURATIONS

Features		AUC score	Threshold score
Handcrafted	HOG	0.622	0.812
	HOGCN	0.671	0.874
Handcrafted+CNN	HOGCN+Conv-1	0.673	0.877
	HOGCN+Conv-2	0.677	0.881
	HOGCN+Conv-3	0.682	0.891
	HOGCN+Conv-4	0.691	0.905
	HOGCN+Conv-5	0.687	0.900

Deep Feature Configuration Analysis: In the experiment, we choose the 19-layer deep convolutional neural network VGG19 (Visual Geometry Group). VGGNet [66] explors the relationship between the depth of a convolutional neural network and its performance. VGGNet is still often used to extract image features. According to the size of the convolutional layer, features of different contexts can be obtained. We compare the specific convolutional layers through experiments. To provide a deep analysis for our BWRR method, we employ 7 feature configurations to perform experiments on OTB100 using the AUC and threshold metrics. The AUC

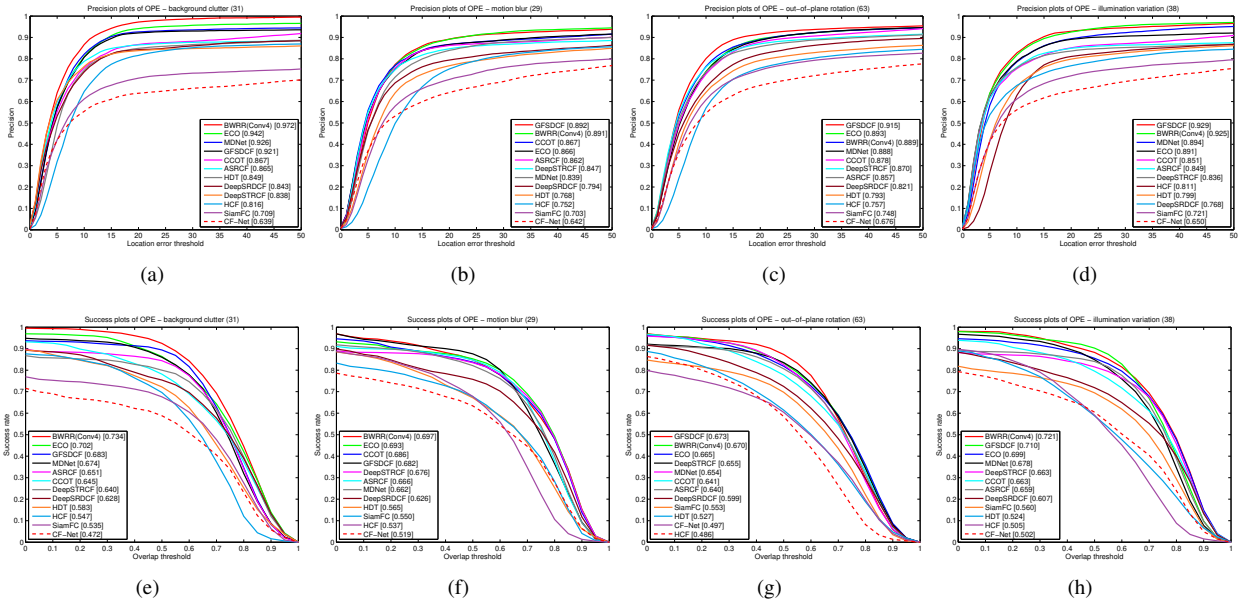


Fig. 11. The comparison of precision and success plots on the OTB100 with deep feature for the subset of challenging attributes: background clutter, motion blur, out of plane rotation and illumination variation.

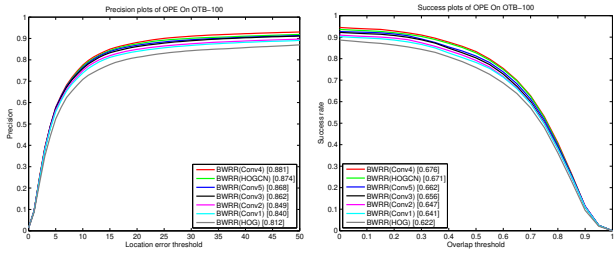


Fig. 12. The results of the OPE on the OTB100 with different feature configurations.

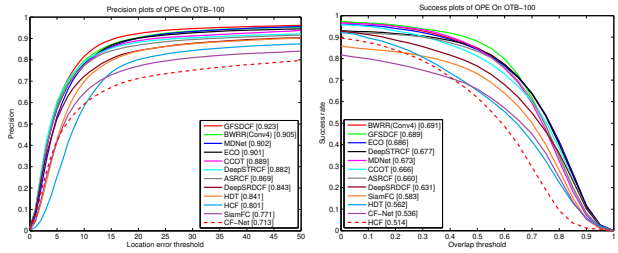


Fig. 13. Precision and success plots of the OPE on the OTB100 dataset with deep features.

and threshold scores under different feature configurations are shown in Table IV. The BWRR with the colour names and HOG features achieves a 4.9% improvement in the AUC and a 6.2% improvement in the threshold score. To better analyze the influence of the CN features on visual tracking, we set up different layers of neural networks in the learning process of the CN features. It can be seen from Table IV that the AUC and threshold scores for tracking also increase in more deeper network representation. As shown in Fig. 12, the middle convolutional layer (Conv-4) has the best performance. The high layers (Conv-3, Conv-4 and Conv-5) significantly improve the performance compared with the low layers(Conv-1 and Conv-2).

Results on OTB100 with Deep Features: To show the tracking performance of BWRR with deep features, we compare it with other 11 trackers, including the GFSDCF, ECO, MDNet, CCOT, ASRCF, HDT, HCF, DeepSTRCF, DeepSRDCF, SiamFC, and CF-Net. Among them, the deep network used by the BWRR is the optimal network mentioned in the previous section, that is, the deep network based on Conv-4.

As shown in Fig. 13, BWRR performs quiet well in terms of tracking. On the one hand, the tracking performance of the BWRR ranks first among all compared models, with a score of 0.691, regarding the success plots, which is 0.29% higher than that of the GFSDCF model (second best) and 0.73% higher than that of the ECO (third best). On the other hand, BWRR ranks second, with a score of 0.905, regarding the precision plots, which is very similar to that of the top-ranked GFSDCF. In addition, the precision score of the BWRR is 0.44% higher than that of the ECO. Therefore, it is reasonable to believe that BWRR still performs better than other trackers with deep features.

To better discuss the tracking performance of BWRR in a complex environment, we select four attributes for analysis, the results of which are shown in Fig. 11. Except for the out-of-plane rotation, our BWRR achieves the best performance in terms of the success plots among all trackers, which proves that BWRR can capture the target well in complex environments without tracking drift or failure, especially with regard to the background. BWRR has the highest precision and success scores in the case of background clutter. As shown in Fig. 11(a), the precision of our BWRR is 3.18% higher than that of the ECO, achieving a score of 0.972. Although the rank of the BWRR in the precision plot is not as good as that in the success plot for the other three attributes, its score is still very competitive compared to those of the other trackers. Except

for the out-of-plane rotation, BWRR maintains the second-best performance under motion blur and illumination variation, with scores close to those of the GFSDCF, which ranks first.

V. Conclusions

In this paper, a bilateral weighted regression ranking (BWRR) algorithm that introduces two weighted matrices into the data fidelity term and a sparse term to achieve a more stable model and more robust visual tracking is proposed in this paper. The update of the weight matrix is obtained by ranking and numerically transforming the matrix elements. In the process of model optimization, the least squares regression equation is used to solve the filter update problem, and the ADMM algorithm is employed to solve the whole iterative process to reduce the computational complexity. The experiment results demonstrate the superiority of our model.

References

[1] H. Yang, L. Shao, F. Zheng, L. Wang, and Z. Song, "Recent advances and trends in visual tracking: A review," *Neurocomputing*, vol. 74, no. 18, pp. 3823–3831, 2011.

[2] Y. Wu, J. Lim, and M. H. Yang, "Online object tracking: A benchmark," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2411–2418, 2013.

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

[4] M. Kristan, R. Pflugfelder, J. Matas, A. Leonardis, M. Felsberg, L. Cehovin, G. Fernandez, T. Vojir, G. Hager, and G. Nebehay, "The visual object tracking vot2015 challenge results," in *IEEE International Conference on Computer Vision Workshop*, pp. 564–586, 2015.

[5] M. Kristan, A. Leonardis, J. Matas, M. Felsberg, R. Pflugfelder, L. ehovin, T. Vojr, G. Hger, A. Lukei, and G. Fernndez, "The visual object tracking vot2016 challenge results," *Computer Vision ECCV 2016 Workshops*, vol. 8926, no. 191-217, 2016.

[6] G. Xu, H. Zhu, L. Deng, L. Han, Y. Li, and H. Lu, "Dilated-aware discriminative correlation filter for visual tracking," in *World Wide Web 22*, vol. 2, pp. 791–805, 2019.

[7] M. Danelljan, G. Hager, F. S. Khan, and M. Felsberg, "Learning spatially regularized correlation filters for visual tracking," in *IEEE international conference on computer vision*, pp. 4310–4318, 2015.

[8] F. Li, C. Tian, W. Zuo, L. Zhang, and M. H. Yang, "Learning spatial-temporal regularized correlation filters for visual tracking," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4904–4913, 2018.

[9] J. Zhang, P. Li, C. Jin, W. Zhang, and S. Liu, "A novel adaptive kalman filtering approach to human motion tracking with magnetic-inertial sensors," *IEEE Transactions on Industrial Electronics*, 2019.

[10] J. Xu, L. Zhang, and D. Zhang, "A trilateral weighted sparse coding scheme for real-world image denoising," *CoRR*, vol. abs/1807.04364, 2018.

[11] Z. Feng, J. Kittler, M. Awais, P. Huber, and X. Wu, "Wing loss for robust facial landmark localisation with convolutional neural networks," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, DOI 10.1109/CVPR.2018.00238, pp. 2235–2245, 2018.

[12] X. Lu, D. Yuan, Z. He, and D. Li, "Sparse selective kernelized correlation filter model for visual object tracking," in *2017 International Conference on Security, Pattern Analysis, and Cybernetics (SPAC)*, pp. 100–105, 2017.

[13] W. Zhong, H. Lu, and M. Yang, "Robust object tracking via sparse collaborative appearance model," *IEEE Transactions on Image Processing*, vol. 23, no. 5, pp. 2356–2368, 2014.

[14] N. Wang, S. Li, A. Gupta, and et al., "Transferring rich feature hierarchies for robust visual tracking," *Computer Science*, 2015.

[15] C. Ma, J. B. Huang, X. Yang, and M. H. Yang, "Hierarchical convolutional features for visual tracking," in *IEEE International Conference on Computer Vision*, pp. 1–6, 2015.

[16] Q. Qian, L. Chen, H. Li, and R. Jin, "Dr loss: Improving object detection by distributional ranking," *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, DOI 10.1109/CVPR42600.2020.01218, pp. 12 161–12 169, 2020.

[17] T. Xu, Z. Feng, X. Wu, and J. Kittler, "Learning adaptive discriminative correlation filters via temporal consistency preserving spatial feature selection for robust visual object tracking," *IEEE Transactions on Image Processing*, vol. 28, no. 11, pp. 5596–5609, Nov. 2019.

[18] H. Song, "Robust visual tracking via online informative feature selection," *Electronics Letters*, vol. 50, no. 25, pp. 1931–1933, 2014.

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

[20] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista, "High-speed tracking with kernelized correlation filters," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 3, pp. 583–596, 2014.

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

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

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

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

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

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

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

[28] S. Khan, G. Xu, R. Chan, and H. Yan, "An online spatio-temporal tensor learning model for visual tracking and its applications to facial expression recognition," *Expert Systems with Applications*, vol. 90, pp. 427–428, 2017.

[29] G. Xu, S. Khan, H. Zhu, L. Han, M. Ng, and H. Yan, "Discriminative tracking via supervised tensor learning," in *Neurocomputing*, pp. 33–47, 2018.

[30] Q. Wang, C. Yuan, J. Wang, and W. Zeng, "Learning attentional recurrent neural network for visual tracking," *IEEE Transactions on Multimedia*, vol. 21, no. 4, pp. 930–942, 2019.

[31] H. Hu, B. Ma, J. Shen, H. Sun, L. Shao, and F. Porikli, "Robust object tracking using manifold regularized convolutional neural networks," *IEEE Transactions on Multimedia*, vol. 21, no. 2, pp. 510–521, 2019.

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

[33] N. Liang, G. Wu, W. Kang, Z. Wang, and D. D. Feng, "Real-time long-term tracking with prediction-detection-correction," *IEEE Transactions on Multimedia*, vol. 20, no. 9, pp. 2289–2302, 2018.

[34] W. Zuo, X. Wu, L. Lin, L. Zhang, and M. Yang, "Learning support correlation filters for visual tracking," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, DOI 10.1109/TPAMI.2018.2829180, no. 5, pp. 1158–1172, 2019.

[35] Y. Zheng, L. Sun, S. Wang, J. Zhang, and J. Ning, "Spatially regularized structural support vector machine for robust visual tracking," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 10, pp. 3024–3034, Oct. 2019.

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

[37] X. Dong, J. Shen, D. Yu, W. Wang, J. Liu, and H. Huang, "Occlusion-aware real-time object tracking," *IEEE Transactions on Multimedia*, vol. 19, no. 4, pp. 763–771, 2017.

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

[39] G. Liu, "Robust visual tracking via smooth manifold kernel sparse learning," *IEEE Transactions on Multimedia*, vol. 20, no. 11, pp. 2949–2963, 2018.

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

[41] L. Bertinetto, J. Valmadre, J. F. Henriques, A. Vedaldi, and P. H. Torr, "Fully-convolutional siamese networks for object tracking," in *European Conference on Computer Vision*, p. 6, 2016.

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

[43] Y. Qi, S. Zhang, L. Qin, H. Yao, Q. Huang, J. Lim, and M. H.Yang, "Hedged deep tracking," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2–6, 2016.

[44] M. Danelljan, G. Hager, F. S. Khan, and M. Felsberg, "Convolutional features for correlation filter based visual tracking," in *IEEE International Conference on Computer Vision Workshop*, 2015.

[45] R. Tao, E. Gavves, and A. W. Smeulders, "Siamese instance search for tracking," in *IEEE Conference on Computer Vision and Pattern Recognition*, p. 1420¨C1429, 2016.

[46] B. Hu, Z. Guan, F. Lewis, and C. L. P. Chen, "Adaptive tracking control of cooperative robot manipulators with markovian switched couplings," *IEEE Transactions on Industrial Electronics*, 2020.

[47] X. Mei and H. B. Ling, "Robust visual tracking using ℓ_1 minimization," in *IEEE International Conference on Computer Vision*, pp. 1436–1443, 2009.

[48] K. Crammer, O. Dekel, J. Keshet, S. Shalev-Shwartz, and Y. Singer, "Online passive-aggressive algorithms," *Journal of Machine Learning Research*, vol. 7, no. 3, pp. 551–585, 2006.

[49] Y. Yang, H. T. Shen, Z. Ma, Z. Huang, and X. Zhou, "$\ell_{2,1}$-norm regularized discriminative feature selection for unsupervised learning," in *international joint conference on artificial intelligence*, vol. 1, p. 1589, 22 2011.

[50] J. Eckstein and D. P. Bertsekas, "On the douglas-rachford splitting method and the proximal point algorithm for maxmal monotone operators," *Mathematical Programming*, vol. 55, no. 1-3, pp. 293–318, May. 1992.

[51] Y. Wu, J. Lim, and M. Yang, "Online object tracking: A benchmark," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2411–2418, 2013.

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

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

[54] M. Mueller, N. Smith, and B. Ghanem, "A benchmark and simulator for uav tracking," in *European conference on computer vision*, p. 445¨C461, 2016.

[55] M. Kristan, J. Matas, A. Leonardis, M. Felsberg, L. Cehovin, G. Fernandez, T. Vojir, G. Hager, G. Nebehay, and R. Pflugfelder, "The visual object tracking vot2016 challenge results," in *Eurpean Conference on Computer Vision Workshops*, vol. 2, p. 9, 2016.

[56] L. Cehovin, M. Kristan, and A. Leonardis, "Is my new tracker really better than yours?" *Applications of Computer Vision*, pp. 540–547, 2014.

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

[58] M. Mueller, N. Smith, and B. Ghanem, "Context-aware correlation filter tracking," in *IEEE Conference on Computer Vision and PatternRecognition*, pp. 1387–1395, 2017.

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

[60] J. Zhang, S. Ma, and S. Sclaroff, "Meem: robust tracking via multiple experts using entropy minimization," in *European Conference on Computer Vision*, pp. 188–203, 2014.

[61] Z. Kalal, K. Mikolajczyk, and J. Matas, "Tracking-learning-detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, DOI 10.1109/TPAMI.2011.239, no. 7, pp. 1409–1422, 2012.

[62] H. Nam and B. Han, "Learning multi-domain convolutional neural networks for visual tracking," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, DOI 10.1109/CVPR.2016.465, pp. 4293–4302, 2016.

[63] T. Xu, Z. Feng, X. Wu, and J. Kittler, "Joint group feature selection and discriminative filter learning for robust visual object tracking," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, DOI 10.1109/ICCV.2019.00804, pp. 7949–7959, 2019.

[64] K. Dai, D. Wang, H. Lu, C. Sun, and J. Li, "Visual tracking via adaptive spatially-regularized correlation filters," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[65] M. Danelljan, G. Hager, F. S. Khan, and M. Felsberg, "Convolutional features for correlation filter based visual tracking," in *2015 IEEE International Conference on Computer Vision Workshop (ICCVW)*, 2016.

[66] K. Simoyan and A.Zisserman, "Very deep convolutional networks for large-scale image recognition," *Computer Science*, 2014.

Hu Zhu (M'17) received the B.S. degree in mathematics and applied mathematics from Huaibei Coal Industry Teachers College, Huaibei, China, in 2007, and the M.S. and Ph.D. degrees in computational mathematics and pattern recognition and intelligent systems from Huazhong University of Science and Technology, Wuhan, China, in 2009 and 2013, respectively. In 2013, he joined the Nanjing University of Posts and Telecommunications, Nanjing, China. His research interests include pattern recognition, image processing, and computer vision.

Hao Peng received B.S. degree in communication engineering from Jiangsu University of Technology, Changzhou, China, in 2018. He is now pursuing his master degree in electronic and communication engineering in Nanjing University of Posts and Telecommunications. His current research interest is target tracking based on correlation filtering.

Guoxia Xu (M'19) received the B.S. degree in information and computer science from Yancheng Teachers University, Jiangsu Yancheng, China in 2015, and the M.S. degree in computer science and technology from Hohai University, Nanjing, China in 2018. He was a research assistant in City University of Hong Kong and Chinese University of Hong Kong. Now, he is pursing his Ph.D. degree in Department of Computer Science, Norwegian University of Science and Technology, Gjovik Norway. His research interest includes pattern recognition, image processing, and computer vision.

Lizhen Deng (M'17) received the B.S. degree in electronic information science and technology from Huaibei Coal Industry Teachers College, Huaibei, China, in 2007, and the M.S. degree in communication and information systems from Nanjing University of Aeronautics and Astronautics, Nanjing, China, in 2010. She received her Ph.D. degree in electrical engineering from Huazhong University of Science and Technology, China, in 2014. In 2014, she joined the Nanjing University of Posts and Telecommunications, Nanjing, China. Her current research interests include image processing, computer vision, pattern recognition, and spectral data processing.

Yueying Cheng received B.S. degree in electronic science and technology from Chongqing University of Posts and Telecommunications, Chongqing, China, in 2018. She is now pursuing his master degree in signal and information processing in Nanjing University of Posts and Telecommunications. Her current research interest is visual tracking.

Aiguo Song (M'98-SM'12) received the Ph.D. degree from the School of Instrument Science and Engineering, Southeast University, Nanjing, China, in 1996. He is currently a Professor and Dean with the School of Instrument Science and Engineering, Southeast University. He is also the Director with the Jiangsu Key Laboratory of Remote Measurement and Control. He has authored and coauthored two books, and published more than 160 papers in international journals and conference proceedings. He is the recipient of the National Science Fund for Distinguished Young Scholars in 2013 and won the Second Prize of 2017 National Award for Technological Invention. Additionally, he hosted one National Key Research and Development Program of China, one National Key Basic Research Program of China (973 Program), ten National High Technology Research and Development Program of China (863 Program), two Key Program of the National Natural Science Foundation of China, etc.