IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received March 22, 2021, accepted April 16, 2021, date of publication April 22, 2021, date of current version May 3, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3075089

Multi-Channel Feature Dimension Adaption for Correlation Tracking

LINGYUE WU[1,2], **TINGFA XU**[1,2,3], **YUSHAN ZHANG**[1,2], **FAN WU**[1,2], **CHANG XU**[1,2], **XIANGMIN LI**[1,3], **AND JIHUI WANG**[1,3]

[1] Image Engineering and Video Technology Laboratory, School of Optics and Photonics, Beijing Institute of Technology, Beijing 100081, China
[2] Beijing Institute of Technology Chongqing Innovation Center, Chongqing 401120, China
[3] Key Laboratory of Photoelectronic Imaging Technology and System, Ministry of Education of China, Beijing 100081, China

Corresponding author: Tingfa Xu (xutingfa@163.com)

This work was supported by the Key Laboratory Foundation under Grant TCGZ2020C004.

ABSTRACT Recent discriminative trackers especially based on Correlation Filters (CFs) have shown dominant performance for visual tracking. This kind of trackers benefit from multi-resolution deep features a lot, taking the expressive power of deep Convolutional Neural Networks (CNN). However, distractors in complex scenarios, such as similar targets, occlusion, and deformation, lead to model drift. Meanwhile, learning deep features results in feature redundancy that the increasing number of learning parameters introduces the risk of over-fitting. In this paper, we propose a discriminative CFs based visual tracking method, called dimension adaption correlation filters (DACF). First, the framework adopts the multi-channel deep CNN features to obtain a discriminative sample appearance model, resisting the background clutters. Moreover, a dimension adaption operation is introduced to reduce relatively irrelevant parameters as possible, which tackles the issue of over-fitting and promotes the model effectively adapting to different tracking scenes. Furthermore, the DACF formulation optimization can be efficiently performed on the basis of implementing the alternating direction method of multipliers (ADMM). Extensive evaluations are conducted on benchmarks, including OTB2013, OTB2015, VOT2016, and UAV123. The experiments results show that our tracker gains remarkable performance. Especially, DACF obtains an AUC score of 0.698 on OTB2015.

INDEX TERMS Correlation filters, multi-channel feature learning, object tracking.

I. INTRODUCTION

Visual tracking is one of the fundamental computer vision tasks that has received much attention [1]–[3]. It is a task aiming to continually detect a target in a video sequence only its initial position is given. It has numerous real-world applications, including vehicle tracking [4], automatic surveillance [5], and pedestrian tracking [6], [7]. However, it is suffering from some challenging visual attributes, such as background clutters, occlusions [8], motion changes, and size changes. Therefore, an ideal tracker is designed to be robust and efficient.

Most tracking methods observation models are based on either generative or discriminative models. Generative methods aim to find the best-matching candidate of the object. Meanwhile, discriminative trackers distinguish the object from the surrounding background by training a discriminator. These methods concern more about the difference between object and surrounding while generative models care about the exact similarity between the candidates and the target. Overall, discriminative trackers are more accurate than generative trackers. Recently, Discriminative Correlation Filters (DCFs) based tracking algorithms have achieved more advancing performance [9]–[11] than traditional discriminative approaches. Two advantages contribute to the advancement of DCF approaches. First, the DCF approaches can make use of a large number of samples benefiting from the circulant structure for training and prediction. Second, the DCF based trackers learn models quickly in the frequency domain. The convolution operation in the time domain corresponds to the element-wise product in the frequency domain that makes the calculation more efficient. This transformation from the time domain to the Fourier domain can be implemented through the Fast Fourier Transform (FFT). Moreover, it contributes to the combination of multiple features such as Histogram of Oriented Gradient (HOG) [12], Color

The associate editor coordinating the review of this manuscript and approving it for publication was Junchi Yan.

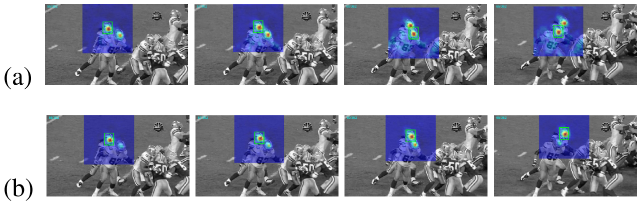


FIGURE 1. Tracking results with confidence score maps of sequence *Football* where there are similar targets. (a) Similar targets in the searching region disturb the observation model of the tracker employing shallow features presentation. (b) Our method with multi-channel deep features has great discriminative power and detects the right target.

Names [13], and deep convolution features learned from VGGnet [14].

However, the DCF approach still has some inherent deficiencies to address. The circulant structure produces sufficient samples through circular shifting from a base search area that includes the object. Setting an appropriate search area size is significant in standard DCF for producing sufficient samples. If the search area is small, objects with fast motion may easily move to or even outside the research boundary that usually results in inaccurate detection. To deal with this problem, an extension of the image patch is introduced that the base sample patch is bigger than the true boundary box of the original object. But in this way, extra background information is included in samples, which would cut down the tracking performance of the learned model. The observation model is often distracted by wrong information of the background like similar targets or background clutters in the search area. As shown in Fig. 1, the tracker cannot discriminate the right target among multi similar targets, leading to inferior tracking results.

Advanced DCF based trackers benefit a lot from rich features representations and perplexing learning formulations. The deep features will provide a robust appearance model to deal with severe appearance variations. Meanwhile, DCF based trackers obtain discriminative power from learning deep features to resist distractors in complex scenarios. More attention has been shifted from handcraft features towards deeper features learned from deep Convolutional Neural Networks (CNN) networks to capture multi-level information. However, most present DCF based trackers use only relatively shallow pretrained CNN networks like VGGnet. Using deeper networks (e.g., ResNet [15], DenseNet [16], and SE-ResNet [17]) to learn convolutional features is not completely researched.

Nevertheless, most of the DCF based trackers fuse all the multi-resolution deep features directly, introducing severe data redundancy. This comes at the high cost of numerous parameters learning and frequent online updating. Limited samples and increasing training parameters raise the harm of over-fitting. This issue is tackled in ECO, under a factorized convolution approach. But in ECO, filter channels are reduced to fixed dimensions for different objects, which cannot fully take advantage of the diversity information of different objects tracking scenes.

In this work, we develop a robust and efficient tracker called dimension adaption correlation filters (DACF) for making full use of multi-level deep CNN features and address the problems of over-fitting. The framework of this proposed method is depicted in Fig. 2.

In the feature representation stage, we utilize multi-channel deep convolutional features for visual object tracking with the aim of alleviating model drift caused by complex scenes. We investigate the multi-channel deep convolutional features to exploit the great power of it. DACF tracker only makes use of deep features to obtain sample appearance model, without any handcraft features such as HOG or Color Names.

Moreover, we propose a dimension adaption component to adaptively adopt part effective multi-level features during different tracking scenarios. Benefiting from this dimension adaption method, our DACF reduces the number of parameters without excessive information loss, which simultaneously remits the problems of over-fitting.

In optimization process, we solve the problem efficiently by alternating direction method of multipliers (ADMM) within very few iterations.

We perform extensive experiments on OTB2013 [19], OTB2015 [20], VOT2016 [40], and UAV123 [41] benchmarks. The experiments results demonstrate that our DACF tracker achieves notable performance in comparison to the state-of-the-art trackers. Our approach outperforms the baseline STRCF both in accuracy and robustness on OTB2013 and OTB2015 benchmarks.

II. RELATED WORK

CNNs [14] have been widely employed in computer vision tasks. Visual tracking tasks with CNNs perform excellently in recent years. Many tracking methods design deep architectures to learn end-to-end trackers that need to provide a large amount of training data. MDNet [21] combines multi-domain networks to differentiate between the target and background. Some methods train a Siamese net to distinguish whether objects in two images are the same or not. SiamRPN [22] introduces Siamese subnetwork for feature extraction and specialized subnetwork including two branches of classification and regression to generate region proposal. Based on SiamRPN, DaSiamRPN [23] designs a distractor-aware module to improve the discriminant ability. SiamRPN++ [24] applies deep benchmark networks such as ResNet [15] and Inception [25] into tracking networks based on Siamese Network.

The DCFs for visual tracking have been popularized in recent years and many discriminative trackers have been proposed. The earliest proposed CF tracker is learned by minimizing the output sum of squared error (MOSSE) [26]. MOSSE uses only gray-scale samples to train the filter at high speed. CSK [27] drives a circulant structure for dense sampling and uses the kernel matrix in ridge regression. The circulant structure produces sufficient samples and simplifies the ridge regression problem. Only providing the initial position of the object limits obtaining positive samples in

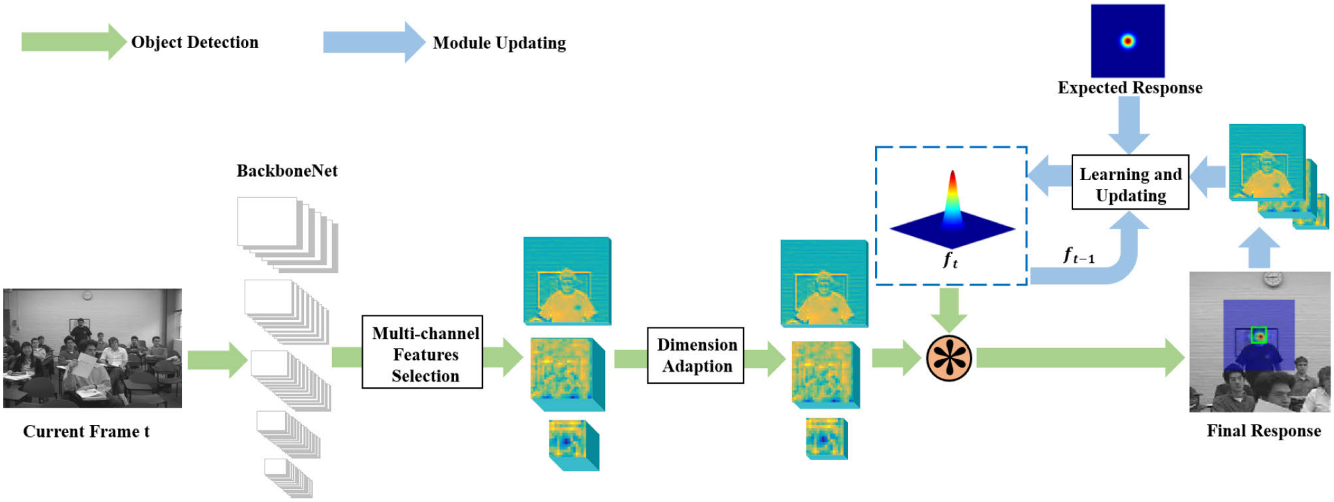


FIGURE 2. Overview of the framework of our tracking method. The whole architecture includes multi-channel features selection and the dimension adaption operation. The framework extracts multi-channel deep convolution features with discriminating power. Through the dimension adaption operation, the features dimensions are drastically reduced, adapting to various scenario. The green arrows indicate the process of object detection. The blue arrows indicate the process of module updating that learns the filter of current frame t using the filter learned in previous frame $t - 1$.

visual tracking. It will be useful to utilize more samples. The DCF based trackers can make use of more samples than conventional discriminative approaches, profiting from the circulant structure. KCF [28] adopts a linear kernel in a fast multi-channel extension of linear CFs, thereby improving the precision and robustness. More discriminative features are used for incorporating multi-channel features in the Fourier domain, such as HOG [12], color names [13], and deep CNN features. C-COT [29] learns filters from a integration of multi-resolution deep future maps in a continuous-domain. DeepSRDCF extracts pre-trained CNN features for superior performance. ECO [18] tracker improves both speed and performance on the basis of C-COT by way of alleviating over-fitting and enhancing compactness of the generative model. CFWCR [30] upgrades ECO by normalizing every independent feature extracted from different CNN layers and getting the weighted sum of convolution responses for all layers to generate the final confidence score. To deal with the boundary effects, SRDCF [31] presents a spatial regularization component to penalize filters coefficients on boundary and solves the formulations by the Gauss-Seidel algorithm. BACF [32] trains the tracker from real negative training examples instead of negative examples in prior CFs which are limited to circularly shifted samples. This work learns the filter on multi-channel features using ADMM. STRCF [33] introduces a temporal regularization into correlation filters for obtaining a more robust appearance model in comparison with SRDCF.

III. PROPOSED METHOD

A. BASELINE APPROACH

STRCF: Usually, the model updating happens in each frame that only utilizes information of the current frame, ignoring the information of the previous frames. The lack of previous memory information degrades the robustness of the trackers, leading to model drift over time. Some trackers take a unique strategy that updates the model every few frames [18]. This updating scheme results in advanced tracking performance but slows down the model convergence speed. To merge historical information, other DCF trackers update the model via interpolating current frame model with previous model parameters, using a learning rate to control the updating degree. Thus usage of historic information could mitigate the effect of model drift. STRCF [33] introduces a temporal regularization component to avoid model drift over time, leading to a more robust and discriminative model.

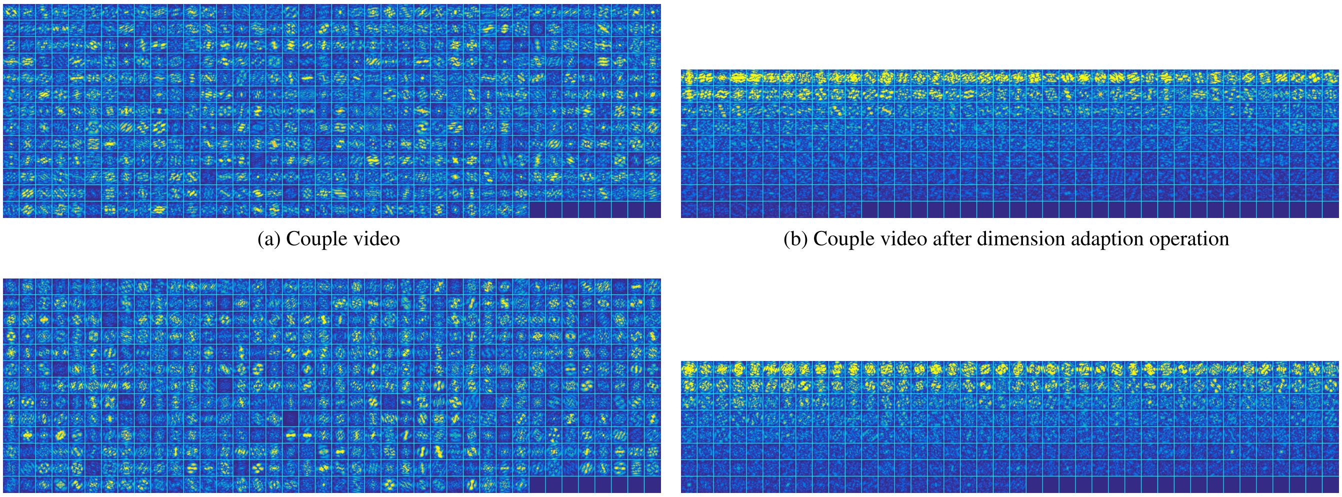
We first review the STRCF [33] model. Each training sample $x_t = \{x_t^d\}_{d=1:D}$ contains D channels feature maps with size of $M \times N$ and y is the Gaussian response label. In STRCF model, a multi-channel correlation filter $f = (f^1 \cdots f^D)$ can be trained through the minimization of the following as,

$$\arg\min_f \frac{1}{2}\left\|\sum\nolimits_{d=1}^{D} x_t^d * f^d - y\right\|^2 + \frac{1}{2}\sum\nolimits_{d=1}^{D}\left\|w \cdot f^d\right\|^2 + \frac{\mu}{2}\|f - f_{t-1}\|^2, \quad (1)$$

where $\sum_{d=1}^{D}\left\|w \cdot f^d\right\|^2$ denotes the introduced spatial regularization term, $\|f - f_{t-1}\|^2$ is a temporal regularization term motivated by Passive-Aggressive (PA) [34], f_{t-1} denotes the correlation filters obtained in the $(t-1)$-th frame and μ denotes the regularization parameter. For predicting the detection scores of the target, filter f is trained as,

$$S_f(x) = x * f = \sum_{d=1}^{D} x^d * f^d. \quad (2)$$

ECO: The ECO [18] method introduces a factorized convolution approach aiming at the reduction of parameters number. The construction of the d-th channel filter f^d is indicated as a linear combination $f^d = \sum_{c=1}^{C} p_{d,c} f^c$ using



(a) Couple video

(b) Couple video after dimension adaption operation

(c) Skiing video

(d) Skiing video after dimension adaption operation

FIGURE 3. Visualization of the learned filters to *Res3d* layer extracted from SE-ResNet50. (a) and (c) are all the 512 filters learned from *Res3d* layer of two videos while (b) and (d) demonstrate the reduced filters of 512 filters obtained by dimension adaption operation. Most of the baseline filters contain negligible energy, indicating irrelevant information in the corresponding feature layers. Our dimension adaption operation learns filters with significant energy.

a basis filters set $f^1 \cdots f^C$, where $C < D$. A $D \times C$ matrix $P = \left(p_{d,c}\right)$ is introduced to represent the learned coefficients $p_{d,c}$ compactly. The factorized convolution operator can be shown as,

$$S_f\left(x\right) = x * Pf = \sum_{c,d} x^d * p_{d,c} f^c = P^\top x * f. \tag{3}$$

The dimension of feature map x is reduced from D to C as it is multiplied with the matrix $P^\top$, which seems like reducing the linear dimensionality. Usually, the size (dimension) of matrix $P^\top$ is identical for different objects that cannot fully adapt to tracking scenes variation. Thus, it is reasonable to introduce an adaptive factorized convolution into the CF model.

B. DIMENSION ADAPTION OPERATION

Motivated by above discussion, we introduce a dimension adaption operation to reduce the number of parameters adapting to different tracking videos. Many filters learned from high-dimensional deep features contain negligible energy, as shown in Fig. 3.

We employ PCA at the beginning of each frame for purpose of gaining D principal components $P^1 \cdots P^D$ sorted according to their corresponding eigenvalues in descending order. For each track sequence, we select top C principal components which contain $n_{th} = 99\%$ of the total information to form the matrix $P^\top = \left(P^1 \cdots P^C\right)$. The dimension number C can be obtained by (4):

$$\frac{\sum_{i=1}^{i=C} \lambda_i}{\sum_{i=1}^{i=D} \lambda_i} \geq n_{th}, \tag{4}$$

where λ_i means the eigenvalue corresponding to the eigenvector P^i. The number C varies adaptively over different tracking scenes and is constant during the tracking process.

We use a dimension adaption operation in our track framework to reduce feature parameters number. The dimension numbers of three layers *Conv1x*, *Res3d*, and *Res4f* extracted from SE-ResNet50 are 64, 512, and 1024, respectively. The reduced dimensions of *Res3d* and *Res4f* layers in each tracking sequence of OTB2015 are visualized in Fig. 4. Apparently, dimension reducing adapts well to different videos.

In addition, Fig. 5 demonstrates a per-video comparison, comparing our method with STRCF in terms of overlap score. Obviously, our method performs more favorably than STRCF in major videos.

Our Objective Function: According to the above discussion, we propose to learn DACF filters by minimizing the following objective,

$$\arg\min_f \frac{1}{2}\left\|P^\top x * f - y\right\|^2 + \frac{1}{2}\sum\nolimits_{c=1}^{C}\left\|w \cdot f^c\right\|^2 + \frac{\mu}{2}\left\|f - f_{t-1}\right\|^2, \tag{5}$$

where the sample model x is replaced with the C-dimensional projected feature map $P^\top x$.

C. OPTIMIZATION ALGORITHM

To solve (5), We convert it into the equality constrained optimization form by introducing an auxiliary variable g:

$$\begin{aligned} &\arg\min_f \frac{1}{2}\left\|P^\top x * f - y\right\|^2 + \frac{1}{2}\sum\nolimits_{c=1}^{C}\left\|w \cdot g^c\right\|^2 \\ &\quad + \frac{\mu}{2}\left\|f - f_{t-1}\right\|^2 \\ &s.t.\;\; f = g. \end{aligned} \tag{6}$$

Equation (6) can be solved alternately using the ADMM technique. The Augmented Lagrangian form of (6) can be

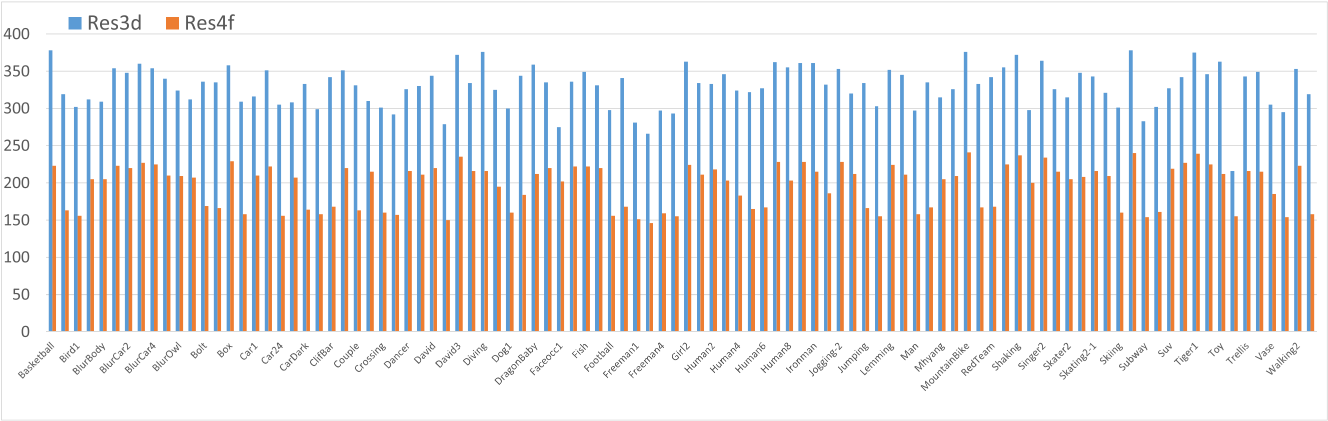


FIGURE 4. Dimension reducing results of *Res3d* and *Res4f* layers on OTB2015 dataset for 100 videos.

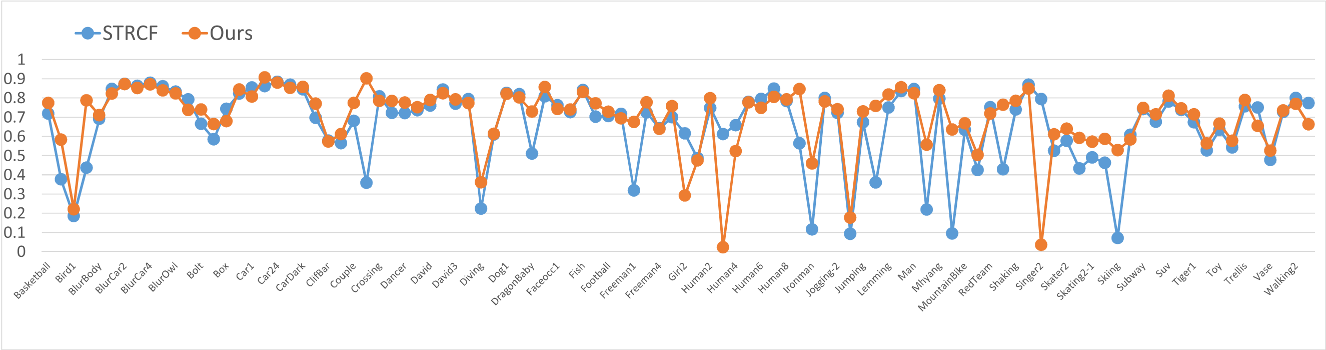


FIGURE 5. Overlap score results on OTB2015 dataset for 100 videos, comparing our proposed method and STRCF.

formulated as:

$$\begin{aligned}\mathcal{L}(f,g,h) = {} & \frac{1}{2}\left\|P^{\top}x * f - y\right\|^{2} + \frac{1}{2}\sum\nolimits_{c=1}^{C}\left\|w\cdot g^{c}\right\|^{2} \\ & + \sum\nolimits_{c=1}^{C}\left(f^{c}-g^{c}\right)^{\top}h^{c} + \frac{\gamma}{2}\sum\nolimits_{c=1}^{C}\left\|f^{c}-g^{c}\right\|^{2} \\ & + \frac{\mu}{2}\left\|f-f_{t-1}\right\|^{2}, \end{aligned} \tag{7}$$

where h and γ are the Lagrange multiplier and the penalty factor, respectively. Equation (7) can be reformulated as:

$$\begin{aligned}\mathcal{L}(f,g,h) = {} & \frac{1}{2}\left\|P^{\top}x * f - y\right\|^{2} + \frac{1}{2}\sum_{c=1}^{C}\left\|w\cdot g^{c}\right\|^{2} \\ & + \frac{\gamma}{2}\sum_{c=1}^{C}\left\|f^{c}-g^{c}+\frac{1}{\gamma}h^{c}\right\|^{2} + \frac{\mu}{2}\left\|f-f_{t-1}\right\|^{2}. \end{aligned} \tag{8}$$

The closed-form solutions of (8) can be obtained via alternatingly solving the following subproblems:

$$\begin{cases} f^{(i+1)} = \underset{f}{\arg\min}\ \left\|P^{\top}x * f - y\right\|^{2} + \gamma\left\|f-g+\frac{1}{\gamma}h\right\|^{2} \\ \qquad\qquad +\mu\left\|f-f_{t-1}\right\|^{2} \\ g^{(i+1)} = \underset{g}{\arg\min}\ \sum\nolimits_{c=1}^{C}\left\|w\cdot g^{c}\right\|^{2} + \gamma\left\|f-g+\dfrac{1}{\gamma}h\right\|^{2} \\ h^{(i+1)} = h^{(i)} + \gamma\left(f^{(i+1)}-g^{(i+1)}\right). \end{cases} \tag{9}$$

The solution to each subproblem is detailed as follows: *Subproblem f :*

Using Parseval's theorem, the first objective function of (9) can be expressed in the frequency domain as:

$$\begin{aligned}\underset{f}{\arg\min}\ & \left\|\sum\nolimits_{c=1}^{C}\widehat{x}_{t}^{c}\cdot\widehat{f}^{c}-\widehat{y}\right\|^{2} + \gamma\left\|\widehat{f}-\widehat{g}+\frac{1}{\gamma}\widehat{h}\right\| \\ & + \mu\left\|\widehat{f}-\widehat{f}_{t-1}\right\|^{2}, \end{aligned} \tag{10}$$

where the symbol ˆ means the discrete Fourier transform (DFT) of a signal. Considering processing on all channels of each pixel, we decompose (10) into MN subproblems, each of which is defined as:

$$\begin{aligned} & \left\|\mathcal{V}_{j}\left(\hat{x}_{t}\right)^{\top}\mathcal{V}_{j}(\hat{f})-\hat{y}_{j}\right\|^{2} \\ \underset{\mathcal{V}_{j}(\hat{f})}{\arg\min} & +\gamma\left\|\mathcal{V}_{j}(\hat{f})-\mathcal{V}_{j}(\hat{g})+\tfrac{1}{\gamma}\mathcal{V}_{j}(\hat{h})\right\|^{2} \\ & +\mu\left\|\mathcal{V}_{j}(\hat{f})-\mathcal{V}_{j}\left(\hat{f}_{t-1}\right)\right\|^{2}, \end{aligned} \tag{11}$$

where $\mathcal{V}_{j}(\hat{f}) \in \mathbb{R}^{C}$ denotes the vector consisting of all C channels of $\hat{f}$ on pixel j. The solution for $\mathcal{V}_{j}(\hat{f})$ is gotten by setting the derivative of (11) with respect to $\mathcal{V}_{j}(\hat{f})$ equal to zero:

$$\begin{aligned}\mathcal{V}_{j}(\hat{f}) = {} & \left((\mu+\gamma)I+\mathcal{V}_{j}\left(\hat{x}_{t}\right)\mathcal{V}_{j}\left(\hat{x}_{t}\right)^{\top}\right)^{-1} \\ & \cdot\left(\mathcal{V}_{j}\left(\hat{x}_{t}\right)\hat{y}_{j}+\gamma\mathcal{V}_{j}(\hat{g})-\mathcal{V}_{j}(\hat{h})+\mu\mathcal{V}_{j}\left(\hat{f}_{t-1}\right)\right). \end{aligned} \tag{12}$$

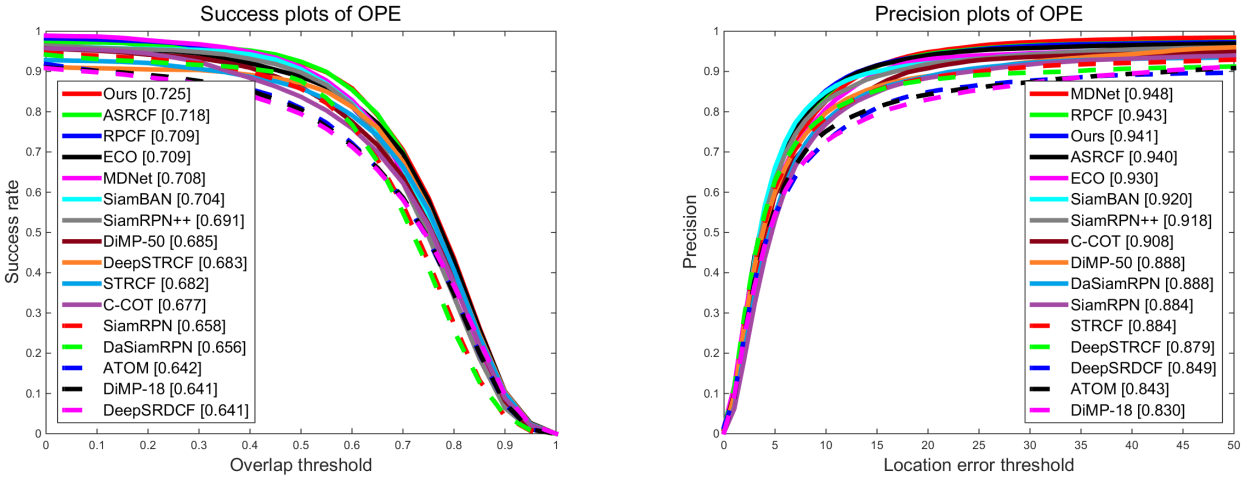


FIGURE 6. **Success and precision plots on OTB2013 dataset.**

Since $\mathcal{V}_j\left(\hat{x}_t\right)\mathcal{V}_j\left(\hat{x}_t\right)^\top$ is a rank-one matrix, we can calculate $\left((\mu+\gamma)I+\mathcal{V}_j\left(\hat{x}_t\right)\mathcal{V}_j\left(\hat{x}_t\right)^\top\right)^{-1}$ rapidly according to the Sherman-Morrison formula [35], stating that $\left(A+uv^\top\right)^{-1}=A^{-1}-\frac{A^{-1}uv^\top A^{-1}}{1+v^\top A^{-1}u}$, and we rewrite (12) as:

$$\mathcal{V}_j(\hat{f})=\frac{1}{\mu+\gamma}\left(I-\frac{\mathcal{V}_j\left(\hat{x}_t\right)\mathcal{V}_j\left(\hat{x}_t\right)^\top}{\mu+\gamma+\mathcal{V}_j\left(\hat{x}_t\right)^\top\mathcal{V}_j\left(\hat{x}_t\right)}\right)\cdot\left(\mathcal{V}_j\left(\hat{x}_t\right)\hat{y}_j+\gamma\mathcal{V}_j(\hat{g})-\mathcal{V}_j(\hat{h})+\mu\mathcal{V}_j\left(\hat{f}_{t-1}\right)\right). \quad (13)$$

We should solve MN subproblems separately over C channels. Thus, the computing complexity of solving $\hat{f}$ is $\mathcal{O}\left(CMN\right)$. Taking inverse DFT account, it costs $\mathcal{O}\left(CMN\log MN\right)$ to solve f.

Subproblem g:

According to the second row of (9), the closed-form solution of g can be obtained as:

$$g=\left(W^\top W+\gamma I\right)^{-1}(\gamma f+h), \quad (14)$$

where $W=diag\left(w\right)$ indicates the diagonal matrix with C diagonal matrices $diag\left(w\right)$ concatenated as elements on the diagonal.

Lagrangian Multiplier Update:

Lagrangian multiplier is updated using present solutions $f^{(i+1)}$ and $g^{(i+1)}$ as:

$$h^{(i+1)}=h^{(i)}+\gamma\left(f^{(i+1)}-g^{(i+1)}\right), \quad (15)$$

where $f^{(i+1)}$ and $g^{(i+1)}$ are the current solutions to the above two subproblems at iteration $(i+1)$. We select penalty factor γ following a scheme as:

$$\gamma^{(i+1)}=\min\left(\gamma_{\max},\rho\gamma^{(i)}\right), \quad (16)$$

where $\gamma_{\max}$ and ρ denote the maximum value of γ and the increment factor, respectively.

Our track model is convex, and it satisfies the Eckstein-Bertsekas condition [36]. Thus, it can converge to global optimum and has closed-form solution.

IV. EXPERIMENT

A. EXPERIMENTAL SETUP

We implement our tracker on MATLAB2014b using MatConvNet and AutoNN toolboxes. We run all the experiments on a PC machine equipped with an Intel i5 4570 CPU, 16GB RAM, and a single NVIDIA GTX 1080 GPU. We apply *Conv1x*, *Res3d*, and *Res4f* three layers of SE-ResNet50 to extract features for training and prediction. The size of searching area is set between 200 × 200 to 250 × 250. For the ADMM optimization, we choose the regularization parameter as $\mu=15$ and the number of iterations as 2. The initial penalty factor is set to $\gamma^{(0)}=0.1$ and updated by $\gamma^{(i+1)}=\min\left(\gamma_{\max},\rho\gamma^{(i)}\right)$, where $\gamma_{\max}=1$ and $\rho=10$.

B. QUANTITATIVE EVALUATION

1) COMPARISONS ON OTB BENCHMARKS

a: EVALUATION ON OTB2013

The OTB2013 dataset [19] is one of the most popular tracking datasets, which includes 50 fully annotated image sequences with various challenging attributes, such as occlusion (OCC), scale variation (SV), and deformation (DEF). We employ the One Pass Evaluation (OPE) to evaluate different trackers. In this benchmark, the quantitative analysis is based on two evaluation metrics: precision rate plot and success rate plot. The precision plot shows the percentage of frames whose distance between the estimated location with the ground truth is within the given range of threshold distance. The overlap score is defined as the intersection over union (IoU) ratio of predicted and ground truth bounding boxes. The success plot shows the ratios of frames whose overlap score is larger than the thresholds varied from 0 to 1. The Area Under the Curve (AUC) of success plots is used to rank the trackers.

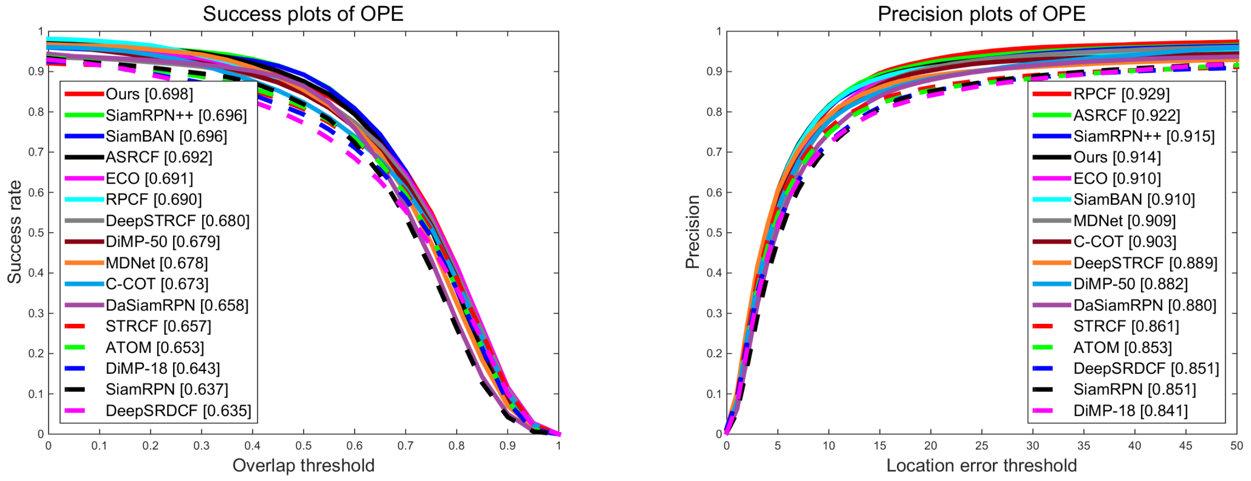


FIGURE 7. Success and precision plots on OTB2015 dataset.

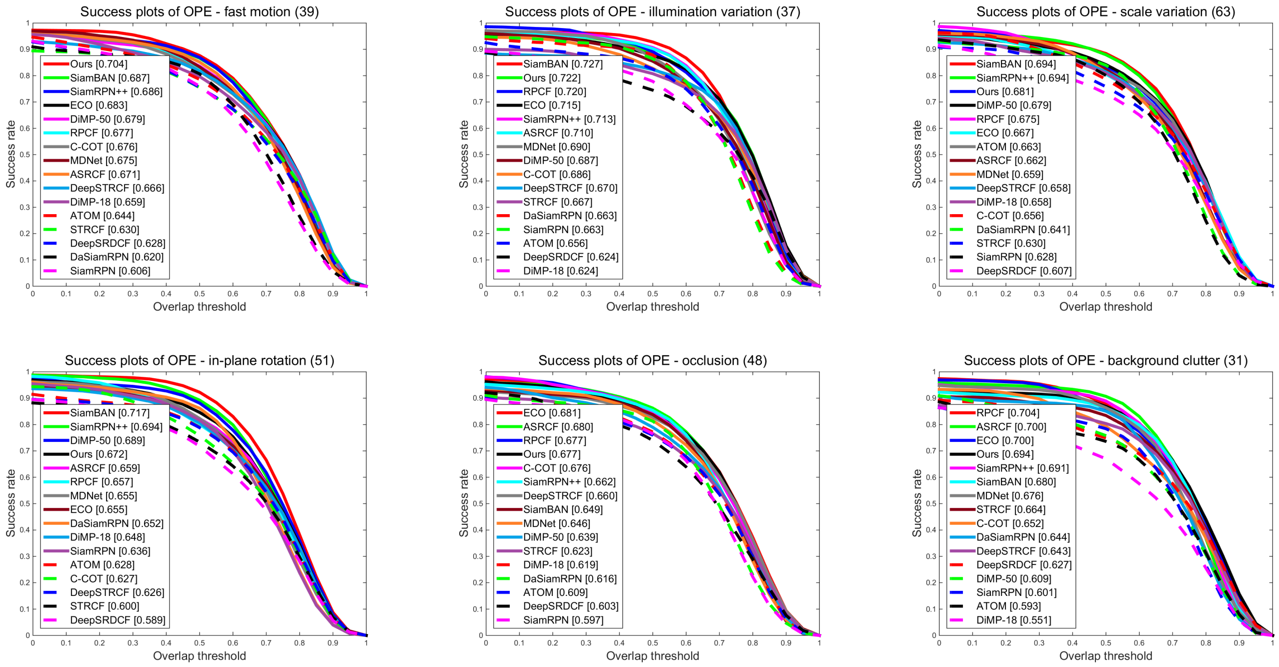


FIGURE 8. Success plots of all competing trackers with 6 attributes on OTB2015 dataset.

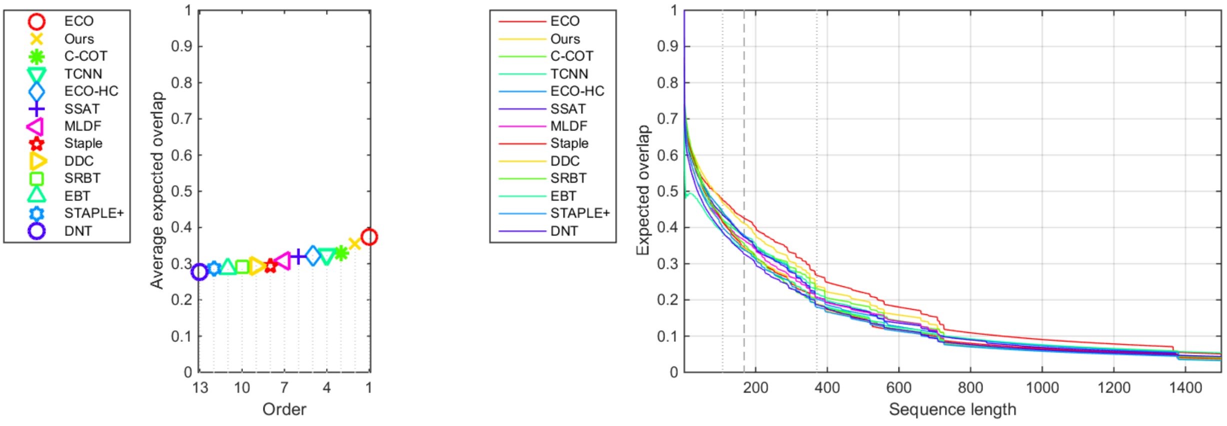
We compare our DACF with 15 state-of-the-art trackers including ECO [18], DiMP-18 [42], DiMP-50 [42], ATOM [43], ASRCF [37], STRCF [33], DeepSTRCF [33], DeepSRDCF [31], C-COT [29], SiamRPN [22], DaSiamRPN [23], SiamRPN++ [24], SiamBAN [44], MDNet [21], and RPCF [39]. The success plots and precision plots of comparison are shown in Fig. 6. Overall, our proposed DACF achieves almost the best performance with a 0.725 AUC score and a 0.941 distance precision rate.

b: EVALUATION ON OTB2015

The OTB2015 benchmark [20] is an extension of OTB2013, which includes 50 added video sequences. We evaluate the performance of the proposed DACF on OTB2015 benchmark and show the results of comparison with top-performing trackers, i.e., ECO, DiMP-18, DiMP-50, ATOM, ASRCF, STRCF, DeepSTRCF, DeepSRDCF, C-COT, SiamRPN, DaSiamRPN, SiamRPN++, SiamBAN, MDNet, and RPCF (see Fig. 7).

Our DACF ranks the best performance among these out-standing trackers with an AUC score of 0.698 and a distance precision rate of 0.914. DACF outperforms STRCF by a gain of 6.24% and 6.16% respectively in terms of success and precision.

Fig. 8 illustrates performance evaluation with 6 attributes on OTB2015. It is apparent that our DACF achieves almost



(a)

(b)

FIGURE 9. Experimental report plots on VOT2016. (a) Average expected overlap graph with trackers ranked from right to left. (b) Expected overlap curve.

TABLE 1. Evaluation results on VOT2016 in terms of Expected Average Overlap (EAO), Accuracy (A), and Robustness (R). The best three are indicated by red, blue, and green, respectively.

Tracker	EAO	A	R
ECO	**0.374**	0.553	**0.200**
Ours	0.354	**0.589**	0.233
C-COT	0.331	0.539	0.238
TCNN	0.325	0.554	0.268
ECO-HC	0.322	0.542	0.303
SSAT	0.321	0.577	0.291
MLDF	0.311	0.490	0.233
Staple	0.295	0.544	0.378
DDC	0.293	0.541	0.345
EBT	0.291	0.465	0.252
SRBT	0.290	0.496	0.350
STAPLE+	0.286	0.557	0.368
DNT	0.278	0.515	0.329

the best results on these attributes. In the case of Illumination Variation (IV) and Scale Variation (SV), trackers learning CFs from numerous features layers (e.g., C-COT) suffer from over-fitting due to large parameters. Our DACF adapts well to such variations profiting from dimension adaption operation and obtains 8.2% and 8.1% gains respectively than its baseline STRCF. Furthermore, DACF is robust to In-plane Rotation owed to the use of multi-level deep features and is superior to STRCF by 12%.

2) COMPARISONS ON VOT2016

The Visual Object Tracking (VOT) [40] is a popular performance evaluation toolkit in the research area of visual tracking. The VOT kit resets the tracker once it fails to track the target, which is different from OTB benchmark only initialing the tracker at the first frame. We perform comparisons on VOT2016 dataset that consists of 60 short-term video sequences.

In Table 1, we report the tracking results of our method compared with 12 representative competitors, including ECO, C-COT, TCNN [45], ECO-HC, SSAT [40], MLDF [40], Staple [38], DDC [40], SRBT [40], EBT [46], STAPLE+ [40], and DNT [47]. Our tracker achieves expected average overlap (EAO) score of 0.354 and robustness (R) score of 0.233, ranking second and just behind ECO both on these two metrics. Our method achieves accuracy (A) score of 0.589 that outperforms all the other trackers.

According to Fig. 9(a), our tracker ranks second in terms of average expected overlap. It can be seen in Fig. 9(b) that our tracker achieves superior performance against other trackers in expected overlap curve.

3) COMPARISONS ON UAV123

Aerial tracking with unmanned aerial vehicles (UAVs) [41] has become more and more popular nowadays. UAV123 dataset is composed of 123 HD video sequences captured from a low-altitude aerial perspective. We evaluate the performance of our tracker on UAV123 compared with 8 state-of-the-art methods: SiamRPN++, ECO, ECO-HC, ATOM, SAMF [48], DSST [49], SiamBAN, and SRDCF. We display the tracking results in Table 2. It can be seen that our tracker obtains a success score of 0.545 and a precision score of 0.764. Our tracker achieves more appealing performance against DCF based trackers but it falls behind the networks-based end-to-end methods such as ATOM, SiamBAN, and SiamRPN++.

C. QUALITATIVE EVALUATION

For better visualizing the tracking performance of the proposed method, we display tracking results of our proposed tracker compared with 4 trackers, i.e., C-COT, DeepSDRCF, ECO, and Staple on 6 challenging video sequences on OTB2015 benchmark in Fig. 10.

Fig. 10 illustrates that our proposed tracker performs well on all challenging sequences against other trackers. In CarScale, most trackers do not adapt well to scale variation during tracking except our tracker and Staple tracker. In the sequence *Freeman4*, the man's face is occluded in frame #158 but our tracker could capture the target after occlusion during tracking. Most trackers fail to handle in-plane rota-

TABLE 2. Success and Precision tracking results of our tracker and compared trackers on UAV123 dataset.

Tracker	DSST	SAMF	SRDCF	ECO-HC	ECO	Ours	ATOM	SiamBAN	SiamRPN++
Success	0.356	0.396	0.464	0.506	0.525	0.546	0.609	0.631	**0.642**
Precision	0.586	0.592	0.676	0.725	0.741	0.763	0.809	0.833	**0.840**

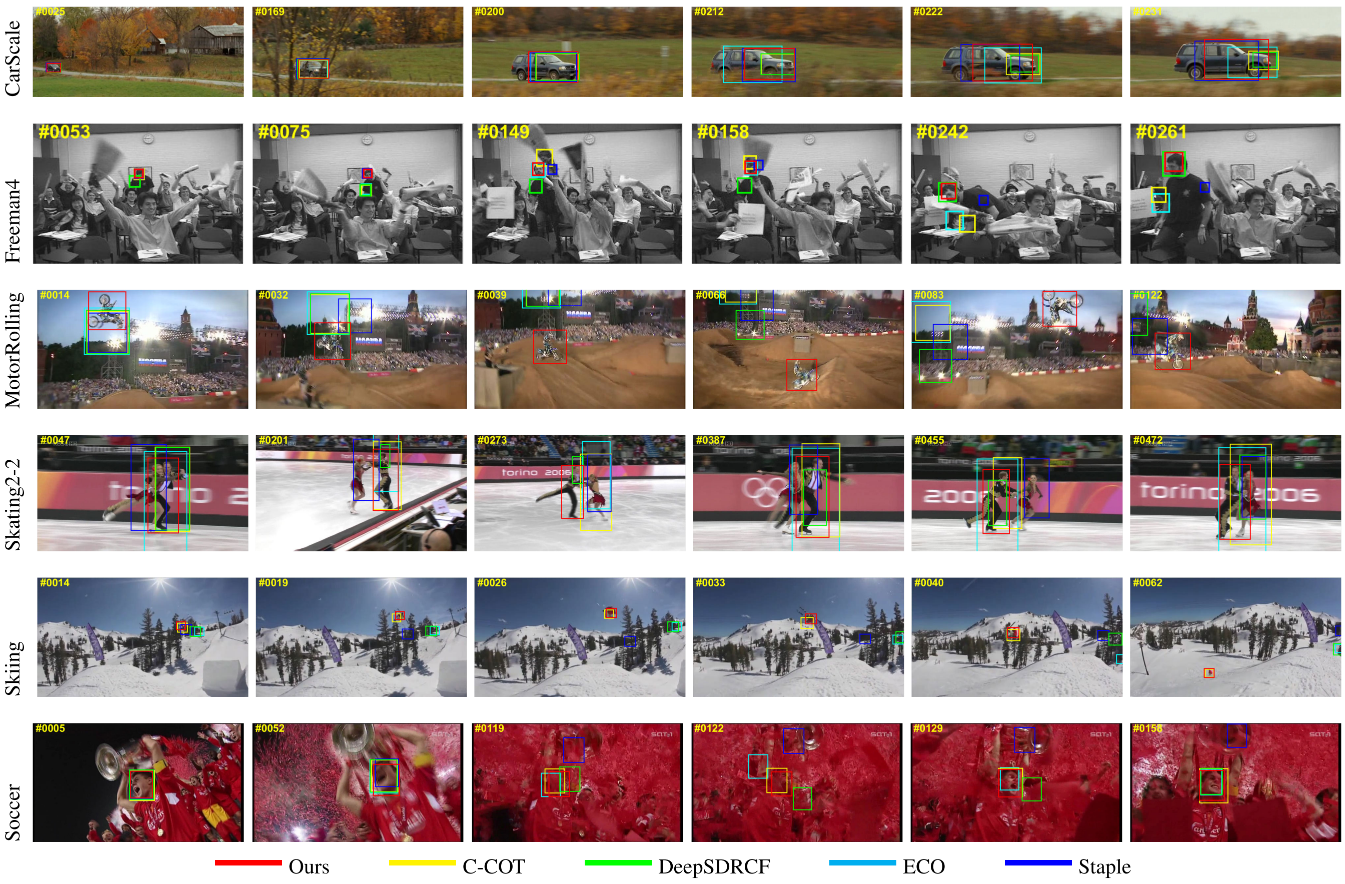


FIGURE 10. Qualitative tracking results of our proposed tracker compared with other 4 trackers, i.e., C-COT, DeepSDRCF, ECO, and Staple on 6 challenging sequences (Top to down: *Carscale*, *Freeman4*, *MotorRolling*, *Skating2-2*, *Skiing*, and *Soccer*) in OTB2015.

tion even since frame #14 in *MotorRolling*, but our tracker reliably tracks the target in subsequent frames. The target in *Skating2-2* undergoes out-of-plane rotation and deformation due to the male skater changing his movements. Only our tracker could deal with these challenges well while other trackers gradually drift away from the target. In *Skiing*, our method and C-COT can track the object with the attribution of low resolution. Our proposed method performs slightly better than C-COT in the process of tracking. In *Soccer*, it can be seen that most trackers fail to cope with heavy background clutter, but our tracker perform favorably against others.

D. ABLATION ANALYSES

Module Analyses: We investigate the effectiveness of proposed modules in our tracker and report the results in Fig. 11. The evaluation is performed on the OTB-2013 dataset in terms of success criterion. (1) We take STRCF tracker as the 'Baseline' that utilizes handcraft features only but does not make use of deep futures nor any feature reducing operation. (2) 'Baseline + VGG-M' is the DeepSTRCF tracker that uses VGG-M network on the basis of 'Baseline' without feature reducing operations. (3) 'Baseline + VGG-M + DAO' means joining our dimension adaption operation into 'Baseline +

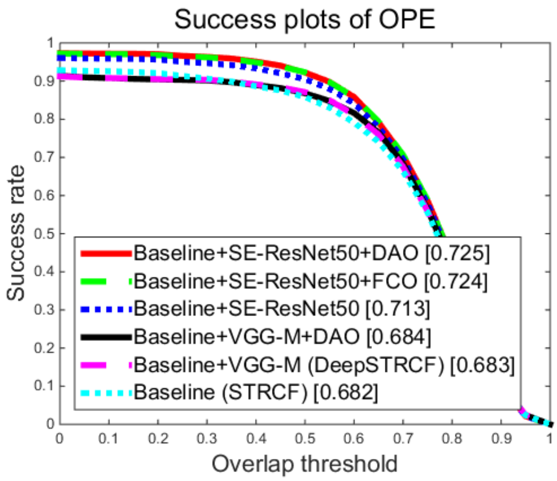


FIGURE 11. Ablation Analyses of different modules on OTB2013 dataset.

VGG-M'. (4) 'Baseline + SE-ResNet50' denotes replacing VGG-M of 'Baseline + VGG-M' with SE-ResNet50 as the backbone net. (5) 'Baseline + SE-ResNet50 + FCO' means introducing factorized convolution operation into the method. (6) 'Baseline + SE-ResNet50 + DAO' is our final tracker that joins dimension adaption operation. According to Fig. 11, 'Baseline + SE-ResNet50' is superior to 'Baseline + VGG-M' (DeepSTRCF) by 4.4%, benefiting from multi-channel deep CNN features extracted from SE-ResNet50 network. 'Baseline + SE-ResNet50 + DAO' has an improved performance than 'Baseline + SE-ResNet50' by 1.7% due to the proposed dimension adaption operation. Overall, our tracker surpasses the baseline method (STRCF) by 6.3%. Moreover, our method outperforms 'Baseline + VGG-M' (DeepSTRCF) using deep features by a gain of 6.1%.

V. CONCLUSION

In this work, we utilize multi-channel deep features in the STRCF [33] based framework for visual tracking to obtain efficient sample representation. Dimension adaption operation is introduced into the framework to reduce the number of parameters. Compared with previous works reducing parameters into fixed numbers, this dimension adaption operation could adapt well to varied tracking scenes and counter the issue of over-fitting.

REFERENCES

[1] X. Sheng, Y. Liu, H. Liang, F. Li, and Y. Man, "Robust visual tracking via an improved background aware correlation filter," *IEEE Access*, vol. 7, pp. 24877–24888, 2019, doi: 10.1109/ACCESS.2019.2900666.

[2] Y. Ming and Y. Zhang, "ADT: Object tracking algorithm based on adaptive detection," *IEEE Access*, vol. 8, pp. 56666–56679, 2020, doi: 10.1109/ACCESS.2020.2981525.

[3] C. Liu, J. Gong, J. Zhu, J. Zhang, and Y. Yan, "Correlation filter with motion detection for robust tracking of shape-deformed targets," *IEEE Access*, vol. 8, pp. 89161–89170, 2020, doi: 10.1109/ACCESS.2020.2993777.

[4] Y. Wang, X. Ban, H. Wang, X. Li, Z. Wang, D. Wu, Y. Yang, and S. Liu, "Particle filter vehicles tracking by fusing multiple features," *IEEE Access*, vol. 7, pp. 133694–133706, 2019, doi: 10.1109/ACCESS.2019.2941365.

[5] X. Gao, G. Xu, S. Li, Y. Wu, E. Dancigs, and J. Du, "Particle filter-based prediction for anomaly detection in automatic surveillance," *IEEE Access*, vol. 7, pp. 107550–107559, 2019, doi: 10.1109/ACCESS.2019.2931820.

[6] K. Chen, X. Song, X. Zhai, B. Zhang, B. Hou, and Y. Wang, "An integrated deep learning framework for occluded pedestrian tracking," *IEEE Access*, vol. 7, pp. 26060–26072, 2019.

[7] C. Neff, M. Mendieta, S. Mohan, M. Baharani, S. Rogers, and H. Tabkhi, "REVAMP2T: Real-time edge video analytics for multicamera privacy-aware pedestrian tracking," *IEEE Internet Things J.*, vol. 7, no. 4, pp. 2591–2602, Apr. 2020.

[8] S. Li, J. Chu, G. Zhong, L. Leng, and J. Miao, "Robust visual tracking with occlusion judgment and re-detection," *IEEE Access*, vol. 8, pp. 122772–122781, 2020, doi: 10.1109/ACCESS.2020.3007261.

[9] K. Chen and W. Tao, "Learning linear regression via single-convolutional layer for visual object tracking," *IEEE Trans. Multimedia*, vol. 21, no. 1, pp. 86–97, Jan. 2019, doi: 10.1109/tmm.2018.2846405.

[10] W. Ruan, J. Chen, Y. Wu, J. Wang, C. Liang, R. Hu, and J. Jiang, "Multi-correlation filters with triangle-structure constraints for object tracking," *IEEE Trans. Multimedia*, vol. 21, no. 5, pp. 1122–1134, May 2019, doi: 10.1109/TMM.2018.2872897.

[11] S. Zhang, W. Lu, W. Xing, and L. Zhang, "Learning scale-adaptive tight correlation filter for object tracking," *IEEE Trans. Cybern.*, vol. 50, no. 1, pp. 270–283, Jan. 2020, doi: 10.1109/TCYB.2018.2868782.

[12] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2005, pp. 886–893.

[13] J. van de Weijer, C. Schmid, J. Verbeek, and D. Larlus, "Learning color names for real-world applications," *IEEE Trans. Image Process.*, vol. 18, no. 7, pp. 1512–1523, Jul. 2009, doi: 10.1109/TIP.2009.2019809.

[14] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2015, *arXiv:1409.1556v6*. [Online]. Available: https://arxiv.org/abs/1409.1556v6

[15] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[16] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2261–2269.

[17] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 8, pp. 2011–2023, Aug. 2020, doi: 10.1109/TPAMI.2019.2913372.

[18] M. Danelljan, G. Bhat, F. S. Khan, and M. Felsberg, "ECO: Efficient convolution operators for tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6931–6939.

[19] Y. Wu, J. Lim, and M.-H. Yang, "Online object tracking: A benchmark," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 2411–2418.

[20] Y. Wu, J. Lim, and M.-H. Yang, "Object tracking benchmark," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1834–1848, Sep. 2015, doi: 10.1109/TPAMI.2014.2388226.

[21] H. Nam and B. Han, "Learning multi-domain convolutional neural networks for visual tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4293–4302.

[22] B. Li, J. Yan, W. Wu, Z. Zhu, and X. Hu, "High performance visual tracking with siamese region proposal network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 8971–8980.

[23] Z. Zhu, Q. Wang, B. Li, W. Wu, J. Yan, and W. Hu, "Distractor-aware siamese networks for visual object tracking," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Cham, Switzerland, 2018, pp. 103–119.

[24] B. Li, W. Wu, Q. Wang, F. Zhang, J. Xing, and J. Yan, "SiamRPN++: Evolution of siamese visual tracking with very deep networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 4277–4286.

[25] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Boston, MA, USA, Jun. 2015, pp. 1–9.

[26] D. Bolme, J. R. Beveridge, B. A. Draper, and Y. M. Lui, "Visual object tracking using adaptive correlation filters," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2010, pp. 2544–2550.

[27] J. F. Henriques, C. Rui, P. Martins, and J. Batista, "Exploiting the circulant structure of tracking-by-detection with kernels," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2012, pp. 702–715.

[28] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista, "High-speed tracking with kernelized correlation filters," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 3, pp. 583–596, Mar. 2015, doi: 10.1109/TPAMI.2014.2345390.

[29] M. Danelljan, A. Robinson, F. S. Khan, and M. Felsberg, "Beyond correlation filters: Learning continuous convolution operators for visual tracking," in *Proc. Eur. Comput. Vis. (ECCV)*, Cham, Switzerland, 2016, pp. 472–488.

[30] Z. He, Y. Fan, J. Zhuang, Y. Dong, and H. Bai, "Correlation filters with weighted convolution responses," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2017, pp. 1992–2000.

[31] M. Danelljan, G. Häger, F. S. Khan, and M. Felsberg, "Learning spatially regularized correlation filters for visual tracking," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 4310–4318.

[32] H. K. Galoogahi, A. Fagg, and S. Lucey, "Learning background-aware correlation filters for visual tracking," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1144–1152.

[33] F. Li, C. Tian, W. Zuo, L. Zhang, and M.-H. Yang, "Learning spatial-temporal regularized correlation filters for visual tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 4904–4913.

[34] K. Crammer, O. Dekel, J. Keshet, S. Shalev-Shwartz, and Y. Singer, "Online passive-aggressive algorithms," *J. Mach. Learn. Res.*, vol. 7, no. 3, pp. 551–585, Dec. 2006.

[35] K. B. Petersen and M. S. Pedersen, "Inverses," in *The Matrix Cookbook*. Lyngby, Denmark: Technical Univ. Denmark, 2012, ch. 3, sec. 2, p. 8.

[36] J. Eckstein and D. P. Bertsekas, "On the Douglas–Rachford splitting method and the proximal point algorithm for maximal monotone operators," *Math. Program.*, vol. 55, nos. 1–3, pp. 293–318, Apr. 1992.

[37] K. Dai, D. Wang, H. Lu, C. Sun, and J. Li, "Visual tracking via adaptive spatially-regularized correlation filters," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 4665–4674.

[38] L. Bertinetto, J. Valmadre, S. Golodetz, O. Miksik, and P. H. S. Torr, "Staple: Complementary learners for real-time tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 1401–1409.

[39] Y. Sun, C. Sun, D. Wang, Y. He, and H. Lu, "ROI pooled correlation filters for visual tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5783–5791.

[40] M. Kristan, R. Bowden, and K. Lebeda, "The visual object tracking VOT2016 challenge results," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Oct. 2016, pp. 777–823.

[41] M. Mueller, N. Smith, and B. Ghanem, "A benchmark and simulator for UAV tracking," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Cham, Switzerland: Springer, Oct. 2016, pp. 445–461.

[42] G. Bhat, M. Danelljan, L. Van Gool, and R. Timofte, "Learning discriminative model prediction for tracking," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Seoul, South Korea, Oct. 2019, pp. 6181–6190.

[43] M. Danelljan, G. Bhat, F. S. Khan, and M. Felsberg, "ATOM: Accurate tracking by overlap maximization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 4655–4664.

[44] Z. Chen, B. Zhong, G. Li, S. Zhang, and R. Ji, "Siamese box adaptive network for visual tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, Jun. 2020, pp. 6667–6676.

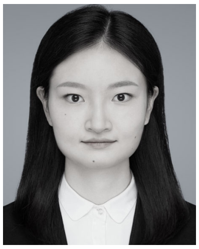
[45] H. Nam, M. Baek, and B. Han, "Modeling and propagating CNNs in a tree structure for visual tracking," Aug. 2016, *arXiv:1608.07242*. [Online]. Available: http://arxiv.org/abs/1608.07242

[46] G. Zhu, F. Porikli, and H. Li, "Beyond local search: Tracking objects everywhere with instance-specific proposals," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 943–951.

[47] Z. Chi, H. Li, H. Lu, and M.-H. Yang, "Dual deep network for visual tracking," *IEEE Trans. Image Process.*, vol. 26, no. 4, pp. 2005–2015, Apr. 2017, doi: 10.1109/TIP.2017.2669880.

[48] Y. Li and J. Zhu, "A scale adaptive kernel correlation filter tracker with feature integration," in *Proc. Eur. Conf. Comput. Vis. Workshops (ECCVW)*. Cham, Switzerland: Springer, 2014, pp. 254–265.

[49] M. Danelljan, G. Häger, F. S. Khan, and M. Felsberg, "Discriminative scale space tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 8, pp. 1561–1575, Aug. 2017, doi: 10.1109/TPAMI.2016.2609928.



LINGYUE WU received the B.E. degree from the School of Optics and Photonics, Beijing Institute of Technology, Beijing, China, in 2018. She is currently pursuing the M.S. degree with the Beijing Institute of Technology. Her research interests include computer vision and machine learning.

TINGFA XU received the Ph.D. degree from the Changchun Institute of Optics, Fine Mechanics and Physics, Changchun, China, in 2004. He is currently a Professor with the School of Optics and Photonics, Beijing Institute of Technology, Beijing, China. His research interests include optoelectronic imaging and detection, and hyper-spectral remote sensing image processing.

YUSHAN ZHANG received the B.E. degree from the School of Information and Electronics, Beijing Institute of Technology, Beijing, China, in 2018. She is currently pursuing the M.S. degree with the Beijing Institute of Technology. Her research interests include computer vision, machine learning, and deep learning.

FAN WU received the B.E. degree from the School of Optics and Photonics, Beijing Institute of Technology, Beijing, China, in 2018. She is currently pursuing the M.S. degree with the Beijing Institute of Technology. Her research interests include computer vision and machine learning.

CHANG XU is currently pursuing the Ph.D. degree with the School of Optics and Photonics, Beijing Institute of Technology, Beijing, China. Her research interests include computational imaging, hyperspectral data acquisition/processing, and computer vision.

XIANGMIN LI received the Ph.D. degree from the Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an, China, in 1995. He is currently a Professor with the School of Optics and Photonics, Beijing Institute of Technology, Beijing, China. His research interests include object tracking/detection and image/video processing.

JIHUI WANG received the Ph.D. degree in optical engineering from the Beijing Institute of Technology, Beijing, China, in 2008. From 2011 to 2012, she worked as a Visiting Fellow with The Pennsylvania State University. She is currently an Assistant Professor with the School of Optics and Photonics, Beijing Institute of Technology. Her research interests include optoelectronic imaging theory and digital image processing.