IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received November 27, 2019, accepted December 12, 2019, date of publication December 23, 2019,
date of current version January 3, 2020.

Digital Object Identifier 10.1109/ACCESS.2019.2961778

Double-Channel Object Tracking With Position Deviation Suppression

JUN CHU[1,2], XUJI TU[1,3], LU LENG[1,3], (Member, IEEE), AND JUN MIAO[2,4]

[1]Key Laboratory of Jiangxi Province for Image Processing and Pattern Recognition, Nanchang Hangkong University, Nanchang 330063, China
[2]Key Laboratory of Lunar and Deep Space Exploration, National Astronomical Observatories, Chinese Academy of Sciences, Beijing 100101, China
[3]School of Software, Nanchang Hangkong University, Nanchang 330063, China
[4]School of Aeronautical Manufacturing Engineering, Nanchang Hangkong University, Nanchang 330063, China

Corresponding author: Lu Leng (leng@nchu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61663031, Grant 61866028, Grant 61661036, Grant 61763033, Grant 61662049, and Grant 61866025, in part by the Foundation of China Scholarship Council under Grant CSC201908360075, in part by the Key Program Project of Research and Development (Jiangxi Provincial Department of Science and Technology) under Grant 20171ACE50024 and Grant 20192BBE50073, in part by the Construction Project of Advantageous Science and Technology Innovation Team in Jiangxi Province under Grant 20165BCB19007, and in part by the Open Foundation of Key Laboratory of Jiangxi Province for Image Processing and Pattern Recognition under Grant ET201680245 and Grant TX201604002.

ABSTRACT The object tracking methods based on multi-domain convolutional neural network (MDNet) commonly fail to track in the case of background clutter. A novel double-channel object tracking (DCOT) is proposed to solve this problem. The discriminative correlation filter (DCF), which has strong discriminative power of low-level features, is employed for the position deviation suppress of the samples generated from MDNet. Firstly the pre-trained deep network is used to learn and classify the target and background in the video frames. If the tracked position of the DCF is judged to be correct, we delete the target candidate samples with high position deviation from MDNet. The position deviation is measured by the distance between the tracked positions of the DCF and MDNet. Finally, MDNet and DCF are updated with a robust update strategy. The experiments are performed on OTB-100 and VOT-2016. The overlap precision and distance precision of DCOT on OTB-100 are 92.2% and 69.5%, respectively, which are higher than those of MDNet by 1.3% and 1.7%. The results of DCOT in background clutter are higher than those of SANet by 0.2% and 2.8%, respectively. DCOT is also superior to other state-of-the-art trackers on VOT-2016.

INDEX TERMS Double-channel object tracking, position deviation suppression, DCF, MDNet.

I. INTRODUCTION

Object tracking is to specify the size and position of the target in the first frame of the video sequence, and then find the size and position of the same target in the subsequent frames [1], [2]. It is widely used in various fields such as autonomous driving, robot navigation, missile guidance, intelligent transportation, etc. [3]. Among them, discriminative correlation filter (DCF) and deep learning become two mainstream technologies for object tracking [4], [5]. The DCF tracking algorithms using traditional features have fast speed and high localization accuracy in simple background. The localization accuracy of the DCF algorithms using deep features is improved, but they still suffer from the problems of target occlusion, deformation, out-of-plane rotation, and so on. The deep-learning-based object tracking algorithms are typically pre-trained on large data sets (such as ImageNet [6]), so the appearance and semantic information are well integrated to improve the performance in the scenarios of target occlusion and deformation. Different sequences involve various labels of class, movements, and object shapes; therefore, tracking algorithms encounter several challenges in different sequences, such as occlusion, deformation, illumination variance, motion blur, and so on.

A new convolution neural network (CNN) architecture is used in multi-domain convolutional neural network (MDNet) [7], in which multiple labeled videos are used for the training of the shared representations among different targets. Different videos represent different domains. Firstly, CNN network is offline trained. Then the fully connected layer is online fine-tuned to train and update the network. Under the existing object classification knowledge, the tracker learns

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

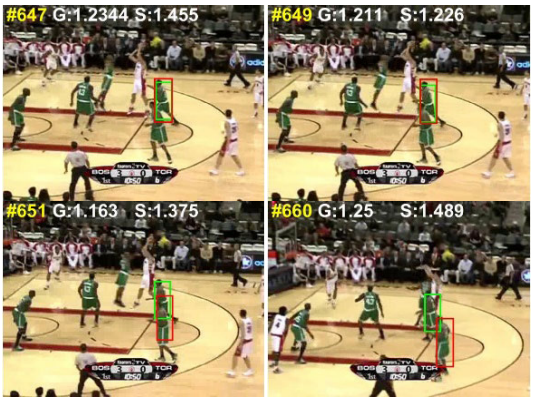

FIGURE 1. The object tracking performance of MDNet in sequence Basketball. The red and green boxes are the prediction results of MDNet and ground truth, respectively. G and S represent the confidences of the ground truth and the prediction results, respectively.

the targets with various poses in the current tracking scene, and adds long-term and short-term learning mechanisms to solve occlusion problem. However, the CNN network is not able to well classify similar objects that are close. As shown in Figure 1, a tracking error occurs when the two basketball players in the similar color clothes are close to each other.

A novel double-channel object tracking (DCOT) is proposed to solve the aforementioned problems. The one channel is MDNet. The other channel is discriminative correlation filter (DCF), which has strong discriminative power of low-level features and is employed for the position deviation suppress of the samples generated from MDNet. Our contributions include:

(1) The proposed two-channel target tracking algorithm combines MDNet and spatial regularization correlation filter (SRDCF). MDNet has the strong processing power for target deformation, while SRDCF has the sensitivity to complex background and low-resolution targets.

(2) The low-level feature of correlation filter (DCF) with strong discrimination ability is used to suppress the positional deviation of the samples in MDNet.

(3) A novel robust update strategy is established. MDNet uses the short-term and long-term mechanisms to update, while SRDCF is updated under the supervision of MDNet. This way makes the two methods complement each other, and improves the overall robustness of the algorithm.

OTB-100 [8] and VOT2016 [9] are used to evaluate DCOT, which is compared with several state-of-the-art deep-learning-based trackers and classic trackers.

II. RELATED WORKS

We just introduce the tracking algorithms based on correlation filter and deep learning since they are two mainstream technologies for object tracking.

A. TRACKING ALGORITHM BASED ON CORRELATION FILTER

The object tracking algorithms based on DCF remarkably accelerate the training and detection process by using fast Fourier transform, and they show satisfactory tracking performance and speed [10]–[15]. The early correlation filter tracking algorithms use the low-level features [12]–[14] or their fusion [16]. To solve the scale problem, the discriminative scale space tracking (DSST) algorithm [13] adds scale filter to the position filter. The object centroid is first localized, and then its size is matched. Guo *et al.* [15] used compressive random projection to get Haar-like features for fast and reliable tracking. For different features in different tracking scenarios, Staple [16] adaptively fuses histogram of oriented gradient (HOG) features and color features to improve the performance in different tracking scenarios. These algorithms easily fail in the cases of target deformation and occlusion. Thanks to the circulant matrix [17], the number of samples is enlarged to enhance the robustness. Unfortunately, because the target search area is limited, the filter has limited ability to learn the information of the background surrounding the target. Accordingly few negative samples are trained, which tends to cause over-fitting and boundary effect. Thus the filter is likely to fail owing to target deformation, fast motion, and occlusion. If the search area is blindly expanded, it is probable that too much background information suppresses the discriminative power of the filter. In order to alleviate the boundary effect, SRDCF [18] expands the filter search range, and adds a regular penalty term to the loss function to suppress the interference of background information. Thus the predicted position of the state-of-the-art SRDCF model can be considered as the basis for the measurement of position deviation. SRDCF tracker was improved by some new spatial regularization methods, such as Feng *et al.*'s dynamic saliency-aware regularization [19], Han *et al.*'s content-related spatial regularization [20], and Zhang *et al.*'s object-adaptive spatial regularization [21].

Currently the combination between correlation filter and deep features has become a new research direction [22]–[24]. Pre-trained deep features for target description can have better discriminative power. However, most correlation filter algorithms consider the position with the highest response value as the target centroid, so they do not yet ideally overcome target occlusion, out-of-plane rotation, and so on, and accordingly the powerful representation ability of CNN cannot be fully utilized.

B. TRACKING ALGORITHM BASED ON DEEP NEURAL NETWORK

The object tracking algorithms of the deep network architecture directly reduce or modify the pre-trained network model to better use the low-level features and high-level semantic information [7], [25]–[27]. In the cases of target occlusion and deformation, the trackers are superior to most correlation filter trackers and have excellent performance in the VOT [7], [23], [28]. Guo *et al.* [26] proposed dynamic Siamese network effectively to learn the temporal variation of target appearance and present element-wise multi-layer fusion. The VGG [29] network-based tracking algorithm represented by MDNet uses the small VGG model to learn the appearance

of the target, that is, the pre-trained convolution layer is used to extract the features and fine-tune the fully connected layer. The network outputs sample confidence, and the sample with the highest confidence is considered as the target. Modeling and propagating CNNs in a tree structure for visual tracking (TCNN) [27] constructs a tree view in which each small VGG network model is a node of the tree, that is, each node predicts the target in the current frame. A robust calculation strategy of node weight is used to add and delete the node for final predicted target position. Convolutional residual learning for visual tracking (CREST) [30] reformulates DCF as a one-layer CNN. It integrates feature extraction, response generation, and model update into the CNN for end-to-end training. To alleviate a rapid model degradation by large appearance changes, residual learning [31], [32] is applied to capture the target appearance changes. Song *et al*. [33] pointed out that the positive samples are highly spatially overlapping, and the number of positive samples is much smaller than that of the negative samples. Their method uses the generative adversarial network (GAN) [34] to amplify positive samples and learn various changes in targets over a long period of time. It also has strong robustness in complex scenes such as occlusion and rotation. The algorithm achieves excellent results on the benchmark data sets.

III. METHOD

A. MOTIVATION

MDNet is inspired by the classic target detection network Region-CNN (RCNN). The network consists of three convolution layers and three fully connected layers. The last full-connection layer is classification layer, which outputs confidence. In the main tracking process of the algorithm, the target samples are produced according to the target position of the previous frame. The samples are input to network and get the confidence degree. The sample with the highest confidence is considered as the target location. MDNet uses a multi-domain learning method to perform offline training with multiple video sequences to obtain common characteristics between different targets. MDNet performs well in the sequence of target occlusion, deformation and illumination variance. MDNet focuses on the feature maps trained for target and background categories to distinguish between foreground and background; however, its discriminative ability between similar targets is weak. If the target is occluded, its confidence is lower than that of the another similar target in the background, as shown in Figure 1.

SRDCF is improved from kernelized correlation filter (KCF) [20], which has KCF characteristics. It benefits from the sampling of the cyclic matrix, and has a large number of training samples. Different from KCF, it adopts HOG features and color features, which have less loss of original image information. It can cope well with the learning of the target appearance at low resolution, and is sensitive to variance of the target appearance. SRDCF introduces the regular terms to training function. The regular terms alleviate the boundary

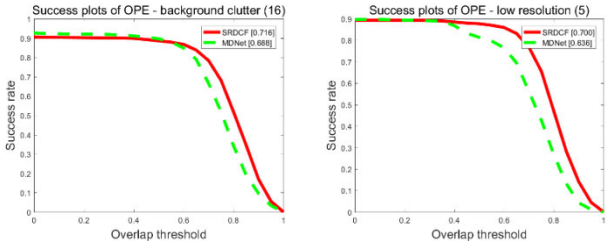

FIGURE 2. The success plots of the SRDCF and the MDNet under Group A object tracking video. Left one: success plot of background clutter. Right one: success plot of lower resolution.

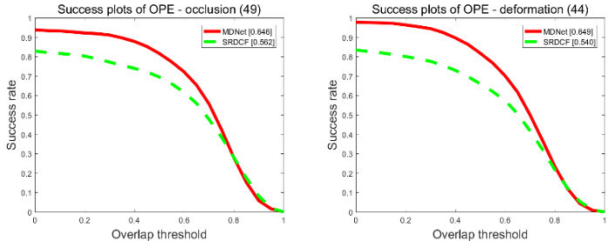

FIGURE 3. The success plots of the SRDCF and the MDNet under Group B object tracking video. Left one: success plot of occlusion. Right one: success plot of deformation.

effect caused by the cyclic matrix. In addition, SRDCF has a larger training area, so the filter learns more information about the target and background, making it easier to distinguish between objects and similar objects near the target.

In the OTB100 dataset, each video is labeled with multiple difficult tags, including target deformation, occlusion, background similarity interference, low resolution, illumination variance, etc. We only selected the sequences with four challenging tags, including target deformation, occlusion, background similarity interference, and low resolution. The selected videos were divided into 2 groups, Group A and Group B. Group A have background clutter and low-resolution. In Group B, there are long-term target occlusions and remarkable target deformations. The reason for this grouping is to show the performance of the MDNet and SRDCF methods under the conditions of Groups A and B, respectively. The tracking performances of the two algorithms, MDNet and SRDCF, are below. In the case of target similarity interference and low resolution (Group A), the overall performance of SRDCF is better than MDNet (Figure 2). MDNet has better tracking performance during long-term occlusion and severe deformation (Group B) (Figure 3).

The above experimental results show that the strengths of the two methods are complementary. We also analyzed the feasibility of putting HOG features and color features into the network in two modes. In the one mode, both HOG features and color features are connected to fully connected layer. It makes no sense, because neither HOG features nor color features are as robust as deep features. In the other mode, HOG features and color features are input into MDNet. The features are convoluted layer-by-layer. The number of feature channel increases from 3 to 34 (31 channels and 3 channels

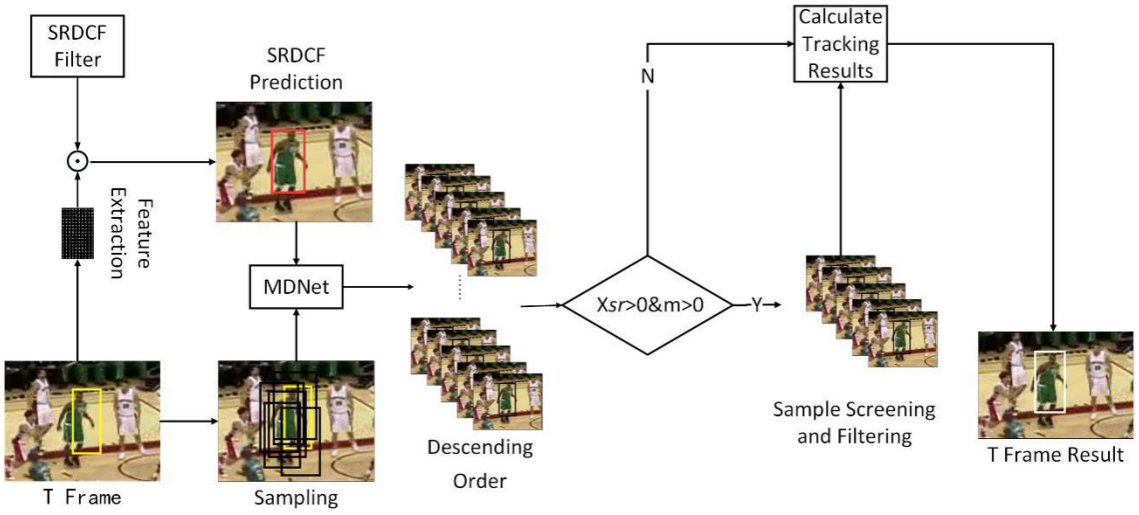

FIGURE 4. The graph of the framework.

for HOG features and color features, respectively), much more parameters are required. Thus the network updating is time-consuming. In order to solve the interference problem of the similar objects in MDNet algorithm without reducing its robustness in target occlusion and deformation, we propose a double-channel object tracking with position deviation suppression. According to the SRDCF result and the sample deviation degree, the target candidate samples of the MDNet are screened to eliminate the interference of the similar target samples.

B. FRAMEWORK

The framework of DCOT is shown in Figure 4, which is implemented as follows.

Firstly, we input the video frames into the filter to extract features, and calculate the feature response map. After that we use SRDCF to predict location of the target centroid as the basis of position deviation measurement, and add it to the target candidate samples of MDNet. Secondly, MDNet generates target candidate samples according to the target position and size of the previous frame, which are input into MDNet together with the SRDCF prediction position in the previous step. Thirdly, MDNet calculates the confidence of each sample, and arranges the target candidate samples in descending order of confidence. We select 3 target candidate samples that have higher confidences than the other candidate samples. m in Figure 4 denotes the average confidence of the 3 selected samples. Fourthly, according to X_{sr} (the sample confidence of the SRDCF prediction position) and m, the target candidate samples with higher sample deviation in the MDNet target candidate sample set are removed. The final output is the mean value of the corresponding samples from the remained candidate samples with high confidences.

C. CONFIDENCE THRESHOLD

Confidence is the positive sample probability of the target sample in MDNet, reflecting the likelihood that the current predicted sample is the target. MDNet uses confidence to judge the condition of the target. During the tracking process, when the target is occluded or deformed sharply, the target is polluted or the appearance changes greatly. And the target model learned from the previous period is quite different, so the positive sample probability of the target is reduced, that is, the target sample confidence becomes lower. Therefore, this paper also uses the target confidence to judge whether the target is occluded or deformed sharply.

We select the Jogging, Matrix, Lemming, and Soccer video sequences in the OTB-100 to compare the confidence of the target before and after occlusion and deformation for finding the optimal confidence threshold. As shown in Figure 5, in these sequences, when the target has occlusion (foreign object occlusion or self-occlusion) or severe deformation, the confidence level is reduced to 0 or negative. In other video sequences, the same rules as above are also found. Therefore, the confidence threshold is set to 0 as the basis for judging whether the target has occlusion or severe deformation.

D. SAMPLE DEVIATION CALCULATION

For MDNet, the target sample resolution for network training is generally low. After the frame is input into the deep learning network, the details of the target appearance are further lost, and the difference between the similar objects cannot be distinguished. The internal discriminative ability is weak, which causes MDNet to track poorly when the target semantic information and the background semantic information are not significantly differentiated (such as the interference of similar objects).

In order to solve this problem, we propose the position deviation suppress (PDS), that is, take the predict results of SRDCF as the basis position to filter out the samples with large deviation from the basis position. PDS is launched only if 2 implementation conditions are both satisfied. The 2 implementation conditions are $m > 0$ and $X_{sr} > 0$.

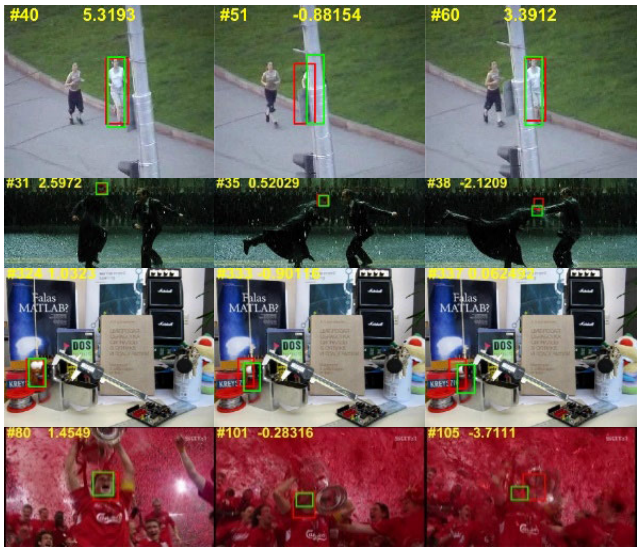

FIGURE 5. The performance of the algorithm in the sequence of Jogging, Matrix, Lemming, and Soccer (top to bottom): The red and green boxes are the prediction result of the algorithm and ground truth, and the upper yellow number is the confidence of the current target prediction value.

When PDS is implemented, the deviation of each target candidate sample in MDNet is computed. The samples with the deviation greater than the threshold are filtered out. Here the threshold is set to 5. Finally, in the remainder sample set, the mean of the 3 target candidate samples with higher confidences than the other samples is computed as the tracking result.

When PDS is not implemented, the mean of the 3 target candidate samples with higher confidences than the other samples in MDNet is directly computed as the tracking result.

DCF locates the target by predicting the position of the target center, and the accuracy evaluation of the tracking algorithm by the protocol for tracker evaluation uses the Euclidean distance [8].Therefore, the Euclidean distance is also used to describe the deviation of the sample. The degree of deviation ρ is the Euclidean distance between the SRDCF predicted position $P_{sr}(a_1, b_1, w_1, h_1)$ and each MDNet target candidate sample $P_{md}(a_2, b_2, w_2, h_2)$..

$$dist = \sqrt{(a_1 - a_2 + \frac{w_1}{2} - \frac{w_2}{2})^2 + (b_1 - b_2 + \frac{h_1}{2} - \frac{h_2}{2})^2} \quad (1)$$

The use of PDS can reduce the possibility of falsely selecting the objects in background, which are similar to the targets, as the predicted targets. As shown in Figure 6, after PDS, many mendacious analog samples in adjacent background are reduced. The deviation threshold is based on the empirical value of 5.

E. UPDATE

The update includes network model update and SRDCF filter update. The network adopts the long-term and short-term update mechanisms of the MDNet, that is, it establishes a long-term and short-term sample library. The long-term and short-term sample library keep the last 50 and 10 frames to track the correct training samples, respectively. MDNet is trained once in every 10 frames with negative samples from short-term and long-term sample library.

FIGURE 6. Comparison before and after the addition of the deviation screening strategy. The red and green boxes are the sample candidate box and ground truth, respectively. Before the screening strategy is added, the sample candidate box is distributed around the target and the mendacious similar targets. After the screening strategy is added, the sample candidate box is only distributed around the target.

When the confidence of the current target candidate sample is lower than 0, the network is trained with the short-term positive and negative samples. Conversely, the current tracking result is sampled into the long-term and short-term sample library.

SRDCF update is determined according to the confidence of the prediction result of SRDCF. When the confidence is lower than 0, the SRDCF prediction result means that the tracker may lose the target. At this time, the SRDCF filter is trained with the MDNet result to prevent the filter from being polluted. Conversely, SRDCF updates filter with its own predicted value.

IV. EXPERIMENTS AND DISCUSSIONS

A. EXPERIMENT SETUP

We compare DCOT with other state-of-the-art algorithms on OTB-100 and VOT-2016. The network parameter settings are the same as those of MDNet [7]. The threshold of the network update is set to 0, and the parameter settings of the correlation filter model are the same as those of SRDCF [18]. The threshold of the deviation is set to 5.

The computing platform contains Intel i5-3470 CPU, 8G memory, graphics card GTX1060. Matlab R2016b is the simulation software platform, and MatConvNet [25] tool library is used by deep learning library.

B. STATE-OF-THE-ART COMPARISON ON OTB-100

The OTB-100 data set contains 100 video test sequences in which the challenges include illumination variance, scale changes, occlusion, deformation, motion blur, fast motion, in-plane rotation, out-of-plane rotation, out-of-view, background clutters, low resolution. Two evaluation indices are used.

1) Success rate plots. The success rate refers to the percentage of the video whose overlap rate is greater than the specified threshold. The overlap rate threshold is 0.5, and the overlap rate is calculated according to the

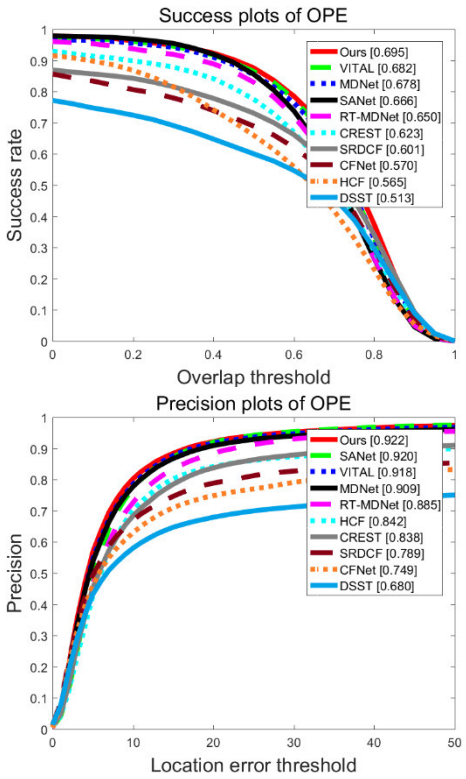

FIGURE 7. Comparison of the overall performance of this algorithm with other algorithms under OTB100. The top is the success rate plot and the bottom is the precision plot.

prediction box X_{pre} and the real box X_{gt} (as shown in Equation 2). The number in the success rate plots is the area under the success rate curve AUC. The larger the area is, the higher the tracking success rate of the algorithm is.

$$Iou = \frac{Aera(X_{pre}) \cap Area(X_{gt})}{Area(X_{pre}) \cup Area(X_{gt})} \tag{2}$$

2) Precision plots. The precision calculation is based on the Euclidean distance between the prediction box and the real box X_{gt}. The comparison method generally adopts the rate of the tracking frames with a precision less than or equal to 20 pix.

In order to verify the performance, DCOT is compared with 9 recent excellent tracking algorithms, in which 6 algorithms are based on deep learning, MDNet, CREST [30], VITAL [33], SANet [35], DCFNet [36], CFNet [37], HCF [38], RT-MDNet [39]. The overall performance is shown in Figure 7. DCOT has better success rate and precision under all sequences than other algorithms.

Compared with MDNet the success rate and precision are increased by 1.7% and 1.3%, respectively. Compared with SRDCF, the two indices are increased by 9.4% and 13.3%. In order to verify whether DCOT can alleviate the problem before the algorithm is improved, this paper selects the all sequences, and compares the performance level of the top ranked algorithms (including two unimproved algorithms).

As shown in Figure 8, the success rate and precision plots in both cases indicate that DCOT is superior to the compared algorithms. On the one hand, DCOT adopts the MDNet framework, which can learn the semantic information of the target. The long-term and short-term sample learning mechanisms are coupled, so the target is not lost in the case of target occlusion, even when the target is deformed. Because VITAL uses GAN to generate more samples to train for learning more variation of target appearance, when target is deformed, it does slightly better than our algorithm, the success rate is increased by 0.01%. On the other hand, based on SRDCF, the filtering strategy of candidate samples is developed. Even under the background clutters and low resolution, the difference between the target and the background appearance details is still obvious, which reduces the misjudgment of the interference sample collection and improves tracking overlap success rate.

DCOT does well for other challenging attributes of OTB dataset. But when a target moves fast, MDNet performs better than DCOT. Sometimes SRDCF fails to track when a target moves fast because the network confidence evaluation is not accurate or reliable for fast motion. It will introduce the negative influence to MDNet, so DCOT fails to track. When a target is out of view, VITAL is best, because VITAL has the robust model which benefits from GANs.

C. STATE-OF-THE-ART COMPARISON ON VOT-2016

VOT-2016, which contains 60 video test sequences, is divided into six types of tracking scenes: camera motion, target empty, illumination change, target motion change, size change, and target occlusion. The evaluation method is supervised. In this mode, if the tracking algorithm fails to track the target for 5 consecutive frames, the real position of the target is given, and the tracking algorithm will be re-initialized to complete the subsequent tracking task. Meanwhile, VOT uses the three main indicators of Accuracy [40], Robustness and EAO (Expected Average Overlap) [41] to evaluate the tracking results.

We compare our tracker with some state-of-the-art trackers on VOT-2016 benchmark, including MDNet, DeepSRDCF [22], SiamFC [42], HCF. The comparison results are shown in Table 1. It can be seen that in terms of accuracy and robustness, DCOT is ahead of other algorithms, and has a certain improvement on the original algorithm. At the same time, it performs very well in EAO. As shown in Figure 9, DCOT is better than the compared methods. The abscissa indicates the robustness, i.e., the ratio of the number of the test sequences with less than 30 tracking failures to the total number of the test sequences, and the ordinate indicates the accuracy. The algorithms have better robustness and accuracy when their results are closer to the upper right corner.

D. QUALITATIVE EVALUATION

Figure 10 shows some results of the top tracker, including MDNet, SANet, SRDCF, CREST and DCOT, on 5 challenging sequences. For the box sequence, the target is the

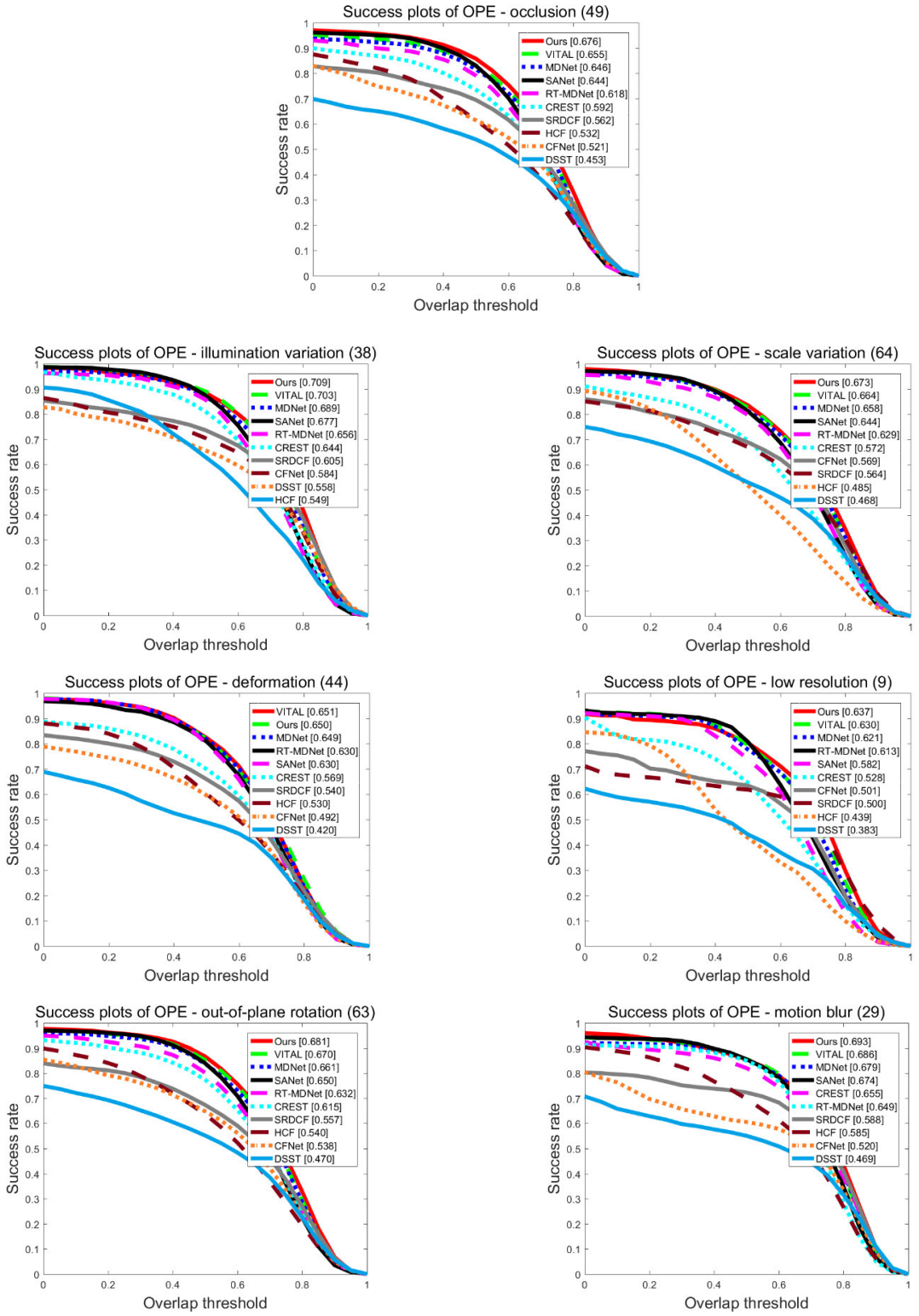

FIGURE 8. Success rate plots for the algorithms in the OTB100 where the targets are occluded, deformed, with low resolution, illumination variation, scale variation, out-of-plane rotation, in-plane rotation, out-of-view, motion blur, fast motion and interfered by the background clutter.

black box. When the box is occluded by a lighter, MDNet wrongly considers the black book is the target. The target models of the other trackers are polluted by the lighter. Due to PDS, the samples including the black book are figured out. Thanks to the long-term and short-term update mechanisms and confidence judgement, DCOT avoids the

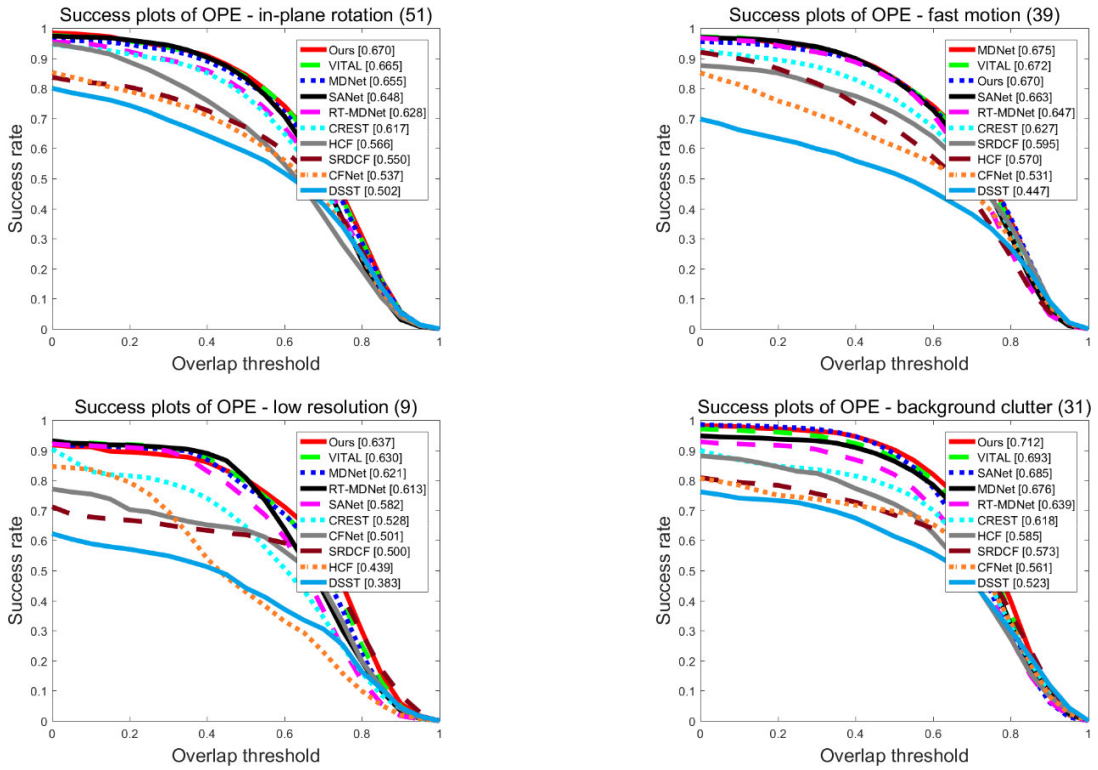

FIGURE 8. ***(Continued)*** **Success rate plots for the algorithms in the OTB100 where the targets are occluded, deformed, with low resolution, illumination variation, scale variation, out-of-plane rotation, in-plane rotation, out-of-view, motion blur, fast motion and interfered by the background clutter.**

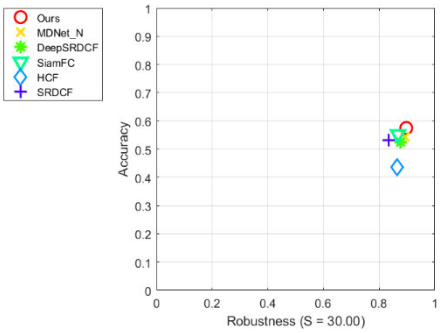

FIGURE 9. **VOT-2016 A-R diagram.**

TABLE 1. **Comparison on VOT-2016, the first and second best results of all indicators are marked with bold and italic.**

	Accuracy	Robustness	EAO
Ours	**0.57**	**17.82**	**0.33**
MDnet	0.54	21.08	0.26
DeepSRDCF	0.52	*20.36*	0.28
SiamFC	*0.55*	24.00	*0.28*
SRDCF	0.53	28.32	0.25
HCF	0.44	23.86	0.22

TABLE 2. **Comparison of various algorithm speeds.**

	DCOT	MDNet	SANet	SRDCF	CREST	HCF
FPS	0.891	0.989	0.005	8.11	1.00	15.3
AUC	0.676	0.646	0.644	0.562	0.592	0.532

interface of occlusion, and tracks the target correctly. For the bolt sequence, the target is the runner with red box in the first frame. The target changes a lot during running. SRDCF loses the target because it uses only low-level features. DCOT tracks the target successfully due to the good learning ability of MDNet that learns the target appearance. For the coupon sequence and the suv sequence, the target is disturbed by similarities. The resolutions of the sequences are low sometimes, MDNet fails to distinguish some mendacious objects, and accordingly fails to track. DCOT uses PDS to filter the mendacious objects out, and accordingly tracks correctly. For the lemming sequence, the target is a cat doll which is occluded by a lighter for quite a while. Due to the confidence judgement, DCOT stops updating, so it is not polluted by the occlusion. Above all, DCOT shows good performance in the challenging sequences.

E. ALGORITHM TIME EFFICIENCY ANALYSIS

Due to the parallel combination of deep network and correlation filter algorithm structure, the deep network needs to be updated online. The algorithm complexity is higher than that of the algorithms combining deep network features and correlation filter. We select the algorithms with similar performance (The success rate AUCs are close in

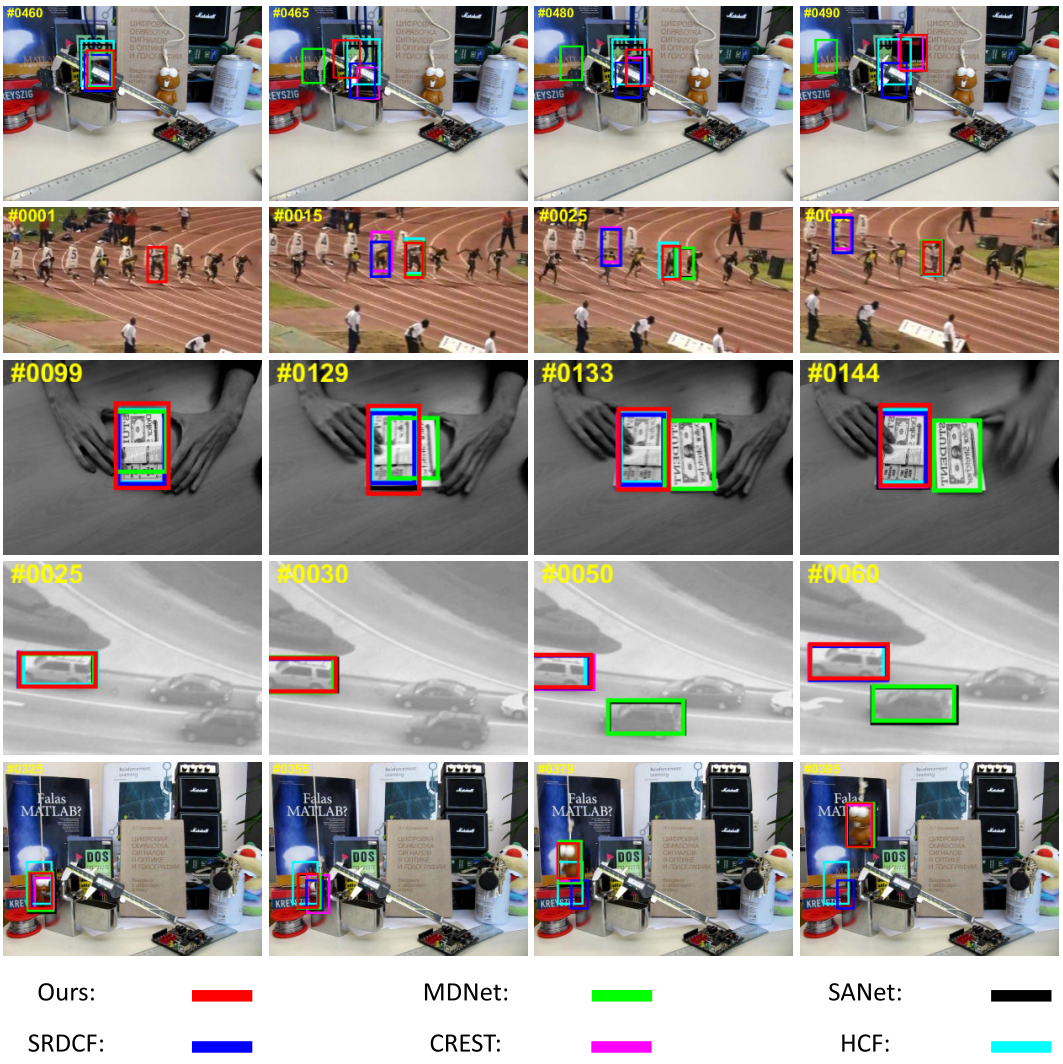

FIGURE 10. Performance of six tracker in box, bolt2, coupon, suv, and lemming(from up to down).

OTB), and compare their time efficiency (unit: FPS) on the same hardware platform. Performance indicator refers to OTB success rate AUC. The comparison results are shown in Table 2. It can be seen that the speed of DCOT is slower than those of the single structure correlation filter algorithms (SRDCF, HCF), and the algorithm speeds of most deep network structures are not remarkably different, but are obviously faster than SANet that is also the improvement on the basis of MDNet and has similar tracking effect. Because SANet adds the Recurrent Neural Networks (RNN) network layer to the original network structure, the network structure becomes more complicated and increases the computation complexity. In general, although DCOT is better than the compared algorithms, the time efficiency needs to be further improved.

V. CONCLUSION

In view of the inability of MDNet to distinguish between true target and the similar mendacious targets surrounding the true target, we propose DCOT to improve tracking performance. The predicted position information of the correlation filter is considered as the basis position to calculate the sample deviation degree. PDS removes the similar mendacious targets. Since the mendacious target interference is effectively limited, the tracking performance is improved. At the same time, the network confidence is used for the target state judgment of SRDCF, and helps SRDCF track correctly when the target is occluded or deformed. MDNet and SRDCF are fused to improve the robustness.

However, the deep network model lacks sufficient training samples to learn the target information. In some cases, some serious target deformation and target fast motion still lead to tracking failure. In addition, the speed of DCOT is similar to that of the general deep learning algorithms, but it is slower than the correlation filtering algorithms with a single structure. We will try to further accelerate the tracking speed in our future works to meet the real-time requirement.

REFERENCES

[1] D. Li, G. Wen, and Y. Kuai, "Collaborative convolution operators for real-time coarse-to-fine tracking," *IEEE Access*, vol. 6, pp. 14357–14366, 2018.

[2] J. Li, X. Zhou, S. Chan, and S. Chen, "Robust object tracking via large margin and scale-adaptive correlation filter," *IEEE Access*, vol. 6, pp. 12642–12655, 2018.

[3] X. Qi, W. Huabin, Z. Jian, and T. Liang, "Real-time online tracking via a convolution-based complementary model," *IEEE Access*, vol. 6, pp. 30073–30085, 2018.

[4] K. R. Reddy, K. H. Priya, and N. Neelima, "Object detection and tracking—A survey," in *Proc. CICN*, 2015.

[5] P. Li, D. Wang, L. Wang, and H. Lu, "Deep visual tracking: Review and experimental comparison," *Pattern Recognit.*, vol. 76, pp. 323–338, Apr. 2018.

[6] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classifcation with deep convolutional neural networks," in *Proc. NIPS*, 2012, pp. 1097–1105.

[7] H. Nam and B. Han, "Learning multi-domain convolutional neural networks for visual tracking," in *Proc. CVPR*, 2016, pp. 4293–4302.

[8] 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.

[9] M. Kristan, A. Leonardis, J. Matas, M. Felsberg, R. Pflugfelder, L. Čehovin, T. Vojír, G. Häger, A. Lukežič, G. F. Dominguez, A. Gupta, A. Petrosino, A. Memarmoghadam, A. Garcia-Martin, A. Montero, A. Vedaldi, A. Robinson, A. Ma, A. Varfolomieiev, and Z. Chi, "The visual object tracking VOT2016 challenge results," in *Proc. ECCV Workshops*, 2016, pp. 777–823.

[10] M. Zhang, J. Xing, J. Gao, X. Shi, Q. Wang, and W. Hu, "Joint scale-spatial correlation tracking with adaptive rotation estimation," in *Proc. CVPR Workshops*, 2015, pp. 32–40.

[11] T. Zhang, C. Xu, and M.-H. Yang, "Learning multi-task correlation particle filters for visual tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, pp. 365–378, Feb. 2019.

[12] D. S. Bolme, J. R. Beveridge, B. A. Draper, and Y. M. Lui, "Visual object tracking using adaptive correlation filters," in *Proc. CVPR*, 2010, pp. 2544–2550.

[13] M. Danelljan, G. Häger, F. Khan, and M. Felsberg, "Accurate scale estimation for robust visual tracking," in *Proc. BMVC*, 2014, pp. 1–11.

[14] M. Danelljan, F. Shahbaz Khan, M. Felsberg, and J. Van de Weijer, "Adaptive color attributes for real-time visual tracking," in *Proc. CVPR*, 2014, pp. 1090–1097.

[15] Q. Guo, W. Feng, C. Zhou, C.-M. Pun, and B. Wu, "Structure-regularized compressive tracking with online data-driven sampling," *IEEE Trans. Image Process.*, vol. 26, no. 12, pp. 5692–5705, Dec. 2017.

[16] L. Bertinetto, J. Valmadre, S. Golodetz, O. Miksik, and P. H. Torr, "Staple: Complementary learners for real-time tracking," in *Proc. CVPR*, 2016, pp. 1401–1409.

[17] 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.

[18] M. Danelljan, G. Hager, F. S. Khan, and M. Felsberg, "Learning spatially regularized correlation filters for visual tracking," in *Proc. ICCV*, 2015, pp. 4310–4318.

[19] W. Feng, R. Han, Q. Guo, J. Zhu, and S. Wang, "Dynamic saliency-aware regularization for correlation filter-based object tracking," *IEEE Trans. Image Process.*, vol. 28, no. 7, pp. 3232–3245, Jul. 2019.

[20] R. Han, Q. Guo, and W. Feng, "Content-related spatial regularization for visual object tracking," in *Proc. ICME*, 2018, pp. 1–6.

[21] P. Zhang, Q. Guo, and W. Feng, "Fast and object-adaptive spatial regularization for correlation filters based tracking," *Neurocomputing*, vol. 337, pp. 129–143, Apr. 2019.

[22] M. Danelljan, G. Hager, F. S. Khan, and M. Felsberg, "Convolutional features for correlation filter based visual tracking," in *Proc. ICCV Workshops*, 2015, pp. 621–629.

[23] F. Li, C. Tian, W. Zuo, L. Zhang, and M.-H. Yang, "Learning spatial-temporal regularized correlation filters for visual tracking," in *Proc. CVPR*, 2018, pp. 1–11.

[24] A. Lukezic, T. Vojir, Z. L. Cehovin Zajc, J. Matas, and M. Kristan, "Discriminative correlation filter with channel and spatial reliability," in *Proc. CVPR*, 2017, pp. 6309–6318.

[25] D. Held, S. Thrun, and S. Savarese, "Learning to track at 100 FPS with deep regression networks," in *Proc. ECCV*, 2016, pp. 749–765.

[26] Q. Guo, W. Feng, C. Zhou, R. Huang, L. Wan, and S. Wang, "Learning dynamic siamese network for visual object tracking," in *Proc. ICCV*, 2017, pp. 1763–1771.

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

[28] Z. Zhu, G. Huang, W. Zou, D. Du, and C. Huang, "UCT: Learning unified convolutional networks for real-time visual tracking," in *Proc. ICCV*, 2017, pp. 1973–1982.

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

[30] Y. Song, C. Ma, L. Gong, J. Zhang, W. L. Rynson, and M.-H. Yang, "CREST: Convolutional residual learning for visual tracking," in *Proc. ICCV*, 2017, pp. 2574–2583.

[31] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR*, 2016, pp. 770–778.

[32] Q. Wang, Z. Teng, J. Xing, J. Gao, W. Hu, and S. Maybank, "Learning attentions: Residual attentional siamese network for high performance online visual tracking," in *Proc. CVPR*, 2018, pp. 4854–4863.

[33] Y. Song, C. Ma, X. Wu, L. Gong, L. Bao, W. Zuo, C. Shen, R. W. Lau, and M.-H. Yang, "Vital: Visual tracking via adversarial learning," in *Proc. CVPR*, 2018, pp. 8990–8999.

[34] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. WardeFarley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. NIPS*, 2014, pp. 2672–2680.

[35] H. Fan and H. Ling, "SANet: Structure-aware network for visual tracking," in *Proc. ICCV Workshops*, Jul. 2017, pp. 42–49.

[36] Q. Wang, J. Gao, J. Xing, M. Zhang, and W. Hu, "DCFNet: Discriminant correlation filters network for visual tracking," 2017, *arXiv:1704.04057*. [Online]. Available: https://arxiv.org/abs/1704.04057

[37] J. Valmadre, L. Bertinetto, J. F. Henriques, A. Vedaldi, and P. H. Torr, "End-to-end representation learning for correlation filter based tracking," in *Proc. CVPR*, 2017, pp. 5000–5008.

[38] C. Ma, J.-B. Huang, X. Yang, and M.-H. Yang, "Hierarchical convolutional features for visual tracking," in *Proc. ICCV*, 2015, pp. 3074–3082.

[39] I. Jung, J. Son, M. Baek, and B. Han, "Real-time mdnet," in *Proc. ECCV*, 2018, pp. 83–98.

[40] M. Kristan *et al.*, "The visual object tracking VOT2013 challenge results," in *Proc. ICCV*, Dec. 2013, pp. 98–111.

[41] 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 VOT2015 challenge results," in *Proc. ICCV*, 2015, pp. 1–23.

[42] L. Bertinetto, J. Valmadre, J. F. Henriques, A. Vedaldi, and P. H. Torr, "Fully-convolutional siamese networks for object tracking," in *Proc. ECCV*, 2016, pp. 850–865.

[43] M. Zhang, Q. Wang, J. Xing, J. Gao, P. Peng, W. Hu, and S. Maybank, "Visual tracking via spatially aligned correlation filters network," in *Proc. ECCV*, 2018, pp. 469–485.

JUN CHU received the Ph.D. degree from Northwestern Polytechnic University, Xi'an, China, in 2005.

She was a Postdoctoral Researcher with the Exploration Center of Lunar and Deep Space, National Astronomical Observatory, Chinese Academy of Sciences, from 2005 to 2008. She was a Visiting Scholar with the University of California, Merced, USA. She is currently the Director of the Key Laboratory of Jiangxi Province for Image Processing and Pattern Recognition, and a Full Professor with the School of Software, Nanchang Hangkong University. Her research interests include computer vision and pattern recognition.

Dr. Chu is a member of Computer Vision Special Committee and China Computer Federation.

XUJI TU received the bachelor's and master's degrees from Nanchang Hangkong University, Nanchang, China, in 2016 and 2019, respectively. His research interests include computer vision and image processing.

LU LENG received the Ph.D. degree from Southwest Jiaotong University, Chengdu, China, in 2012.

He performed his Postdoctoral Research at Yonsei University, Seoul, South Korea, and Nanjing University of Aeronautics and Astronautics, Nanjing, China. He was a Visiting Scholar with West Virginia University, USA. He is currently an Associate Professor with Nanchang Hangkong University. He has published more than 70 international journal and conference papers. He has been granted several scholarships and funding projects for his academic research. His research interests include computer vision, biometric template protection, and biometric recognition.

Dr. Leng is a member of Association for Computing Machinery (ACM), China Society of Image and Graphics (CSIG), and China Computer Federation (CCF). He is a reviewer of several international journals and conferences.

JUN MIAO received the Ph.D. degree from Nanchang University, Nanchang, China, in 2015.

He is currently a Researcher with the Key Laboratory of Jiangxi Province for Image Processing and Pattern Recognition and an Associate Professor with the School of Aeronautical Manufacturing Engineering, Nanchang Hangkong University. He has a Short Visit with the National Astronomical Observatory, Chinese Academy of Sciences. His research interests include computer vision, 3D reconstruction, and pattern recognition.