ROBUST DEEP TRACKING WITH TWO-STEP AUGMENTATION DISCRIMINATIVE CORRELATION FILTERS

Chunhui Zhang[1,2], *Shiming Ge*[1,*], *Yingying Hua*[1,2], *Dan Zeng*[3,*]

[1] Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100195, China
[2] School of Cyber Security, University of Chinese Academy of Sciences, Beijing 100049, China
[3] School of Communication and Information Engineering, Shanghai University, Shanghai 200444, China
{zhangchunhui, geshiming, huayingying}@iie.ac.cn dzeng@shu.edu.cn

ABSTRACT

Recently, deep trackers have proven success in visual tracking due to their powerful feature representation. Among them, discriminative correlation filter (DCF) paradigm is widely used. However, these trackers are still difficult to learn an adaptive appearance model of the object due to the limited data available. To address that, this paper proposes a two-step augmentation discriminative correlation filters (TADCF) approach to improve robustness. Firstly, we propose an online frame augmentation scheme to obtain rich and robust deep features which can effectively alleviate background distractors, leading to better generalization and adaptation of the learned model. Secondly, an object augmentation mechanism is implemented by exploiting rotation continuity restriction, which simultaneously models target appearance changes from rotation and scale variations. Extensive experiments on four benchmarks illustrate that the proposed approach performs favorably against state-of-the-art trackers.

Index Terms— Visual tracking, discriminative correlation filter, frame augmentation, object augmentation

1. INTRODUCTION

Generic visual tracking is one of the challenging and fundamental problem in the field of computer vision and artificial intelligence. It plays a crucial role in intelligent video surveillance, virtual reality, autonomous driving, UAV monitoring, etc. The tracking task is to predict an arbitrary object in the subsequent frames given the target position and size in the first frame, which may face a series of challenges, such as rotation, scale variation, occlusion, and deformation.

In recent years, discriminative correlation filter (DCF) based trackers have drawn great attention such as Staple [1], DSST [2], SAMF [3], SRDCF [4], and BACF [5]. Most of these trackers use handcrafted features, which impede their robustness and accuracy. Inspired by the great success of deep CNN in other computer vision tasks, the visual tracking community has been focus on exploiting the power of deep tracking. Some deep trackers such as ECO [6], DeepSTRCF [7], LADCF [8], and UPDT [9] have provided remarkable results on standard benchmarks [10]. However, it is still an open and activate research area to achieve perfect tracking.

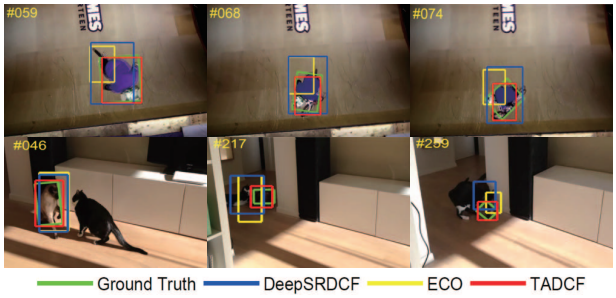

Fig. 1: Comparisons of our approach with two state-of-the-art trackers on VOT2018. TADCF is able to predict the shape more precise than DeepSRDCF [11], ECO [6] when object has large transformations, e.g. rotation or scale variations.

Existing trackers still suffer from the fundamental challenge [9] learning an adaptive appearance model of the object duo to the limited data available. Specifically, the lack of training data to model the background of the target over time will degrade the tracking performance [5]. Although some trackers explore fusing more deep features, the limited data is still an insurmountable gap. For another, deep CNN features capture high-level semantics, while being invariant to target transformations, e.g. rotation and scale variations, thus deep trackers including deep DCF trackers do not perform well when object appearance contains large changes.

To alleviate above problems, we propose a two-step augmentation discriminative correlation filters (TADCF) approach, which contains two simple yet effective mechanisms,

This work was partially supported by grants from National Natural Science Foundation of China (61772513, 61402463), National Key Research and Development Plan (2016YFC0801005), the Open Projects Program of National Laboratory of Pattern Recognition. Shiming Ge is also supported by Youth Innovation Promotion Association, Chinese Academy of Sciences.
*Corresponding authors: Shiming Ge (email:geshiming@iie.ac.cn) and Dan Zeng (email: dzeng@shu.edu.cn)

978-1-5386-9552-4/19/$31.00 ©2019 IEEE
DOI 10.1109/ICME.2019.00305

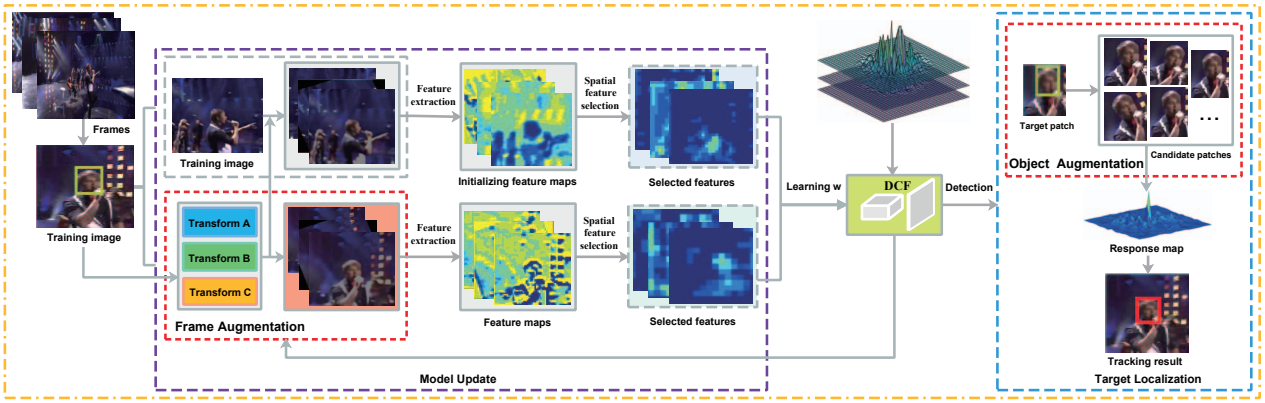

Fig. 2: Pipeline of the TADCF model. First, online frame augmentation is introduced into model update step to preprocess training images before extracting feature maps, then spatial feature selection is used to obtain the selected features. Through online frame augmentation we can extract rich and robust deep features and learn an adaptive model, which can effectively solve the issue of modeling the background of the target due to the limited data available. On this basis, object augmentation is introduced into target localization step to address large object appearance transformations, e.g. rotation and scale variations.

namely online frame augmentation and object augmentation. We first develop online frame augmentation scheme to model the background of the target. In the process of extracting deep features, we conduct several data augmentation methods in advance, and then get the feature maps from original image and augmented images. The discriminative power of extracted features is further guaranteed by spatial feature selection. So that, we can learn a more generalized and adaptive model. In addition, we use object augmentation mechanism to model target appearance variations. Observing the change of target rotation is usually accompanied by the change of scale, we simultaneously estimate angle and scale variations. In this way, the number of candidate pairs is reduced without harming the tracking performance. The robustness is further considered by introducing the rotation continuity restriction. The proposed approach can precisely predict the shape when object appearance has large transformations, as shown in Fig. 1.

The main contributions of this paper are threefold: 1) We develop an online frame augmentation scheme to facilitate the model generalizability. In this way, the dilemma between adaptive model and limited data is alleviated to some extent. 2) An object augmentation mechanism is proposed to handle large object appearance transformations for robust visual tracking. To the best of our knowledge, this is the first work to jointly consider rotation and scale variations in deep DCF tracker. 3) Extensive experiments show the proposed tracker achieves remarkable performance against state-of-the-arts.

2. RELATED WORK

The limited data in generic visual tracking is serious, where object is identified solely by a rectangle in the first frame. Using existing data, by flipping, blurring, rotating, and other data processing methods, more images can be created, thereby improving the adaptation and generalization ability of the tracking model. MOSSE [12] has used augmented gray-scale image samples to train correlation filters. DaSiamRPN [13] introduces motion blur into the data augmentation. Recent UPDT [9] explores the effect of different data augmentation techniques on both deep features and shallow features.

Deep features have been widely used in DCF trackers to boost their performance. DeepSRDCF [11] exploits shallow CNN features in a spatially regularized DCF framework. C-COT [14] solves the problem of training in continuous spatial domain. ECO [6] reduces the dimensions of the convolutional layer filter in the CNN features. However, deep features are usually invariant to image transformations [9, 15], such as in-plane rotation and scale variation. Recently, in order to improve the accuracy of tracking, some works try to exploit rotation estimation. Such as RAJSSC [15] simultaneously estimates angle and scale in Log-Polar domain, Siam-BM [16] performs angle estimation and spatial masking in siamese-based tracker. Inspired by recent temporal consistency preserving spatial feature selection model [8], we propose online frame augmentation and object augmentation to further improve the performance of deep DCF tracker.

3. PROPOSED METHOD

In this section, we present our TADCF approach for dealing background distractors and large object appearance transformations based on deep DCF tracker. To this end, we first introduce the formulation of our TADCF approach, after that discuss the online frame augmentation scheme, and then de-

scribe the object augmentation mechanism in details. The pipeline of proposed TADCF model is illustrated in Fig. 2. We crop a training image L^2 times larger than the bounding box of the target from current frame. In the first training image, we use online frame augmentation to augment image and extract initializing feature maps, after that perform spatial feature selection [8], and then use the selected features to initialize the discriminative correlation filters $\boldsymbol{w}$ [5]. In subsequent training images, online frame augmentation is used to update the learned filters. In object augmentation, candidate patches are obtained from target patch according to rotation and scale variations, then the target prediction is the region with the largest detection score of the response map.

3.1. Formulation

Let $\boldsymbol{x} = [\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_{n^2}]$ denote the feature maps of image samples with the size of $n \times n$, and $\boldsymbol{y} = [\boldsymbol{y}_1, \boldsymbol{y}_2, ..., \boldsymbol{y}_{n^2}]$ are the Gaussian shaped labels. The essence of spatial feature selection is spatial regularization, and temporal consistent constraints is temporal regularization. So, the proposed TADCF model can be equivalently formulated as,

$$\arg\min_{\boldsymbol{w}} \|\boldsymbol{w}\odot\boldsymbol{x}-\boldsymbol{y}\|_2^2+\lambda_1\|\boldsymbol{w}\|_1+\lambda_2\|\boldsymbol{w}-\boldsymbol{w}_{model}\|_2^2 \quad (1)$$

where $\odot$ is the convolution operation, λ_1 and λ_2 are tuning parameters and $\lambda_1 \ll \lambda_2$, $\boldsymbol{w}$ is the estimate filter and $\boldsymbol{w}_{model}$ is the template filter. Here, $\|\boldsymbol{w}\|_1$ and $\|\boldsymbol{w}-\boldsymbol{w}_{model}\|_2^2$ represent the spatial regularizer, and temporal regularizer, respectively.

We first use online frame augmentation (as in Sect. 3.2) to preprocess the training image samples before extracting the feature maps in model update step. Then, we adopt spatial feature selection to select several specific elements in the filter $\boldsymbol{w} \in \mathbb{R}^{n^2\times 1}$ as following,

$$\begin{aligned} &\arg\min_{\boldsymbol{w},\boldsymbol{\phi}} \|\boldsymbol{w}\odot\boldsymbol{x}-\boldsymbol{y}\|_2^2+\lambda_1\|\boldsymbol{w}\|_1 \\ &s.t.\ \boldsymbol{w}=\boldsymbol{w}_{\boldsymbol{\phi}}=\boldsymbol{diag}(\boldsymbol{\phi})\boldsymbol{w} \end{aligned} \quad (2)$$

where $\boldsymbol{diag}(\boldsymbol{\phi})$ represents the diagonal matrix of the selected features $\boldsymbol{\phi}$. Considering the multi-channel feature representations, such as deep features, we denote the multi-channel input as $\boldsymbol{x} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_D\}$ and the corresponding filters as $\boldsymbol{w} = \{\boldsymbol{w}_1, \boldsymbol{w}_2, ..., \boldsymbol{w}_D\}$, where D is the number of channels. $\boldsymbol{x}_i = [x_i^1, x_i^2, ..., x_i^{M^2}]^{\mathrm{T}}$ represents the ith channel, the spatial size of feature map is $M \times M$, and the jth spatial feature element in the ith channel represents as x_i^j. The objective function in Eq.(1) can be extended to multi-channel features,

$$\begin{aligned} \arg\min_{\boldsymbol{w}} &\sum_{i=1}^{D}\|\boldsymbol{w}_i\odot\boldsymbol{x}_i-\boldsymbol{y}\|_2^2+\lambda_1\left\|\sqrt{\sum_{i=1}^{D}\boldsymbol{w}_i\cdot\boldsymbol{w}_i}\right\|_1 \\ &+\lambda_2\sum_{i=1}^{D}\|\boldsymbol{w}_i-\boldsymbol{w}_{model\,i}\|_2^2 \end{aligned} \quad (3)$$

where $\cdot$ denotes the Hadamard product. The TADCF model can be efficiently solved using the Alternating Direction Method of Multipliers (ADMM) [5]. Accordingly, the augmented Lagrangian function can be expressed as,

$$\begin{aligned} \mathcal{L}=&\sum_{i=1}^{D}\|\boldsymbol{w}_i\odot\boldsymbol{x}_i-\boldsymbol{y}\|_2^2+\lambda_1\left\|\sqrt{\sum_{i=1}^{D}\boldsymbol{w'}_i\cdot\boldsymbol{w'}_i}\right\|_1 \\ &+\lambda_2\sum_{i=1}^{D}\|\boldsymbol{w}_i-\boldsymbol{w}_{model\,i}\|_2^2+\frac{\mu}{2}\sum_{i=1}^{D}\left\|\boldsymbol{w}_i-\boldsymbol{w'}_i+\frac{\boldsymbol{\eta}_i}{\mu}\right\|_2^2 \\ &s.t.\ \boldsymbol{w}=\boldsymbol{w'} \end{aligned} \quad (4)$$

where μ is the penalty parameter and $\boldsymbol{h}=\{\boldsymbol{\eta}_1, \boldsymbol{\eta}_2, ..., \boldsymbol{\eta}_D\}$ are the Lagrange multipliers. Then the ADMM algorithm can be applied to alternatingly update $\boldsymbol{w}$, $\boldsymbol{w'}$ and $\boldsymbol{h}$,

$$\begin{cases} \hat{\boldsymbol{w}}_i=\arg\min_{\boldsymbol{w}_i}\|\hat{\boldsymbol{w}}_i^*\cdot\hat{\boldsymbol{x}}_i-\hat{\boldsymbol{y}}\|_2^2+\lambda_2\|\hat{\boldsymbol{w}}_i-\hat{\boldsymbol{w}}_{model\,i}\|_2^2 \\ \qquad\qquad +\frac{\mu}{2}\left\|\hat{\boldsymbol{w}}_i-\hat{\boldsymbol{w}}'_i+\frac{\hat{\boldsymbol{\eta}}_i}{\mu}\right\|_2^2 \\ \boldsymbol{w'}=\arg\min_{\boldsymbol{w'}}\lambda_1\sum_{j=1}^{M^2}\left\|\boldsymbol{w'}^j\right\|_2+\frac{\mu}{2}\sum_{j=1}^{M^2}\left\|\boldsymbol{w}^j-\boldsymbol{w'}^j+\frac{\boldsymbol{\eta}^j}{\mu}\right\|_2^2 \\ \boldsymbol{h}=\boldsymbol{h}+\mu(\boldsymbol{w}-\boldsymbol{w'}) \end{cases} \quad (5)$$

where the hat denotes the frequency domain, $*$ represents the complex conjugate. Once the solution $\boldsymbol{w}$ to Eq.(3) is obtained, the updating of DCF model can then be obtained through the updating rule expressed as,

$$\boldsymbol{w}_{model}=(1-\gamma)\boldsymbol{w}_{model}+\gamma\boldsymbol{w} \quad (6)$$

where γ denotes the updating rate.

3.2. Frame Augmentation

Frame augmentation is used to preprocess images for extracting rich and robust deep features, so that the learned model can obtain high adaptive power in the presence of background distractors. The proposed frame augmentation consists of different data augmentation methods that can alleviate the problem of limited data. However, the usual practice of data augmentation is to train a model offline with a large amount of data, which requires very high quality and quantity of data, but the effect of a trained model on another particular video will be greatly reduced. Therefore, offline data augmentation is not optimal for visual tracking.

Different from many previous method, this paper proposes online frame augmentation (as shown in Fig. 2), which uses different data augmentation techniques before extracting deep features to get rich and robust features, and then synthesizes the features from training image and augmented images to obtain the final feature maps. We perform the following data augmentation methods, represented as Transform A, Transform B, Transform C, etc:

Color Jittering: Randomly change the original saturation and brightness of the image in the HSV color space.
Blur: Use Gaussian blur for the original image.
Flip: The original image is flipped horizontally.
Rotation: Rotate image at different angles from $-45°$ to $45°$.
Shift: Shift of n pixels in horizontal and vertical directions.

3.3. Object Augmentation

Object augmentation is employed to enhance the performance when object appearance has large changes. The change of target angle is usually accompanied by the change of scale. So, we can simultaneously address angle and scale variations. The proposed object augmentation mechanism (as shown in Fig. 2) enumerates some possible angles and increases the number of angle for adapting the object appearance transformations. Supposed that the choice of angle is K, the choice of scale variation is G, we have $K \times G$ possible combinations. In order to reduce the choice of combinations without damaging the performance of the proposed tracker, we can change the properties (angle or scale variation) of the tracked target only one each time. Thus, the number of candidate combinations is reduced from $K \times G$ to $N = K + G - 1$.

Given the template filter $\boldsymbol{w}_{model}$, the tracking process of DCF model is to find the optimal candidate that maximises the discriminate function in the current frame as following,

$$\arg\max_{\boldsymbol{x}_i} f(\boldsymbol{x}_i; \boldsymbol{w}_{model}) = \arg\max_{\boldsymbol{x}_i} \boldsymbol{w}_{model} \odot \boldsymbol{x}_i \quad (7)$$

where the candidate samples are usually generated by different scales around tracking result in previous frame, i.e. $\boldsymbol{x}_i = \boldsymbol{x}_i(x_i, y_i, s_i)$, (x_i, y_i) is the target location, s_i is the scale factor. However, our TADCF approach proposes object augmentation mechanism in target localization step to address different object appearance variations. In other words, we simultaneously consider angle and scale variations, i.e. $\boldsymbol{x}_i = \boldsymbol{x}_i(x_i, y_i, a_i, s_i)$. So, the Eq.(7) can be rewritten as,

$$\arg\max_{\boldsymbol{x}_i} f(\boldsymbol{x}_i; \boldsymbol{w}_{model}) + g(\boldsymbol{x}_i) \quad (8)$$

where $g(\boldsymbol{x}_i)$ is the object appearance detection score. We attempt to find the most credible target prediction (x, y, a, s), given the detection score $R_{a,s}$ of candidate patches $\boldsymbol{x}_i = (x_i, y_i, a_i, s_i), (i = 1, 2, ..., N)$ as following,

$$(x_i, y_i, a_i, s_i) = \arg\max_{\boldsymbol{x}_i} \mathrm{g}(\boldsymbol{x}_i) = \arg\max_{x,y,a,s} R_{a,s} \quad (9)$$

where (x_i, y_i) denotes the center of the tracked target, a_i denotes the target angle and s_i represents the scale factor. We force that $a_i = 0$ (no angle change) when $s_i \neq 1$ (scale change), and $a_i \neq 0$ when $s_i = 1$.

Generally, the change of object angle is continuous. In other words, the target rotation could not exceed a predefined threshold between two successive frames. We denote the absolute value of the target angle change between the $i-1$th frame and the ith frame by Δa , where $\Delta a = |a_i - a_{i-1}|$. Then, the target angle change score is determined by,

$$T_a = \begin{cases} 0.75, & \Delta a < Thr_a \\ 0.25, & \Delta a \geq Thr_a \end{cases} \quad (10)$$

Considering the target angle change score T_a , the Eq.(9) can be rewritten as following,

$$(x_i, y_i, a_i, s_i) = \arg\max_{\boldsymbol{x}_i} \mathrm{g}(\boldsymbol{x}_i) = \arg\max_{x,y,a,s} \lambda_c T_a R_{a,s} \quad (11)$$

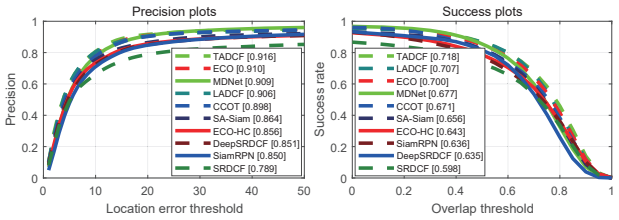

Fig. 3: Precision plots and Success plots on OTB2015 dataset using one-pass evaluation.

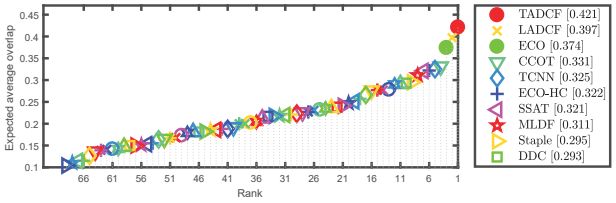

Fig. 4: EAO scores on VOT2016.

where λ_c is the angle change penalty factor.

4. EXPERIMENTS

4.1. Implementation Details

We provide a comprehensive evaluation of TADCF on OTB2015, VOT2016, VOT2017, and VOT2018. We apply HOG, Color-Names, and ResNet50 (the 4th layer) as features. L is set to 5. The λ_1 and λ_2 are set to 1 and 15, respectively. The penalty parameter μ is set to 1. The updating rate $\gamma$is set to 0.13. We set the number of angle as $K = 5$, and the number of scale as $G = 5$. The threshold Thr_a is $45°$. The parameter λ_c is set to 0.83. In Eq.(9), we first fix target location (x_i, y_i) and scale s_i, and then calculate the corresponding detection score $R_{a,s}$ by rotating angle a_i of the candidate region $\boldsymbol{x}_i$. Our tracker is implemented on Matlab2017a with MatConvNet and all experiments are executed on a PC with a 2.9GHz CPU. The average speed of TADCF is 1.1 FPS.

4.2. Results on OTB2015

The OTB2015 benchmark is one of the most widely used dataset in evaluating trackers, which contains 100 challenging videos. We evaluate TADCF with numerous state-of-the-art methods including ECO [6], ECO-HC [6], CCOT [14], MDNet [17], SiamRPN [18], SA-Siam [19], SRDCF [11], DeepSRDCF [11], and LADCF [8]. The precision plots and success plots of one path evaluation are shown in Fig. 3. The area under curve (AUC) and mean distance precision (DP) are summarized in Table 1. It shows that our TADCF achieves the best AUC score 0.718 and the best DP score 0.916.

Table 1: Comparisons with the state-of-the-art trackers on OTB2015 in terms of AUC and DP scores.

Trackers	SRDCF	DeepSRDCF	SiamRPN	ECO-HC	SA-Siam	CCOT	MDNet	ECO	LADCF	TADCF
DP	0.789	0.851	0.850	0.856	0.864	0.898	0.909	0.910	0.906	**0.916**
AUC	0.598	0.635	0.636	0.643	0.656	0.671	0.677	0.700	0.707	**0.718**

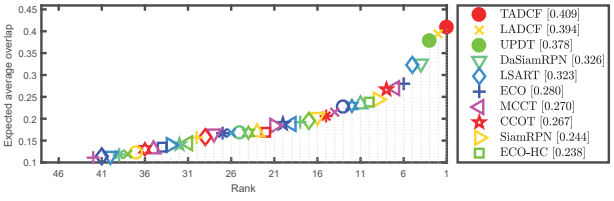

Fig. 5: EAO scores on VOT2017.

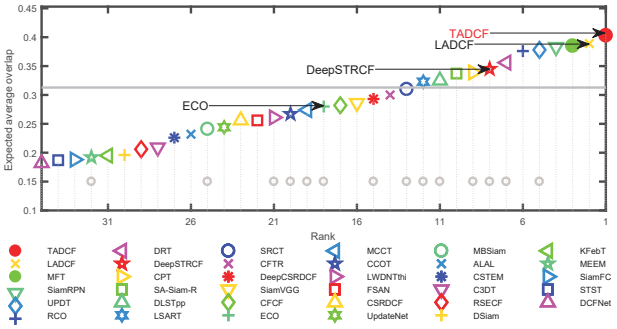

Fig. 6: EAO scores on VOT2018. The gray horizontal line indicates the average performance of fourteen trackers, which are denoted by gray circle in the bottom part of the graph.

4.3. Results on VOT Datasets

Results on VOT2016. There are 60 challenging videos on VOT2016. The EAO curve evaluated on VOT2016 is presented in Fig. 4, and 70 other state-of-the-art trackers are compared. We have added three recent algorithms: ECO [6], ECO-HC [6], and LADCF [8]. The final result shows TADCF ranking $1st$ according to EAO score, outperforming the LADCF by relative 6%, outperforming the winner tracker (CCOT) [14] on VOT2016 by relative 27.2%.

Results on VOT2017. For the evaluation on VOT2017, Fig. 5 reports the results of ours against 51 other state-of-the-art trackers with respect to the EAO score, here only listed prat of results. To further verify the performance of TADCF, we add several newly proposed trackers: UPDT [9], LADCF [8], SiamRPN [18], and DaSiamRPN [13]. TADCF ranks first with an EAO score of 0.409, which significantly outperforms the winer tracker (LSART) [21] with a relative gain of 26.6%. Our approach is based on DCF model, which is theoretically very simple, powerful and is also easy to follow.

Results on VOT2018. Our experiments contain all 72 state-of-the-art trackers on VOT2018. Fig. 6 shows the ranking results in terms of EAO, from which we can observe that TADCF outperforms the top performer (LADCF). For the sake of presentation clarity, we only show some top ranked trackers for evaluation. In Table 2, we list the EAO, Accuracy and Robustness of TADCF, DeepSTRCF [7], CCOT [14], ECO [6], UPDT [9], LADCF [8], SiamFC [20], SA-Siam-R [10], and SiamRPN [18]. SiamRPN obtains best Accuracy score because that region proposal network (RPN) [18] benefits localization accuracy. While our TADCF ranks $1st$ in EAO and Robustness. The top performance can be attributed to frame augmentation and object augmentation in our approach.

Table 2: Comparisons with the state-of-the-art trackers in terms of EAO, Accuracy, and Robustness on VOT2018.

Trackers	EAO	Accuracy	Robustness
CCOT [14]	0.267	0.494	0.318
ECO [6]	0.280	0.484	0.276
DeepSTRCF [7]	0.345	0.523	0.215
UPDT [9]	0.378	0.536	0.184
SiamFC [20]	0.188	0.503	0.585
SA-Siam-R [10]	0.337	0.566	0.258
SiamRPN [18]	0.383	**0.586**	0.276
LADCF [8]	0.389	0.503	0.159
TADCF	**0.403**	0.537	**0.124**

4.4. Ablation Study

Here, we conduct ablation analysis of our tracker TADCF on VOT2015 benchmark. With different experimental settings, we obtain the following four variants of our tracker, which are respectively named as "Baseline","Baseline+F", "Baseline+O", and "TADCF". "Baseline" is the baseline tracker with no online frame augmentation and object augmentation. We adopt the character "F", "O" to denote online frame augmentation and object augmentation, respectively. The results of different variants are described in Table 3.

First, the baseline tracker cannot achieve more satisfying performance (0.379 in EAO, 0.558 in Accuracy, and 0.183 in Robustness). Second, the effectiveness of frame augmentation can be verified comparing "Baseline+F" with "Baseline", which contributes to the relative performance gains of 4.2%, 1.6% in EAO and Accuracy, relative reduction of 14.2% in Robustness. Third, the effectiveness of object augmentation can be validated by comparing "Baseline+O" with

Table 3: Ablation study of our approach.

Trackers	Baseline	Baseline+O	Baseline+F	TADCF
EAO	0.379	0.388	0.395	**0.434**
Accuracy	0.558	0.561	0.567	**0.573**
Robustness	0.183	0.164	0.157	**0.132**

"Baseline". Finally, we can observe that "TADCF" improves the "Baseline" by relative gains of 14.5%, 2.7% in EAO and Accuracy, relative reduction of 27.9% in Robustness.

5. CONCLUSIONS

In this paper, we propose a two-step augmentation DCF-based tracker to learn an adaptive appearance model of the object with limited data available. To alleviate background distractors, we perform online frame augmentation to improve generalizability and adaptability of the learned model. For addressing large object transformations, we use object augmentation simultaneously modeling target appearance changes from rotation and scale variations. The effectiveness of our approach is validated on OTB and VOT datasets.

6. REFERENCES

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

[2] M. Danelljan, G. Hger, F. Khan, and M. Felsberg, "Accurate scale estimation for robust visual tracking," in *BMVC*, 2014.

[3] Y. Li and J. Zhu, "A scale adaptive kernel correlation filter tracker with feature integration," in *ECCVW*, 2014.

[4] M. Danelljan, G. Hager, F. K. Shahbaz, and M. Felsberg, "Learning spatially regularized correlation filters for visual tracking," in *ICCV*, 2015.

[5] K. H. Galoogahi, A. Fagg, and S. Lucey, "Learning background-aware correlation filters for visual tracking," in *ICCV*, 2017.

[6] M. Danelljan, G. Bhat, F. S. Khan, and M. Felsberg, "Eco: Efficient convolution operators for tracking," in *CVPR*, 2017.

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

[8] T. Xu, Z. Feng, X. Wu, and K. Josef, "Learning adaptive discriminative correlation filters via temporal consistency preserving spatial feature selection for robust visual tracking," *arXiv:1807.11348*, 2018.

[9] G. Bhat, J. Johnander, M. Danelljan, F. S. Khan, and M. Felsberg, "Unveiling the power of deep tracking," in *ECCV*, 2018.

[10] M. Kristan, A. Leonardis, J. Matas, and et al, "The visual object tracking vot2018 challenge results," in *ECCVW*, 2018.

[11] M. Danelljan, G. Hager, F. S. Khan, and M. Felsberg, "Convolutional features for correlation filter based visual tracking," in *ICCVW*, 2015.

[12] B. S. David, B. J. Ross, and D. A. Bruce, "Visual object tracking using adaptive correlation filters," in *CVPR*, 2010.

[13] Z. Zhu, Q. Wang, B. Li, W. Wu, J. Yan, and W. Hu, "Distractor-aware siamese networks for visual object tracking," in *ECCV*, 2018.

[14] M. Danelljan, A. Robinson, F. S. Khan, and M. Felsberg, "Beyond correlation filters: Learning continuous convolution operators for visual tracking," in *ECCV*, 2016.

[15] M. D. Zhang, J. Xing, J. Gao, X. Shi, Q. Wang, and W. Hu, "Joint scale-spatial correlation tracking with adaptive rotation estimation," in *ICCVW*, 2015.

[16] A. He, C. Luo, X. Tian, and W. Zeng, "Towards a better match in siamese network based visual object tracker," in *ECCVW*, 2018.

[17] H. Nam and B. Han, "Learning multi-domain convolutional neural networks for visual tracking," in *CVPR*, 2016.

[18] B. Li, J. Yan, W. Wu, Z. Zhu, and X. Hu, "High performance visual tracking with siamese region proposal network," in *CVPR*, 2018.

[19] A. He, C. Luo, X. Tian, and W. Zeng, "A twofold siamese network for real-time object tracking," in *CVPR*, 2018.

[20] L. Bertinetto, J. Valmadre, J. F. Henriques, A. Vedaldi, and P. H. S. Torr, "Fully-convolutional siamese networks for object tracking," in *ECCV*, 2016.

[21] C. Sun, D. Wang, H. Lu, and M. H. Yang, "Learning spatial-aware regressions for visual tracking," in *CVPR*, 2018.