EasyChair Preprint

№ 7004

Deep Siamese Network with Co-Channel and Cr-Spatial Attention for Object Tracking

Fan Gao, Ying Hu and Yan Yan

EasyChair preprints are intended for rapid dissemination of research results and are integrated with the rest of EasyChair.

November 7, 2021

Deep Siamese Network with Co-Channel and Cr-Spatial Attention for Object Tracking

Fan Gao, Ying Hu, and Yan Yan

Nanjing University of Science and Technology, Nanjing 210094 , China
gaofan@njust.edu.cn

Abstract. Siamese trackers with offline training strategies have recently drawn great attention because of their balanced accuracy and speed. However, some limitations still remain to overcome, i.e., trackers cannot robustly discriminate target from similar background so far. In this paper, we propose a novel real-time co-channel and spatial attention based deeper Siamese network (DCANet). Our approach aims at dealing with some challenging situations like appearance variations, similar distractors, etc. Different from replacing the backbone network Alexnet with VGG16 directlty, we modified the structure of VGG16 which has no fully connective layer and padding operation. In addition, co-channel and spatial attention mechanisms were applied to our method to enhance feature representation capability. Channel attention and spatial attention were proposed towards computer vision problems before. However, considered the special structure of siamese network, we designed Co-channel attention module which helps to emphasize the important areas in the two branches simultaneously. When we directly add spatial attention to our tracker, the tracking effect falls. However with a crop operation placed after spatial attention our tracker can tracking better. We perform extensive experiments on three benchmark datasets, including OTB-2013, OTB-2015, VOT-2017, LaSOT and GOT-10k, which demonstrate that our DCANet gains a competitive tracking performance, with a running speed of more than 60 frames per second.

Keywords: Siamese network · Single object tracking · Attention mechanism.

1 Introduction

Visual object tracking is a critical issue in many areas of computer vision such as visual surveillance [1], pose estimation [2], etc. Trackers demand to efficiently locate the target object defined in the initial frame with a bounding box. In order to achieve robust tracking results, challenging cases should be taken into consideration, including occlusion, appearance variation, background distractors, etc.

Siamese network-based trackers have recently drawn great attention. The pioneering work SiamFC [3] treats tracking as learning similarity. By adopting offline tracking strategies, SiamFC achieves real-time speed. Because of its simple architecture and high speed, SiamFC has become the cornerstone of most Siamese network based trackers. However, the tracker cannot adapt to challenging cases. Further research is needed to improve the adaptability and discriminability of tracking network.Many trackers have extended the SiamFC architecture [4–8]. EAST [4] formulates tracking as a decision-making process to speed up the tracker. CFNet [5] embeds a correlation filter layer into the Siamese network so that it can benefit from historical information and meanwhile keep high speed. SA-Siam [6] introduces a semantic branch which comes from classification task to learn semantic features. DSiam [7] proposes a dynamic Siamese network, which can effectively learn the appearance variation. SiamRPN [9] introduces Region proposal network (RPN) and place it after a siamese network. The RPN structure transforms tracking into a binary classification task and a bounding box regression task. SiamRPN and its succeeding works [9–11] usually can achieve a more accurate output but with too many parameters introduced which need a long-time training process. Although trackers above have been improved a lot, there remains one problem to be addressed that is how to enhance the discriminative ability of the neural network and keep in a high running speed.

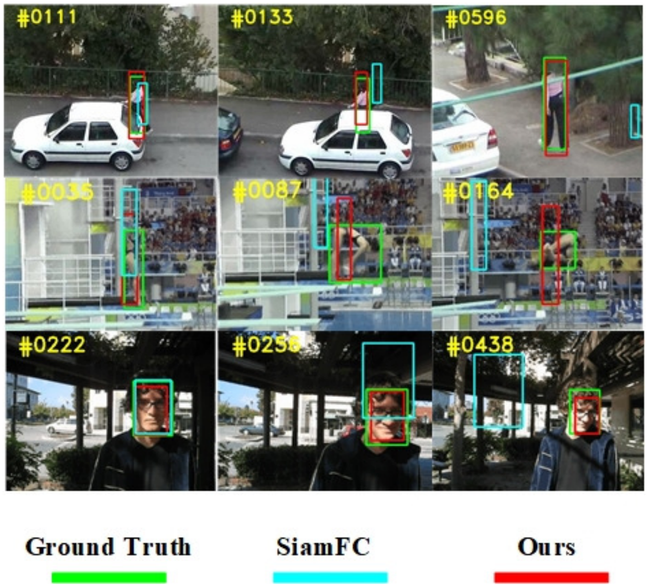


Fig. 1. Comparison of tracking performance results between SiamFC and DCANet. Benefiting from deeper backbone and Co-channel and Cr-spatial attention mechanisms, our tracker can locate the target successfully when SiamFC fails, especially towards some challenging tracking scenarios like scale change, occlusion, etc.

In this paper, we propose a novel Co-channel and Cr-spatial attention mechanism for visual object tracking. We squeeze the template feature as the channel weight and use the learned weight to excite both of the two branches. This modification is designed to make the instance branch keep pace with the template branch and enhance the discriminative capacity of the tracker. A spatial attention module modified from CBAM [12] is also adopted to take advantage of spatial information.

Our contributions are summarized as follows. (i) We propose a Co-channel attention module to associate two branches and emphasis channels related to the target category, which will enhance the discriminative capacity of the tracker. (ii) We place a Cr-spatial attention module into the instance branch to take advantage of spatial information. An additional crop operation is placed after the attention module to mitigate the effect of padding. (iii) Fig.1 shows the comparison of tracking results between SiamFC and DCANet, it seems that our tracker gains better performance towards some challenging tracking scenarios like scale change, occlusion, etc. To show the effectiveness of the proposed idea more qualitatively, we evaluate our DCANet on OTB-2015, VOT-2017 and LaSOT [13] datasets, which demonstrate that our proposed DCANet tracker can gain competitive performance on these tracking benchmarks.

2 Related Work

2.1 Trackers based on Siamese Network

Siamese trackers [3–8] convert the tracking problem into a similarity learning problem to overcome the low-frame-rate challenge in traditional tracking methods. Siamese network based trackers use Siamese network as feature extraction module. Most trackers use shallow AlexNet as backbone network until SiamDW [10] and SiamRPN++ [11] come out. SiamDW investigates the destruction of translation invariance caused by padding operation and introduces cropping-inside residual (CIR) units into the model. SiamRPN++ propose an offset sampling strategy to mitigate the effect of padding. VGGNets [14] shows that deeper network could significantly improve the representative ability of features. Inspired by CIR units, we adopt a modified VGG16 model with cropping operations as the backbone to get a better feature extraction result.

2.2 Attention Mechanisms

In order to enhance the discriminability of network, we take attention mechanisms into consideration. Attention mechanisms originated from neuroscience field and have been widely used in computer vision. DAVT [15] adopts discriminative spatial attention. CSR-DCF [16] introduces color histograms to restrict correlation filter learning, which finally constructs a target spatial reliability map. Squeeze-and-Excitation (SE) block [17] models the interdependencies between channels to adaptively recalibrate channel-wise feature responses. RASNet [8] exploits SEblock to emphasize useful channel information. SASiam [6] also adopts the SEblock as a channel attention module in the semantic branch. CBAM [12] takes both SEblock and spatial attention mechanism into consideration.

3 The proposed algorithm

We propose a real-time tracker with Co-channel and Cr-spatial attention mechanism based on SiamFC. Fig.2 shows the architecture of our proposed tracker. The fundamental idea behind this design is that better feature representation, better output.

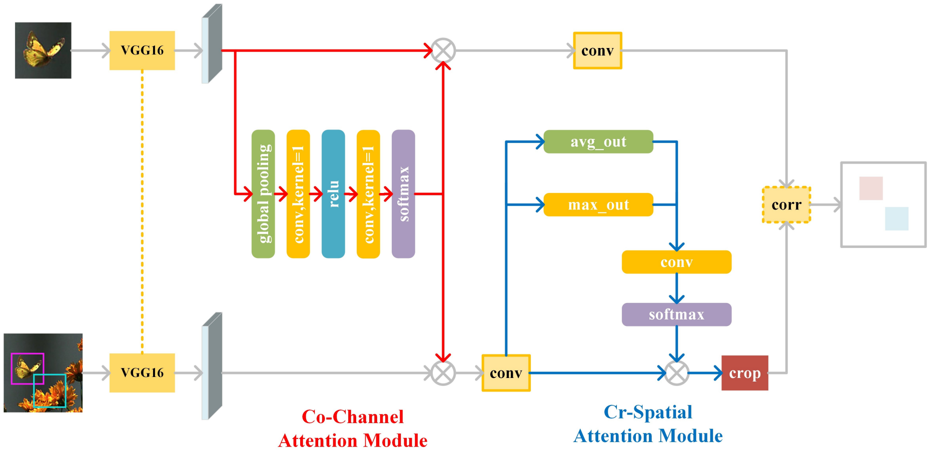


Fig. 2. The overall architecture of our DCANet tracker, which can be divided into four parts. First, a deeper backbone was applied to get better-extracted features. Second, we propose a Co-channel attention module to correlate the exemplar and search images and enhance channel information related to the target in parallel. Third, we adopt the spatial-attention module with a crop operation to take advantage of spatial information and mitigate the effect of padding. The last part is a cross-correlation operation, which plays a role in combination and finally calculates the output score map.

3.1 Siamese Tracker with Modified Backbone

Object tracking can be addressed as a learning similarity problem. Trackers are required to calculate the value function $f(x, z)$, which makes a comparison between template image z and instance image x.

Towards fully-convolutional Siamese architecture like SiamFC, the function φ represents the feature extraction part. $*$ denotes a cross-correlation operation which generates the output score map f. Thus, the final score map can be calculated by:

$$f(z, x) = \varphi(z) * \varphi(x) \tag{1}$$

the maximum score refers to the position of the tracking target by frames.

In our proposed approach, we change the backbone from AlexNet to modified VGG16, which helps us get more discriminative features by a deeper but not too complex feature extraction network. Some recent studies apply more deeper backbone like ResNet, which means it will need expand the training set and take a much longer time to train the network. With a single RTX 2080Ti GPU, it will need about a month to train the network with ResNet backbone. However, the modified VGG16 which selects the first layers without padding operations from VGG16 can gain a better tracking result while not extend the training time much. The training process of our tracker can be completed in less than a day with a single RTX 2080Ti GPU.

3.2 Co-Channel Attention Module

It is mentioned that objects belong to the same type will have a high response on particular channels. Meanwhile, the responses of other channels are suppressed. To enhance the learning of convolutional features by explicitly modeling channel interdependencies, SEblock [17] was proposed. This representative module takes advantage of the channel information and has been widely used in single-branch networks. Siamese network is a two-branch network so that we propose a Co-attention module to multiply the channel weight not only on the template branch but also on the instance branch. The two branches deal with the same object, which obviously belongs to the same categories. Thus we can process the channel information on two branches in parallel. Synchronous weight operation makes the tracker pay attention to the discriminative information in the tracking process. The proposed approach can be described as follows.

Given the template feature $Z \in \mathbb{R}^{H\times W\times C}$, first, we squeeze Z by embedding global information to obtain channel weight $V = [v_1, v_2, ..., v_c]$, where v_c refers to the information learned from the $c-$th filter parameters. Then use a convolution operator to recalibrate the channel weight of input X and input Z in parallel. Final output of proposed Co-channel attention mechanism are two feature maps $U \in \mathbb{R}^{H\times W\times C} = [u_1, u_2, ..., u_c]$ and $N \in \mathbb{R}^{H\times W\times C} = [n_1, n_2, ..., n_c]$. The transformation can be presented as follows.

$$u_c = v_c * X = \sum_{s=1}^{C} v_c^s * x^s \tag{2}$$

$$n_c = v_c * Z = \sum_{s=1}^{C} v_c^s * z^s \tag{3}$$

3.3 Cr-Spatial Attention Module

In this part, we propose a module to take advantage of the spatial information. With a given input $X \in \mathbb{R}^{H\times W\times C}$, Cr-spatial attention module needs to generate a weight map $Q_s(X) \in \mathbb{R}^{H\times W}$ which records spatial information and indicates whether the place should be emphasized or suppressed. However, spatial attention module includes a padding operation which harms the translation invariance property of network. In our proposed model, we employ a crop operation placed after spatial attention module to overcome this problem. The proposed Cr-spatial attention can be computed as:

$$\begin{aligned} X' &= Cr(X * Q_s(X)) \\ &= Cr(X * \sigma(f^{3\times3}([X^s_{avg}; X^s_{max}]))) \end{aligned} \tag{4}$$

where X^s_{avg} and X^s_{max} respectively refer to average-pooled features and max-pooled features. $f^{3\times3}$ represents a 3×3 filter size convolution operation with padding equals to 1. σ refers to a sigmoid function. Cr means crop 1 element around the tensor of input.

4 Experiments

4.1 Implementation Details

We apply a pretrained VGG16 as the backbone network. The training dataset is GOT-10k [18] which is a large database for generic object tracking in the wild. In order to obtain robust results, we adopt data enhancement method. In detail, we generate images in advance by adding transformations to sequences. During training, we randomly select two pictures from the same sequence for each time. One as the template image resized to 127 × 127 pixels, another as the instance image resized to 255 × 255 pixels.

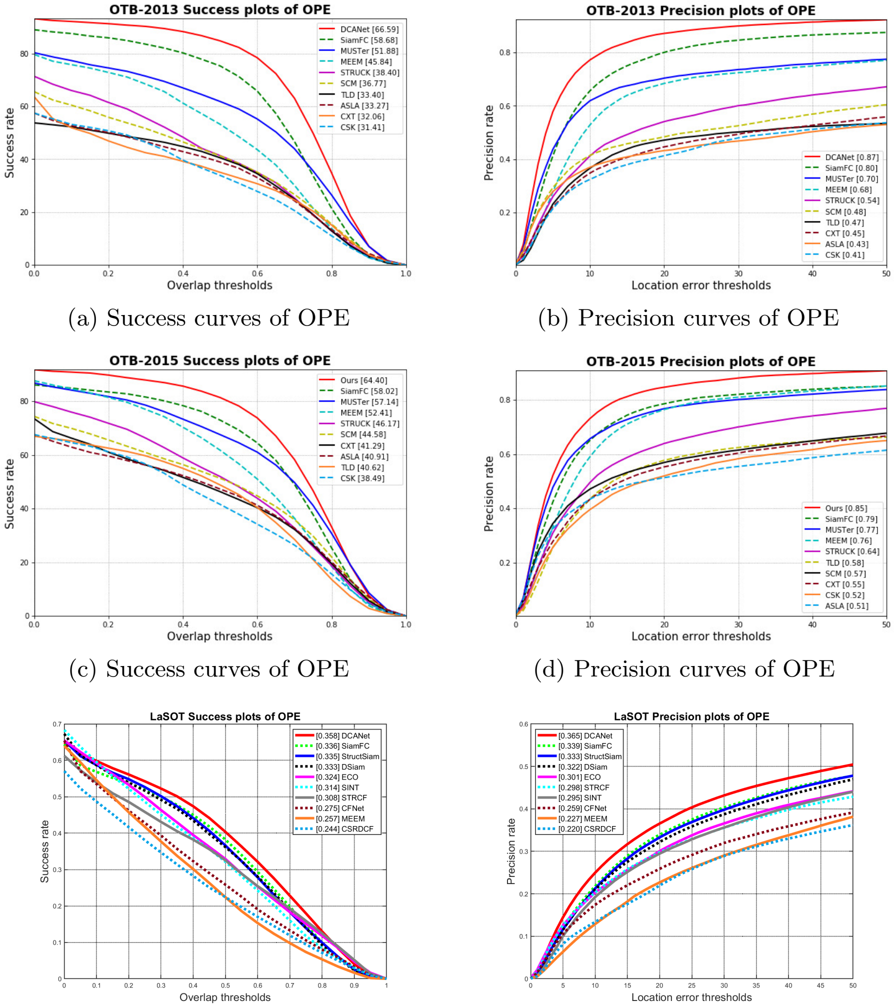
After offline end-to-end training, we perform our tracker and evaluate online tracking with the help of got-10k toolkit [18].

4.2 Evaluation Results of Visual Object Tracking

We evaluate the proposed DCANet on three different benchmarks including OTB-2013, OTB-2015, LaSOT, VOT-2017 and GOT-10k. The evaluation results are shown in Fig.3 and Fig.4. All of illustrations are obtained by OTB, VOT and LaSOT toolkit.

OTB2013 and OTB-2015 Dataset OTB-2013 dataset has 50 image sequences and OTB-2015 dataset has 100 image sequences in the benchmark. We make a comparison between the DCANet tracker and others, including SiamFC, Muster, MEEM, STRUCK, SCM, CXT, ASLA, TLD and CSK. As illustrated in Fig.3, our method gains a better performance than SiamFC.

LaSOT Dataset We compare the proposed tracker with several representative approaches, including SiamFC, StructSiam, DSiam, ECO, SINT, STRCF, CFNet, MEEM and CSRDCF. The results showed in Fig.3 demonstrate that our DCANet outperforms other trackers.



(a) Success curves of OPE

(b) Precision curves of OPE

(c) Success curves of OPE

(d) Precision curves of OPE

(e) Success curves of OPE

(f) Precision curves of OPE

Fig. 3. Experiment results on the OTB2013, OTB-2015 and LaSOT dataset. (a) and (b) are the results of the evaluation on OTB-2013 with other trackers, (c) and (d)are the results of the evaluation on OTB-2015 with other trackers, (e) and (f) are the results of the evaluation on LaSOT with other trackers.

VOT-2017 Dataset VOT2017 dataset consists of 60 sequences and the performance on this dataset is measured by accuracy, robustness and a comprehensive measurement named EAO. We make a comparison between our tracker and others on VOT-2017. The EAO results are shown in Fig.4, which demonstrates that the DCANet tracker gains a competitive performance and runs in real-time.

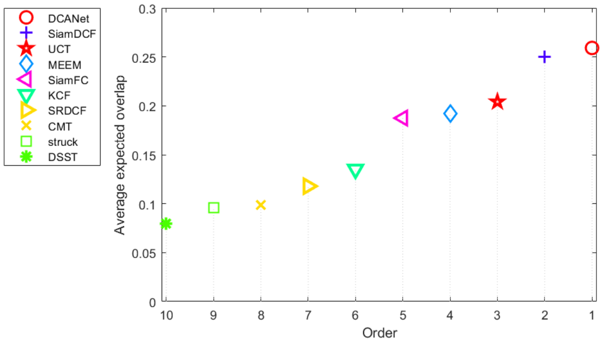
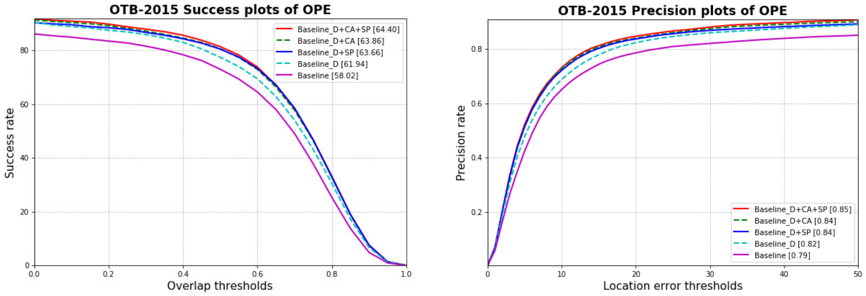


Fig. 4. EAO rank results on the VOT-2017 dataset. We only choose 10 trackers [19] to show for clarity. The measurement EAO involved in this dataset can show us the balance between tracking speed and precision. The higher the better.

GOT-10k dataset GOT-10k is a large, high-diversity, one-shot database for generic object tracking in the wild. Different from other datasets above, we use this dataset as our training data. The test set embodies 84 object classes and 32 motion classes with only 180 video segments, allowing for efficient evaluation. As shown in table.1, DCANet achieves the best performance among these trackers.

Table 1. Comparative results on GOT10k dataset.

Tracker	AO	SR $_{0.5}$	SR $_{0.75}$
KCF	0.203	0.177	0.065
CSK	0.205	0.174	0.056
DSST	0.247	0.223	0.081
CFNet	0.293	0.235	0.068
MDNet	0.299	0.303	0.099
ECO	0.316	0.309	0.111
GOTURN	0.347	0.375	0.124
SiamFC	0.348	0.353	0.098
DCANet	**0.403**	**0.466**	**0.150**



(a) Success curves of OPE

(b) Precision curves of OPE

Fig. 5. (a) and (b) are the results of ablation studies on OTB-2015. As shown above, our tracker DCANet gain the best performance with all components.

4.3 Ablation Studies

We ablate our method on several components to verify the effectiveness of each part. Take the SiamFC tracker as a baseline. First, we change the backbone of the network from Alexnet to a pretrained VGG16 (Baseline_D). Second, we employ the Co-channel attention module in both of the two branches(Baseline_D+CA). Third, we adopt a Cr-spatial attention module in the instance branch (Baseline_D+SP). In the end, we combine both of the two parts and get the final output (Baseline_D+CA+SP). As shown in Fig.5, the results on OTB-2015 demonstrate the effectiveness of each component in DCANet. Obviously, our DCANet achieves the best performance and gains a great improvement than the baseline SiamFC tracker by more than 6%.

5 Conclusion

In this paper, we propose an effective DCANet tracker. First, we propose a Co-channel attention mechanism to emphasize the channels related to target in both of template branch and instance branch. Second, we employ a Cr-spatial attention mechanism to take full advantage of spatial information. Benefits from modifications above, our DCANet achieves competitive tracking performance and runs in real-time. The results of experiments on OTB2013, OTB-2015, VOT-2017, and LaSOT datasets show the superiority of our proposed DCANet tracker.

References

1. Ali, A., Jalil, A., Niu, J., Zhao, X., Rathore, S., Ahmed, J., Iftikhar, M.A.: Visual object tracking—classical and contemporary approaches. Frontiers of Computer Science pp. 167–188 (2016)
2. Xiao, B., Wu, H., Wei, Y.: Simple baselines for human pose estimation and tracking. In: Proceedings of the European Conference on Computer Vision (ECCV) (September 2018)

3. Bertinetto, L., Valmadre, J., Henriques, J.F., Vedaldi, A., Torr, P.H.: Fully-convolutional siamese networks for object tracking. In: European conference on computer vision. pp. 850–865. Springer (2016)
4. Huang, C., Lucey, S., Ramanan, D.: Learning policies for adaptive tracking with deep feature cascades. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (October 2017)
5. Valmadre, J., Bertinetto, L., Henriques, J., Vedaldi, A., Torr, P.H.S.: End-to-end representation learning for correlation filter based tracking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (July 2017)
6. He, A., Luo, C., Tian, X., Zeng, W.: A twofold siamese network for real-time object tracking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2018)
7. Guo, Q., Feng, W., Zhou, C., Huang, R., Wan, L., Wang, S.: Learning dynamic siamese network for visual object tracking. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (October 2017)
8. Wang, Q., Teng, Z., Xing, J., Gao, J., Hu, W., Maybank, S.: Learning attentions: Residual attentional siamese network for high performance online visual tracking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2018)
9. Li, B., Yan, J., Wu, W., Zhu, Z., Hu, X.: High performance visual tracking with siamese region proposal network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2018)
10. Zhang, Z., Peng, H.: Deeper and wider siamese networks for real-time visual tracking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)
11. Li, B., Wu, W., Wang, Q., Zhang, F., Xing, J., Yan, J.: Siamrpn++: Evolution of siamese visual tracking with very deep networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)
12. Woo, S., Park, J., Lee, J.Y., Kweon, I.S.: Cbam: Convolutional block attention module. In: Proceedings of the European Conference on Computer Vision (ECCV) (September 2018)
13. Fan, H., Lin, L., Yang, F., Chu, P., Deng, G., Yu, S., Bai, H., Xu, Y., Liao, C., Ling, H.: Lasot: A high-quality benchmark for large-scale single object tracking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)
14. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
15. Fan, J., Wu, Y., Dai, S.: Discriminative spatial attention for robust tracking. In: European Conference on Computer Vision. pp. 480–493. Springer (2010)
16. Lukezic, A., Vojir, T., Cehovin Zajc, L., Matas, J., Kristan, M.: Discriminative correlation filter with channel and spatial reliability. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (July 2017)
17. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2018)
18. Huang, L., Zhao, X., Huang, K.: Got-10k: A large high-diversity benchmark for generic object tracking in the wild. IEEE Transactions on Pattern Analysis and Machine Intelligence (2019)

19. Mishra, D., Matas, J.: The visual object tracking vot2017 challenge results. In: 2017 IEEE International Conference on Computer Vision Workshops (ICCVW). pp. 1949–1972 (October 2017). https://doi.org/10.1109/ICCVW.2017.230