Residual Attention SiameseRPN for Visual Tracking

Xu Cheng(✉), Enlu Li, and Zhangjie Fu

School of Computer and Software, Nanjing University of Information Science and Technology,
Nanjing 210044, China
xcheng@nuist.edu.cn

Abstract. Visual tracking demands to perform the accurate object location given the object state of the first frame. The existing methods have proposed various ways to handle the challenging problems, yet few of them take the relationship between shallow features and deep semantic features into account. Based on an extensive analysis, we first propose a residual attention SiameseRPN visual tracking method for accurate object state estimation, which introduces the correlation filter in a Siamese network framework. A novel loss function is presented to enhance the discriminative capability. Our approach is derived from three different loss terms that is capable of training a model in a few iterations. Second, we present channel attention mechanism to improve the tracking performance, which is offline trained to capture the general features in the tracking. Third, the proposed tracking model is trained in end-to-end manner and takes full advantage of both low-level representation for correlation filter and high-level semantic features for deep object representation by using multi-task learning strategy which can mine the relationship from both levels. Our approach benefits from two complementary effects. Finally, extensive evaluation and ablation studies demonstrate the effectiveness of the proposed tracking approach. Our tracker achieves state-of-the-art performance on five challenging benchmarks, which proves great potentials in balancing accuracy and speed.

Keywords: Surveillance · Deep learning · Correlation filter · Siamese network · Attention

1 Introduction

Visual tracking is one of the fundamental tasks in computer vision, and has many practical applications, such as human-computer interaction, action recognition, scene understanding, visual navigation, automatic driving and so on. Although much progress has been done in the past decade, it still remains challenging for a tracker to work at a high speed and is robust to complex scenarios including occlusion, illumination variations, low resolution, background clutter, and motion blur.

Recent deep learning based trackers and correlation filter based trackers have shown great potential for robust and fast tracking. Although basic CF has a high running speed due to their element-wise multiplications using Fast Fourier Transform. For complex

Y. Peng et al. (Eds.): PRCV 2020, LNCS 12306, pp. 407–419, 2020.
https://doi.org/10.1007/978-3-030-60639-8_34

scenarios, however, the accuracy of basic CF trackers often drops considerably. Deep network model has been widely developed to improve tracking performance due to their strong feature representation. Most existing approaches rely on the amount of the training data. The deep network model is extensively trained on large benchmarks offline and aggressively learned the object sequences online. These approaches have achieved very good results on some recent challenges.

Despite all these significant progress, most trackers suffer from several weaknesses and still can't attain consummate results. First, the training datasets are far smaller than other visual datasets such as ImageNet. The insufficient training data may cause the deep network model ineffective when facing all kinds of tracking challenges. Second, deep features learned offline can't adapt to specific object or unseen categories well during the tracking. Third, model updating schemes from these methods inevitably affect the network model adaptability, which degrades the tracking accuracy and increases the computationally expensive. These limitations lead to inferior accuracy.

To tackle the above limitations, our contributions can be summarized as follows.

(1) We propose a residual attention SiameseRPN method for visual tracking, which is an end-to-end deep network architecture. A novel loss function from three different aspects is presented to enhance the discriminative capability. Correlation filter layer and semantic feature layer are used to mine the relationship both low-level and high-level features in multi-task learning framework.
(2) An effective attention mechanism is utilized within the Siamese network architecture, which offline learns feature representations to adapt online object tracking.
(3) Numerous experimental results on five challenging benchmarks show that the proposed tracking method achieves state-of-the-art performance.

The rest of the paper is organized as follows. In Sect. 2, we review related work of existing object tracking algorithms. Section 3 briefly introduces the generative adversarial network. In Sect. 3.3, we introduce our approach for visual tracking. In Sect. 4, we present experimental results in two tracking benchmarks. Finally, Sect. 5 concludes this paper.

2 Related Work

There are extensive surveys of visual tracking in literature [1, 2]. We mainly discuss the representative trackers based on deep learning and correlation filters.

Deep Learning Tracking. Deep learning has been widely used to improve tracking performance. Some tracking methods combine deep learning models with correlation filters such as HCF [3], DeepSRDCF [4], ECO [5]. Another method formulates tracking task as a classification or regression problem, including CNN-SVM [6], DeepTrack [7], FCNT [8], TSN [9]. The advantage of these trackers is that they utilize the superior representation power of deep features. However, tracking speed is reduced due to online updating of the deep network model.

Recently some deep model based approaches are trained on videos offline and used to track the object online through an end-to-end deep network learning such as MDNet

[10], CFNet [11], RTT [12], ACFN [13]. The aforementioned problems have been most successfully addressed by Siamese network architecture [14–20]. SINT [14] formulates tracking task as a verification problem and trains a Siamese network model for object matching during the tracking. Similar methods include SiamFC [15], SiamRPN++[16], SiamRPN [17], SiamMN [18], Deeper and Wider Network [19] and SINT++ [20], etc. The VITAL tracker [21] generates hard samples by using adversarial learning and leverages the class imbalance with an effective loss. These methods advance the development of end-to-end deep network model and achieve the promising results on some challenging benchmarks. However, deep network model may suffer from over-fitting due to deficiency of training data.

Correlation Filter Tracking. Recent advances of correlation filter (CF) have achieved great success in terms of speed and accuracy [22–34]. We arrange these algorithms in a hierarchy and classify them into two categories: Basic correlation filter based trackers and regularized correlation filter based trackers. Some basic CF trackers have been developed to boost performance in tracking by using scale estimation [23], spatial constraints [24], reducing boundary effects [25], and long-term tracking [26]. However, basic CF trackers are limited in their detection range since they require the filter size and patch size to be equal. To address this issue, several regularized CF based trackers are proposed, including SRDCF [24], STRCF [27], ACFN [13], DeepSRDCF [4], ECO [5], DMSRDCF [28], C-COT [29], DCFNet [31], CSR-DCF [30], SAMF [32], MCPF [33], ATOM [34], etc. Among others, some trackers combine CF with deep features, which have shown significant improvement.

In this work, we focus on residual attention SiameseRPN for visual tracking. Different from the goals of the above mentioned approaches, our multi-stream network architecture is proposed to address the problem of object drift by using attention mechanism in the sequences.

3 The Proposed Tracking Method

In this section, we first introduce the network architecture of the proposed approach, and then give a detailed training process and loss function. Finally, we apply our model to visual tracking task.

3.1 ResNet Based Siamese Tracking Method

The Siamese network based object tracking methods [14] formulate visual tracking as a matching problem between the object template and the search area. The similarity measure is learned from Siamese deep network structure. The object state is usually given in the first frame of the sequence and can be used as object template z. The goal is to find the most similar candidates from the following frame $\boldsymbol{x}$.

$$f(z, \boldsymbol{x}) = \phi(z) * \phi(\boldsymbol{x}) + b \tag{1}$$

where $\phi()$ is a semantic embedding space; $f()$ denotes a similarity function; b is bias.

Furthermore, SiameseRPN [17] is trained with a ResNet-50 backbone by the spatial aware sampling scheme, which can overcome the translation invariance, asymmetrical features for classification and regression.

Network Architecture: We utilize ResNet-50 as base network architecture and feature extractor. Different from the original ResNet which has a large stride of 32 pixels, we reduce the strides at the last block from 16 pixels and 32 pixels to 8 pixels by modifying the conv5 block to have unit spatial stride. We crop the center 7 × 7 regions as the object template features to reduce a heavy computational burden on the correlation module. Furthermore, we fine-tune ResNet to improve the tracking accuracy. The parameters of deep network model are jointly trained in an end-to-end manner. The flowchart of the proposed tracker is shown in Fig. 1.

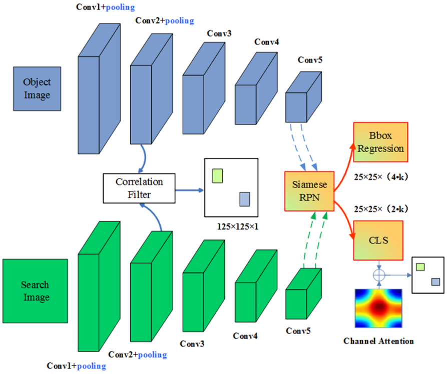

Fig. 1. The flowchart of the proposed tracking method. The intermediate layers in the common feature extractor have been omitted for clarity.

For the ResNet-50 network model, we utilize multi-level features extracted from the first residual block and the last residual block for tracking, respectively. The outputs of two-level features (Conv1, Conv5) are denoted as $\boldsymbol{F}_1(\cdot)$, and $\boldsymbol{F}_5(\cdot)$, respectively. On the one hand, object localization is obtained using the correlation filter working on the low-level fine-grained representations. Correlation filter is carried out as a differentiable layer. On the other hand, high-level semantic features are extracted from Conv5 and fed into the SiameseRPN module to achieve classification and regression tasks.

Offline Training: ResNet-50 deep network architecture is pre-trained on the training datasets of ImageNet, COCO, ImageNet DET and ImageNet VID to learn a general object feature for object representation. We employ single scale images with 127 pixels for template patches and 255 pixels for searching regions, respectively. To enhance the capacity to distinguish distracters of deep network model, we randomly obtain shifting and scaling following a uniform distribution on the search image as data augmentation techniques.

Our network model is trained with stochastic gradient descent (SGD). We use a warm-up learning rate of 0.001 for first 10 iterations to optimize the RPN branches.

For the last 15 iterations, the whole network is end-to-end trained with learning rate exponentially decayed from 0.005 to 0.0005. Weight decay of 0.0005 and momentum of 0.9 are used. We first train our model with for 5 warm-up epochs with learning rate linearly increased from 10^{-7} to 2×10^{-3}, then use a cosine annealing learning rate schedule for the rest of 45 epochs.

Discriminative Loss Function: The training loss is optimized from three different aspects. First, the classification loss and regression loss are written as follows.

$$L_{SiamRPN} = \frac{1}{N_{\text{pos}}} \sum_{x,y} L_{\text{cls}}(p_{x,y}, \boldsymbol{c}^*_{x,y}) + \frac{\lambda}{N_{\text{pos}}} \sum_{x,y} 1\{\boldsymbol{c}^*_{x,y} > 0\} \cdot L_{\text{reg}}(\boldsymbol{t}_{x,y}, \boldsymbol{t}^*_{x,y}) \tag{2}$$

where L_{cls} denotes the focal loss for the object classification; L_{reg} is the IoU loss for the object location; we assign 1 to $\boldsymbol{c}^*_{x,y}$ if it is a positive sample; N_{pos} is the number of positive samples; $1_{\{.\}}$ denotes the indicator function that takes 1 if the condition holds and takes 0 if not. $\boldsymbol{t}_{x,y}$ and $\boldsymbol{t}^*_{x,y}$ stands for the object position and ground-truth, respectively.

Second, different from CF that uses hand-crafted features for visual tracking, we develop to learn low-level feature representation fitting a CF. The features are obtained by a low-level convolutional layer of CNN model. The loss function is designed by

$$L_{low} = \|g(\boldsymbol{x}) - y\|_2^2 = \|\mathbf{X}\mathbf{w} - y\|_2^2 \tag{3}$$

where $\boldsymbol{x}$ is a search image; $\mathbf{X}$ is the circulate matrix of $\boldsymbol{x}$ for the search image patch; $\mathbf{w}$ is the learned CF.

Third, the high-level semantic is used to measure the similarities between the object template and the search image. The problem can be further written as the minimization of the following logistic loss.

$$L_{high} = \sum_{\boldsymbol{x},z} \log(1 + \exp(-y(\boldsymbol{x}, z) f(\boldsymbol{x}, z))) \tag{4}$$

The whole network is trained from end-to-end based on a multi-task learning strategy. The final loss can be overall formulated as follow.

$$L = L_{SiamRPN} + L_{low} + L_{high} \tag{5}$$

3.2 Channel Attention

A convolutional feature channel corresponds to certain visual information. In some certain circumstances, some feature channels are more important than others. The channel attention scheme is to keep the adaptation ability of deep network model to adapt the object appearance changes. To share a common attention, we propose a channel attention scheme to assist the object location.

The architecture of the channel attention is shown in Fig. 2, which is composed by a dimension reduction layer, a ReLU and a dimension increasing layer with sigmoid activation. Given a set of M channel features $\boldsymbol{F} = [\boldsymbol{f}_1, \boldsymbol{f}_2, \ldots, \boldsymbol{f}_M]$, the output of attention net is obtained by computing channel-wise re-scaling on the input in Eq. (7) where β is the parameter of the channel attention.

$$\bar{q}_i = \beta_i \cdot q_i \qquad i = 1, 2, \ldots, d \tag{6}$$

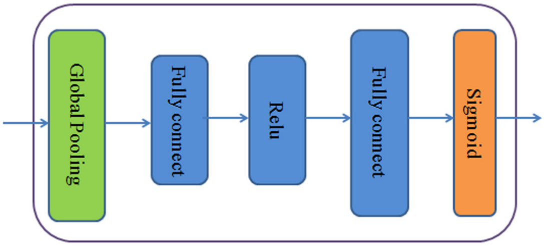

Fig. 2. The architecture of channel attention mechanism.

3.3 Online Tracking

Given the first frame with annotation, we utilize data augmentation strategies to construct an initial training set containing 20 positive samples. The object template is then obtained using the ResNet network architecture which is fine-tuned with the initial training set.

In the tracking phase, when a new input video frame arrives, we crop some large search patches centered at the previous target position with multiple scales. These search patches are fed into the ResNet-50 to get their feature representations. The fine-grained object representation is fed into the correlation filter layer. The semantic representation is evaluated based on the channel attention mechanism. The object candidate states $\boldsymbol{x} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N\}$ are randomly drawn based on the object position in the last frame. The candidates are estimated by finding the maximum of the fused correlation response in Eq. (7).

$$\boldsymbol{x}_t^* = \underset{i=1,....,M}{\arg\max}\, S(\boldsymbol{x}_i) \tag{7}$$

The candidate with the maximum object confidence score is considered as the tracking result. The parameters of deep network model are updated every 20 frames using positive and negative samples collected in previous tracking frames.

4 Experiments

4.1 Implementation Details

In this work, the proposed method is carried out in Python using Tensorflow and Keras deep learning libraries. We test our tracker on a PC machine with an Intel i7 CPU (32G RAM) and an NVIDIA GTX 1080Ti GPU (11G memory), which runs in real-time with 24.8 frames per second (fps). The quantitative analysis and ablation studies are evaluated in this section.

In the initial training phase, the convergence loss threshold is set to 0.02 and the maximum iteration number is 50. For the Siamese network framework, we use the initial object of the first frame as the object template and crop the search region with 3 times of the object size from the current frame. For the scale evaluation, we generate a proposal pyramid with three scales, i.e., 45/47, 1, and 45/43 times of the previous object size.

4.2 Overall Performance

We evaluate our approach with other competing trackers on five challenging tracking benchmarks, including OTB-100 [40], UAV123 [41], VOT2018 [35], LaSOT [37] and TrackingNet [45]. The proposed approach is compared with the state-of-the-art trackers, including the correlation filter based trackers, such as SRDCF [24], MCPF [33], C-COT [29], ECO [5], and STRCF [27]; the non-real-time deep tracking algorithms such as MDNet [10], CREST [38], LSART [43], VITAL [21], and DAT [44]; and the real-time deep learning tracking methods such as ACT [42], SiamFC [15], ATOM [34], CFNet [11], SiamRPN++ [16], LTMU [36], DaSiamRPN [39], and UPDT [46]. In the following, we will report the quantitative analysis on these benchmarks.

OTB-100: Table 1 shows the success overlap rate in the dataset. Among the compared trackers, our tracker obtains an AUC score of 68.1%, competitive with UPDT tracking method.

UAV123: The benchmark includes 123 low altitude aerial videos captured from a UAV. The AUC score on this benchmark is reported in Table 1. SiamRPN++ achieves an AUC score of 61.3%. Our tracker significantly outperforms SiamRPN++ and obtains AUC score of 63.4%.

Table 1. State-of-the-art trackers on OTB-100 and UAV123 benchmarks in terms of AUC score.

	ECO	CCOT	DaSiam RPN	ATOM	UPDT	MDNet	SiamRPN++	Our
OTB-100	64.3	68.2	65.8	66.9	69.2	67.8	68.9	68.1
UAV123	50.6	51.3	58.6	64.4	54.5	52.5	61.3	63.4

VOT2018: We evaluate our tracker on this challenging dataset which consists of 60 video sequences. Accuracy and robustness are used as measures to evaluate the tracking performance. EAO (Expected Average Overlap) is obtained to rank trackers. Results are given in Table 2. We can see that SiamRPN++ achieves the best performance in terms of accuracy. However, it obtains inferior robustness compared with ACT and ATOM. Our tracker has a 15.1% lower failure rate, while achieving compatible accuracy.

Table 2. Comparison of state-of-the-art trackers on VOT2018 benchmark.

	DAT	ACT	DaSiam-RPN	ATOM	UPDT	SiamRPN	SiamRPN++	Our
Accuracy	0.505	0.519	0.586	0.590	0.536	0.586	0.600	0.587
Robustness	0.140	0.201	0.276	0.204	0.184	0.276	0.234	0.151
EAO	0.385	0.356	0.383	0.401	0.378	0.383	0.414	0.440

LaSOT: We evaluate the proposed tracker on this dataset consisting of 280 sequences which have longer sequences with an average of 2500 frames per sequence. Therefore, it is important to adapt the object appearance variations. Figure 3 shows the success rate plot. ATOM tracker employs the pre-trained ResNet-18 to discriminate the object from the background. Our approach uses end-to-end trained method and further improves the performance with an AUC score of 51.6%. The experiment evaluations demonstrate that model adaption capabilities of the proposed tracking method on video sequences.

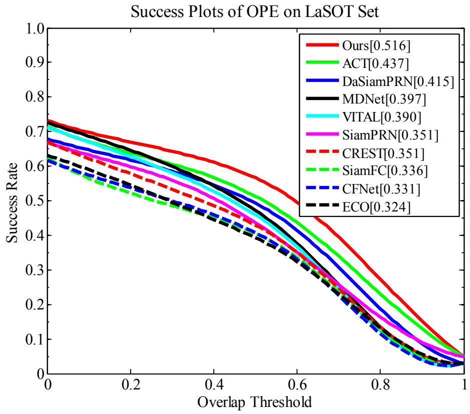

Fig. 3. Success plot on the LaSOT benchmark.

TrackingNet: We carry out our approach on the large-scale TrackingNet dataset. Table 3 shows the tracking evaluation results. SiamRPN++ reports a satisfied AUC score of 73.3%. Our method achieves AUC score of 74.1% with the same ResNet-50 as in SiamRPN++.

Table 3. Comparison of state-of-the-art trackers on TrackingNet benchmark.

	ECO	CFNet	MDNet	CSRDCF	UPDT	SiamFC	SiamRPN++	Our
AUC	55.4	57.8	60.6	53.4	61.1	57.1	73.3	74.1
P	49.2	53.3	56.5	48.0	55.7	53.3	69.4	69.6
P_{norm}	61.8	65.4	70.5	62.2	70.2	66.3	80.0	80.4

4.3 Ablation Studies

We conduct ablation evaluation to verify the contributions of different components and different layer features using OTB-100 and VOT2018 benchmarks. Table 4 shows the AUC scores of each variation.

Table 4. Ablation study of our method on OTB-100 and VOT2018 benchmarks. L_4 and L_5 denote conv4 and conv5, respectively. Finetune is whether the backbone is trained offline.

BackBone	L_4	L_5	Finetune	OTB-100	VOT2018
AlexNet				0.666	0.355
ResNet-50	√		√	0.679	0.347
		√	√	0.675	0.337
ResNet-50	√	√		0.676	0.392
	√	√	√	**0.700**	**0.408**

Feature Selection. The choice of features from different layers plays a significant role in visual tracking. The number of parameters and type of network layers directly affect the speed and accuracy of the tracking algorithms. First, different deep network architectures are evaluated on two popular benchmarks. AlexNet and ResNet-50 are used as backbones to verify the tracking performance. The AUC score is shown in Table 4. Our tracker and SiamRPN++ can benefit from the deeper layers network architecture. In addition, the tracking performance can obtain a great improvement by finetuning the backbone. Furthermore, the experiment results show that conv4 alone obtains a satisfying performance with 0.347 in EAO. Deeper layer and shallow layer perform with 5% drops. We combine conv4 and conv5 to obtain the improvements.

Effectiveness of Different Components. The proposed tracking approach consists of SiamRPN (S), correlation filter layer (CF), and channel attention module (A). To evaluate the importance of different components, we carry out the following variants: (1) ours (S) is our tracker merely using SiamRPN to track the object in every frame; (2) ours (CF) stands for our method by combining low-level correlation filter and high-level semantic representation to obtain the object location in every frame; (3) our tracker (A) denotes the proposed tracker with the channel attention module; and (4) our (S+CF+A) is the final tracker. The effectiveness of different components is evaluated in Table 5.

Table 5. Effectiveness of different components for our tracking algorithm.

Tracker	F-score	Pr	Re	fps
Ours (S)	0.553	0.551	0.541	34.7
Ours (CF)	0.583	0.584	0.557	30.6
Ours (A)	0.597	0.607	0.565	21.4
Ours (S+CF+A)	0.603	0.613	0.596	24.8

Table 5 shows the experimental results of the variants and illustrates that all components can boost the tracking performance. Removal of the channel attention module from the proposed model causes a 6.4% performance drop, while removal of the correlation filter layer reduces the performance by 7.8%. The accuracy of both variants is comparable to the original SiamRPN, while the number of failures increases. Therefore, the attention module is crucial for a robust object selection strategy during the tracking process. The proposed tracking approach leads to a 9.5% EAO and a 8.7% accuracy reduction. Therefore, it benefits from the rotated bounding box estimation.

5 Conclusions

In this paper, we propose an end-to-end deep network architecture for visual object tracking. Channel attention mechanism is introduced to the Siamese network framework. Low-level features are used to learn correlation filter and high-level semantic features are used to deep object representation. Then both two-level features are jointly represented in multi-task learning framework. The loss function is designed to optimize the deep network parameters. Experiments show that the proposed tracking approach significantly improves tracking performance in terms of accuracy and speed. In future work, we plan to incorporate spatial-temporal attention module representation in our model framework to further improve its effectiveness.

Acknowledgments. This work is supported in part by the National Natural Science Foundation of China (Grant No. 61802058, 61911530397); in part by the Equipment Advance Research Foundation Project of China (Grant No. 61403120106); in part by the Project funded by the China Postdoctoral Science Foundation (Grant No. 2019M651650); in part by the Startup Foundation for Introducing Talent of Nanjing University of Information Science and Technology (Grant No. 2018r057).

References

1. Li, P., Wang, D., Wang, L., et al.: Deep visual tracking: review and experimental comparison. Pattern Recogn. **76**, 323–338 (2018)
2. Zhu, P., Wen, L., Du, D., et al.: Vision meets drones: past, present and future. arXiv preprint arXiv:2001.06303 (2020)
3. Ma, C., Huang, J., Yang, X., et al.: Hierarchical convolutional features for visual tracking. In: Proceedings of the IEEE International Conference on Computer Vision, Piscataway, NJ, pp. 3074–3082. IEEE (2015)
4. Danelljan, M., Hager, G., Khan, F., et al.: Convolutional features for correlation filter based visual tracking. In: Proceedings of the IEEE International Conference on Computer Vision Workshops, Piscataway, NJ, pp. 58–66. IEEE (2015)
5. Danelljan, M., Bhat, G., Khan, F., et al.: ECO: efficient convolution operators for tracking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Piscataway, NJ, pp. 6638–6646. IEEE (2017)
6. Hong, S., You, T., Kwak, S., et al.: Online tracking by learning discriminative saliency map with convolutional neural network. In: International Conference on Machine Learning, New York, NY, pp. 597–606. ACM (2015)
7. Li, H., Li, Y., Porikli, F.: DeepTrack: learning discriminative feature representations online for robust visual tracking. IEEE Trans. Image Process. **25**(4), 1834–1848 (2015)

8. Wang, L., Ouyang, W., Wang, X., et al.: Visual tracking with fully convolutional networks. In: Proceedings of the IEEE International Conference on Computer Vision, Piscataway, NJ, pp. 3119–3127. IEEE (2015)
9. Teng, Z., Xing, J., Wang, Q., et al.: Robust object tracking based on temporal and spatial deep networks. In: Proceedings of the IEEE International Conference on Computer Vision, Piscataway, NJ, pp. 1144–1153. IEEE (2017)
10. Nam, H., Han, B.: Learning multi-domain convolutional neural networks for visual tracking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Piscataway, NJ, pp. 4293–4302. IEEE (2016)
11. Valmadre, J., Bertinetto, L., Henriques, J., et al.: End-to-end representation learning for correlation filter based tracking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Piscataway, NJ, pp. 2805–2813. IEEE (2017)
12. Cui, Z., Xiao, S., Feng, J., et al.: Recurrently target-attending tracking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Piscataway, NJ, pp. 1449–1458. IEEE (2016)
13. Choi, J., Chang, H., Yun, S., et al.: Attentional correlation filter network for adaptive visual tracking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Piscataway, NJ, pp. 4807–4816. IEEE (2017)
14. Tao, R., Gavves, E., Smeulders, A.: Siamese instance search for tracking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Piscataway, NJ, pp. 1420–1429. IEEE (2016)
15. Bertinetto, L., Valmadre, J., Henriques, João F., Vedaldi, A., Torr, P.H.S.: Fully-convolutional siamese networks for object tracking. In: Hua, G., Jégou, H. (eds.) ECCV 2016. LNCS, vol. 9914, pp. 850–865. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-48881-3_56
16. Li, B., Wu, W., Wang, Q., et al.: SiamRPN++: evolution of Siamese visual tracking with very deep networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Piscataway, NJ, pp. 4282–4291. IEEE (2019)
17. Li, B., Yan, J., Wu, W., et al.: High performance visual tracking with Siamese region proposal network. In; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Piscataway, NJ, pp. 8971–8980. IEEE (2018)
18. Zhou, W., Wen, L., Zhang, L., et al.: SiamMan: Siamese motion-aware network for visual tracking. arXiv preprint arXiv:1912.05515 (2019)
19. 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, Piscataway, NJ, pp. 4591–4600. IEEE (2019)
20. Wang, X., Li, C., Luo, B., et al.: SINT++: robust visual tracking via adversarial positive instance generation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Piscataway, NJ, pp. 4864–4873. IEEE (2018)
21. Song, Y., Ma, C., Wu, X., et al.: VITAL: visual tracking via adversarial learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Piscataway, NJ, pp. 8990–8999. IEEE (2018)
22. Henriques, J., Caseiro, R., Martins, P., et al.: High-speed tracking with kernelized correlation filters. IEEE Trans. Pattern Anal. Mach. Intell. **37**(3), 583–596 (2014)
23. Danelljan, M., Häger, G., Khan, F., et al.: Discriminative scale space tracking. IEEE Trans. Pattern Anal. Mach. Intell. **39**(8), 1561–1575 (2016)
24. Danelljan, M., Hager, G., Khan, F., et al.: Learning spatially regularized correlation filters for visual tracking. In: Proceedings of the IEEE International Conference on Computer Vision, Piscataway, NJ, pp. 4310–4318 IEEE (2015)
25. Kiani Galoogahi, H., Sim, T., Lucey, S.: Correlation filters with limited boundaries. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Piscataway, NJ, pp. 4630–4638. IEEE (2015)

26. Ma, C., Yang, X., Zhang, C., et al.: Long-term correlation tracking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Piscataway, NJ, pp. 5388–5396. IEEE (2015)
27. Li, F., Tian, C., Zuo, W., et al.: Learning spatial-temporal regularized correlation filters for visual tracking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Piscataway, NJ, pp: 4904–4913. IEEE (2018)
28. Gladh, S., Danelljan, M., Khan, F., et al.: Deep motion features for visual tracking. In: 23rd International Conference on Pattern Recognition, Piscataway, NJ, pp. 1243–1248. IEEE (2016)
29. Danelljan, M., Robinson, A., Shahbaz Khan, F., Felsberg, M.: Beyond correlation filters: learning continuous convolution operators for visual tracking. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9909, pp. 472–488. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46454-1_29
30. Lukezic, A., Vojir, T., Zajc, L., et al.: Discriminative correlation filter with channel and spatial reliability. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Piscataway, NJ, pp. 6309–6318. IEEE (2017)
31. Wang, Q., Gao, J., Xing, J., et al.: DCFNet: discriminant correlation filters network for visual tracking. arXiv preprint arXiv:1704.04057 (2017)
32. Li, Y., Zhu, J.: A scale adaptive kernel correlation filter tracker with feature integration. In: Agapito, L., Bronstein, M.M., Rother, C. (eds.) ECCV 2014. LNCS, vol. 8926, pp. 254–265. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-16181-5_18
33. Zhang, T., Xu, C., Yang, M.: Multi-task correlation particle filter for robust object tracking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Piscataway, NJ, pp. 4335–4343. IEEE (2017)
34. Danelljan, M., Bhat, G., Khan, F., et al. Atom: accurate tracking by overlap maximization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Piscataway, NJ, pp. 4660–4669. IEEE (2019)
35. Kristan, M., et al.: The sixth visual object tracking VOT2018 challenge results. In: Leal-Taixé, L., Roth, S. (eds.) ECCV 2018. LNCS, vol. 11129, pp. 3–53. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-11009-3_1
36. Dai, K., Zhang, Y., Wang, D., et al.: High-performance long-term tracking with meta-updater. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE, Piscataway, NJ (2020)
37. Fan, H., Lin, L., Yang, F., et al.: LaSOT: a high-quality benchmark for large-scale single object tracking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Piscataway, NJ, pp. 5374–5383. IEEE (2019)
38. Song, Y., Ma, C., Gong, L., et al.: CREST: convolutional residual learning for visual tracking. In: Proceedings of the IEEE International Conference on Computer Vision, Piscataway, NJ, pp. 2555–2564. IEEE (2017)
39. Zhu, Z., Wang, Q., Li, B., Wu, W., Yan, J., Hu, W.: Distractor-aware Siamese networks for visual object tracking. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11213, pp. 103–119. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01240-3_7
40. Wu, Y., Lim, J., Yang, M.: Object tracking benchmark. IEEE Trans. Pattern Anal. Mach. Intell. **37**(9), 1834–1848 (2015)
41. Mueller, M., Smith, N., Ghanem, B.: A benchmark and simulator for UAV tracking. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9905, pp. 445–461. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46448-0_27
42. Chen, B., Wang, D., Li, P., Wang, S., Lu, H.: Real-time 'actor-critic' tracking. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11211, pp. 328–345. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01234-2_20

43. Sun, C., Wang, D., Lu, H., et al.: Learning spatial-aware regressions for visual tracking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Piscataway, NJ, pp. 8962–8970. IEEE (2018)
44. Pu, S., Song, Y., Ma, C., et al.: Deep attentive tracking via reciprocative learning. In: Neural Information Processing Systems, pp. 1931–1941. MIT Press, Cambridge (2018)
45. Müller, M., Bibi, A., Giancola, S., Alsubaihi, S., Ghanem, B.: TrackingNet: a large-scale dataset and benchmark for object tracking in the wild. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11205, pp. 310–327. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01246-5_19
46. Bhat, G., Johnander, J., Danelljan, M., Khan, F.S., Felsberg, M.: Unveiling the power of deep tracking. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11206, pp. 493–509. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01216-8_30