Siamese visual tracking with enriched semantics and dynamic template*

WANG Hui-san (王汇三) **and ZHANG Hong-ying** (张红颖)**

College of Electronic Information and Automation, Civil Aviation University of China, Tianjin 300300, China

(Received 25 April 2020; Revised 9 July 2020)

Siamese tracking methods have recently drawn extensive attention due to their balanced accuracy and efficiency. However, most Siamese-based trackers use shallow backbone network, in which extracting high-level semantic features is difficult. When the appearance of distractors and targets is particularly similar, these methods may lead to tracking drift or even failure. Considering this deficiency, we propose a Siamese network with enriched semantics, named ESDT. First, a semantic enrichment module (SEM) comprising dilated convolution layers is designed to improve the classification capability of the siamese tracker. In addition, the target template is updated adaptively to cope with the target texture information changes caused by illumination and blur and further promote the tracking performance. Finally, exhaustive experimental analysis on the public datasets shows that the proposed algorithm outperforms several state-of-the-art algorithms and could track the target stably despite disturbances.

Document code: A **Article ID:** 1673-1905(2021)04-0241-6

DOI https://doi.org/10.1007/s11801-021-0073-y

Object tracking has always been a major topic in computer vision research and is widely used in various fields[1], such as intelligent monitoring, medical diagnosis, and military strike[1]. High-performance visual tracking algorithms with good tracking accuracy and efficiency are required by many applications. However, these algorithms remain challenging due to practical factors, such as scale variation, fast motion, occlusions, deformation, and background clutter[2].

The key to designing a tracker with outstanding performance is to select effective features and corresponding classifiers. Trackers based on Siamese networks[3-5] have recently drawn considerable attention due to their high speed and accuracy.

Features from deep layers have strong semantic information and are invariant to object appearance changes, such as rotation and deformation of the target. Thus, these features are suitable for classifying different objects in the frame. Meanwhile, these semantic features are an ideal complement to appearance features learned in a similarity matching task. However, most Siamese trackers use shallow backbone networks, such as Refs.[6—8], in which extracting high-level semantic features is difficult. Thus, trackers have low robustness in complex scenarios where the target rapidly moves or the illumination changes. The tracking may drift or even fail when the distractors in the image are particularly similar to the target. In addition, most existing Siamese trackers use the target in the initial frame as the template and calculate the similarity between the subsequent frame tracking. However, these trackers use a fixed template, which could not adapt to appearance changes in the target.

Considering the lack of deep features, Sa-Siam[9] utilizes the following two sets of Siamese networks: semantic and appearance branches. The response maps of superficial appearance and deep semantic branches are added in a certain proportion to obtain the final response graph and achieve feature fusion. This approach achieves effective performance but with remarkable speed drops. FlowTrack[10] uses optical flow motion information in Siamese networks to improve feature representation and tracking accuracy, which is computationally expensive and a rather complex system.

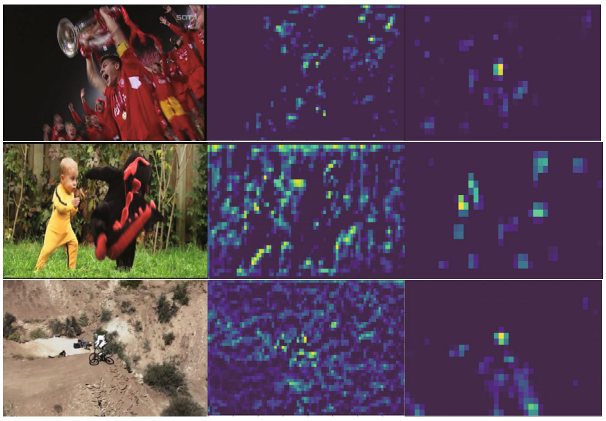
Siamese visual tracking with enriched semantics and dynamic template is explored in this paper to achieve real-time tracking with high accuracy and robustness. Inspired by the state-of-the-art single-shot object detection with enriched semantics[11], an improved semantic enrichment network is proposed to enhance the semantic information of targets during visual tracking. This network boosts the capability to distinguish foreground and semantic backgrounds of ESDT. In addition, target templates are updated adaptively in terms of the average peak-to-correlation energy (APCE)[12] update strategy, and the template learning rate is calculated on the basis of the average difference between the images. The main contributions of this study are listed as follows.

The semantic enrichment module (SEM) is introduced

* This work has been supported in part by the National Key Research and Development Project of China (No.2018YFB1601200), and the Fundamental Research Funds for the Central Universities (No.3122018C004).

** E-mail: carole_zhang@vip.163.com

into the Siamese network to enrich the semantic information of the target features during tracking. These semantic features could guarantee robust tracking despite the dramatic changes in the appearance of the target. The comparison of convolutional features visualization used in SiamFC and ESDT are shown in Fig.1. The target template is updated adaptively, and the template learning rate is calculated on the basis of the average difference between the images to learn the apparent changes in the target in the movement process. These appearance features are an ideal complement to the semantic features. Several experiments are conducted on mainstream public datasets. The results show that the proposed tracker outperforms several state-of-the-art algorithms and could track the target stably despite disturbances.



(a) (b) (c)

Fig.1 Comparison of convolutional features visualization used in SiamFC and ESDT (Different pixel colors indicate that the pixel belongs to different object categories): (a) Example images selected from Soccer, DragonBaby and MountainBike sequences; (b) Extracted features of the input image by convolutional layers in SiamFC; (c) Semantic enriched features used in the later stages for ESDT

The proposed tracking method is comprehensively presented in this section. First, the different characteristics between the features from shallow and deep convolutional layers are analyzed for visual tracking. Then, the semantic enrichment model mainly comprising dilated convolution layers is presented. Subsequently, the template update strategy is introduced.

SiamFC calculates similar matching images with the offline training and does not require online learning. Thus, this calculation can achieve excellent effects in real-time. The use of full convolutional networks achieves unrestricted input image size. Target features are extracted through the convolutional network with shared parameters. Finally, the output feature graphs are cross-correlated to calculate the response map as

$$f(Z,X)=g(\varphi(Z),\varphi(X)) , \tag{1}$$

where $g(\cdot)$ presents the cross-correlation function, and $\varphi(\cdot)$ represents the features extracted by the convolution network. Z and X are used to represent the input sample image and the search image, respectively. After calculating the cross-correlation score, the formula shown as Eq.(2) is used to calculate the loss function

$$\arg\min_{\theta}\frac{1}{N}\sum\nolimits_{i=1}^{N}\{L(f(z_i,X_i;\theta),Y_i\} . \tag{2}$$

SiamFC uses the AlexNet network[6], in which extracting high-level semantic features is difficult and target template updating is ignored. This deficiency leads to low robustness of trackers in complex scenarios when the target moves rapidly, the background of the target is similar to the foreground, or the illumination changes. Thus, SEM is proposed, and the update scheme is presented to achieve robust and accurate tracking.

The structure of the SEM network is shown in Fig.2. The SEM mainly comprises four dilated convolution layers[13] with 3×3 kernel size. The dilated convolution layer introduced the dilation rate into the ordinary convolution layer as a new parameter. The dilation rate defines the distance between the values during data processing of the convolution kernel. Considering maintaining the number of convolutional layers or the amount of network computation, the use of dilated convolution can enlarge the receptive field of the convolution kernel and reduce the complexity of the network model fundamentally. By contrast, the dilated convolution layer can aggregate the multiscale context information of the target flexibly.

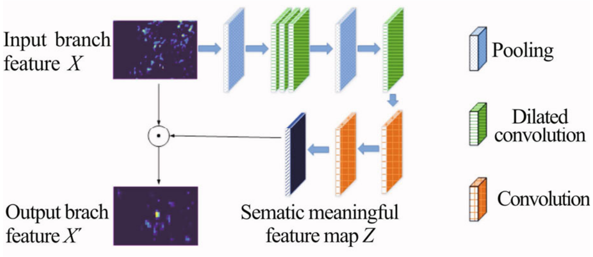


Fig.2 Structure of the SEM network which generates a semantically meaningful feature map *Z* to activate input *X* to be *X′* which is then used in the network for tracking

The first three dilated convolutional layers have a dilation rate of 2, and the last dilated convolutional layer has a dilation rate of 4. Another 1×1 convolutional layer is then deployed to generate $G(X)\in R^{C\times H\times W}$, where C represents the number of channels, H and W define the height and width of the feature map. $G(X)$ is the intermediate result used to generate semantic meaningful feature map as

$$Z=H(G(X))\in R^{C\times H\times W} . \tag{3}$$

The semantic meaningful feature map Z is then used to activate the feature map X by element-wise multiplication as

$$X'=X\Theta Z , \tag{4}$$

where X' is the semantically activated low-level detection feature map, which conveys basic visual patterns and high-level semantic information. X' replaces the original X in the template branch for object tracking.

In the Siamese network, the target in the initial frame is used as the template, and the subsequent tracking is calculated with the target template of the initial frame to estimate the regional feature similarity between the two frames. However, the target deforms and the illumination changes due to the rapid movement of the target. Thus, the tracking may fail when the template is not updated. The APCE referred by LCMF[12] is introduced in this paper as the tracking quality evaluation to stimulate the detector performance preferably in the tracking process. APCE is defined as

$$APCE=\frac{\left|F_{\max}-F_{\min}\right|^2}{mean(\sum_{w,h}(F_{w,h}-F_{\min})^2)}, \quad (5)$$

where $F_{\max}$, $F_{\min}$ and $F_{w,h}$ denote the maximum, minimum, and the (w, h) location response value on the response map, respectively. When the APCE score of the response map of frame t is higher than the set threshold, the target area z_t is clipped and inputted to the feature extraction network, and the target features of the frame are obtained. Then, the area is weighted with the previous $t-1$ frame template by adaptive learning rate.

$$\varphi_t^*=(1-\eta)\cdot\varphi^*_{t-1}+\eta\cdot\varphi(z_t), \quad (6)$$

where φ^*_{t-1} represents the features extracted by the convolution network of frame $t-1$,$\varphi(z_t)$ is the features of the target area z_t and φ_t^* define the new feature which computed by Eq.(6). The learning rate should be dynamically adjusted in accordance with the target changes. A small learning rate is set at this time to ensure tracking stability. When the target appearance dramatically changes, an extensive learning rate should be set to update the model rapidly. The target change of two adjacent frames is measured by calculating the average difference of two adjacent frames.

In the $M\times N$ image, the pixel size is denoted by $P_{i,j}$. The average difference between the images of frames t and $t-1$ can be obtained by

$$d=\frac{\sum_{i,j}^{M,N}\left|P_{ij}^t-P_{ij}^{t-1}\right|}{MN}. \quad (7)$$

When $d<3$, a low learning rate $\eta=0.025$ is set; when $3\leq d<8$, a moderate learning rate $\eta=0.05$ is set; when $d\geq 8$, a high learning rate $\eta=0.1$ is set. This adaptive piece-wise learning rate method ensures the robustness of the improved algorithm in tracking complex scenarios. The pipeline of the proposed ESDT tracker is shown in Fig.3.

This framework comprises a general CNN feature backbone network, an SEM, and a template update module. The pre-trained VGG-19 network is used as φ to extract target features. The semantic information of extracted features is then enriched by SEM. The target template is updated adaptively after the target position is predicted to track the target robustly despite the dramatic changes in the appearance of the target.

First, a square area in the template is cropped in terms of the target center location p and the target scale (w, h). The length of square side $s_z=(w+2p)\times(h+2p)$, where $p=(w+h)/4$ is the context margin. Then, this margin is multiplied by the scale factor s, and the template image is placed as 127×127×3. s satisfies $(w+2p)\times s(h+2p)=127$.

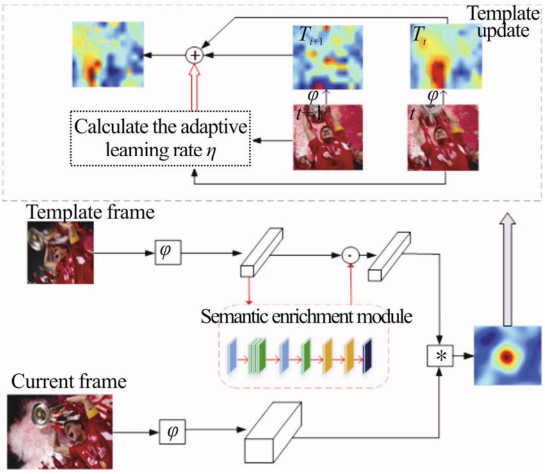


Fig.3 Network architecture of the proposed ESDT tracker

The search region of the current frame is obtained. The region of interest is determined in terms of the center position p_{t-1} of the last frame. The length of the side of region of interest (ROI) is $s_x=(s_z+2\times p_d)\times s$. The ROI image blocks are used as candidate samples to be placed as 255×255×3.

ILSVC-2015[14] is used to train the ESDT network and conduct testing on other benchmarks to avoid training and testing on the same dataset. ILSVC-2015 has many targets occupying the entire frame that is uncommon in real-world tracking tasks. Thus, 2800 ILSVC-2015 video sequences are selected, and 4 000 training clips are randomly generated, with each clip containing 10 successive frames.

The three-scale SiamFC is used as the pre-trained network to save training time, and the parameters in the pre-trained network are utilized as the initial values of the training parameters. The training method uses the stochastic gradient descent (SGD) method. The batch size of the training is set to 8, the number of times the data set is trained through the network is set to 50, and the number of samples for each training is set to 53 200. The total number of training steps is then set to 332 500 steps by $s_t=e\times n/b_s$, where s_t represents the total number of training steps, e is the total number of training, n is the sample number of each training, and b_s is the batch size. The output training loss for every 10 steps of training is set after the total number of training steps is determined. The loss function is defined as $l=\Sigma\log(1+e^{-y\times f})$, and this function is used for each batch of training samples, where l represents the loss of each batch of training samples, y represents the sample value, and f is the result of the output score map. After debugging, the weight decay of the training network is set to 0.000 5, the momentum is set to 0.9, and the learning rate is set to 0.000 1. Meanwhile, SGD is used to update the parameters of each training sample.

The proposed tracker experiment is implemented on a PC with Intel Core i7-9750 CPU (3.6 GHz), NVIDIA RTX2060 GPU, and 16 GB of memory. Several experiments are conducted to evaluate the SiamES tracker against numerous outstanding trackers on OTB2015[15] and VOT2017[16] benchmarks. Algorithm 1 shows the proposed tracking algorithm.

Public dataset OTB2015[15] exhibits 100 sequences with 11 attribute labels, including illumination variation, occlusion, deformation, scale variation, and background clutter. OTB2015 is used in the experiment for performance evaluation. The effect of algorithms is verified and examined through quantitative and qualitative analyses. Subsequently, seven representative and outstanding algorithms are selected for comparison. These algorithms are listed as follows: SiamRPN[17], SiamFC[5], DaSiamRPN[18], Sa_Siam[9], CFNet[19], ECO-HC[20], and BACF[21]. The original parameter settings of the algorithms are retained to ensure fairness.

The precision and success plots are used to evaluate the tracking algorithms comprehensively. Fig.4 illustrates the comparative results of the one-plot evaluation (OPE) over all the 100 sequences of OTB2015. The figure shows that the proposed method performs efficiently.

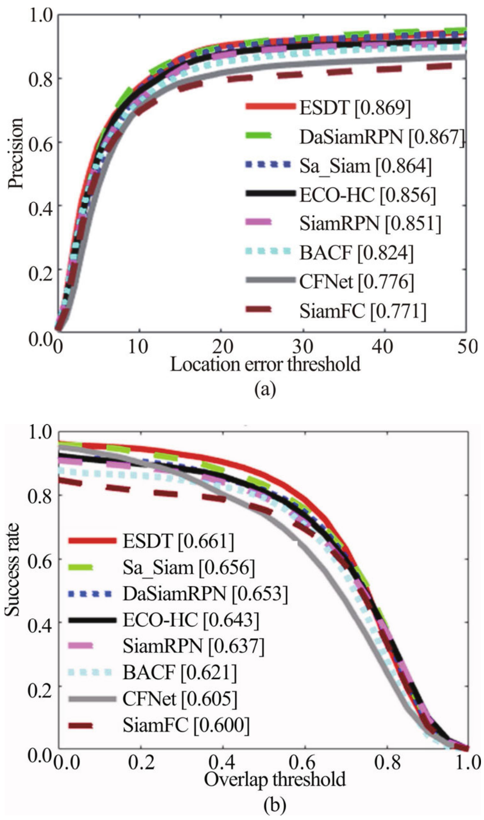


Fig.4 (a) Precision and (b) success plots of OPE for eight outstanding trackers on the OTB-2015 benchmark

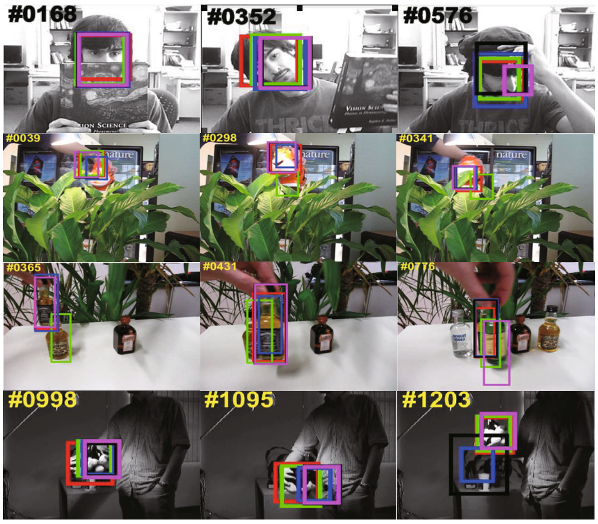
ESDT outperforms all the seven trackers in the precision and success plots of OPE. In the OPE success plot, the area under the curve (AUC) of the proposed approach is 0.661, which is higher than that of the baseline tracker SiamFC by 6.1%. In the OPE precision plot, the proposed tracker gains a precision score of 0.869, which exceeds those of SiamFC[5] and Sa_Siam[9] by 9.8% and 0.5%, respectively.

In addition to the precision and success rates, tracking speed is an important evaluation index. The speed of the tracker reflects tracker use in real-time tracking. Tab.1 shows the comparative results in terms of average frame per second (fps). ESDT achieves a speed of 78 fps, which is slower than the baseline tracker SiamFC[5] with 86 fps. However, ESDT can still satisfy real-time tracking.

Tab.1 Average speed comparison of tracking algorithms

Tracker	ECO-HC	Sa_Siam	SiamFC	SiamRPN	ESDT
Fps	60	50	86	160	78

The accuracy and robustness of the algorithm can be intuitively demonstrated through qualitative analysis. Fig.5 shows the five selected sub-datasets, which address different challenging aspects of tracking, and four representative and good performance approaches, namely ECO-HC[20], Sa_Siam[9], CFnet[19], and SiamFC[5], to test the proposed method.



—: ESDT, —: SiamFC, —: CFNet, —: Sa_Siam, —: ECO-HC

Fig.5 Tracking results of the OTB-2015 dataset (From top to bottom: FaceOcc, Tiger, Liquor, Sylvester)

The “FaceOcc” sequence tracks the head of a person. The target is small and goes through various complex scales and occlusion. The graph shows that Sa_Siam[9], SiamFC[5], CFNet[19], and ESDT can perform stead tracking. However, ECO-HC[20] cannot adaptively change the box size when the man lowers his face. However, ESDT is optimal in terms of tracking accuracy. In the “Liquor” sequences, the scales and appearances of the object change during the tracking process. Only ESDT and Sa_Siam can accurately locate the targets throughout the entire sequence. CFNet[19], SiamFC[5], and ECO-HC[20] lost the target. Illumination seriously changes in the sequence “Sylvester.” Some trackers lost the object when the illumination near the target considerably changed. However, the other methods, including ESDT, can still accurately track the object. The “Tiger” sequences demonstrate that CFNet[19] and ECO-HC[20] lost the object. Although the object is frequently out of view during the entire process, ESDT can accurately locate nearly all the objects.

The VOT2017[16] dataset contains 60 short sequences annotated with six different attributes. In VOT benchmarks, the accuracy (*A*), robustness (*R*), and expected average overlap (*EAO*) are used to evaluate the tracker performance. The trackers are ranked according to the *EAO* scores.

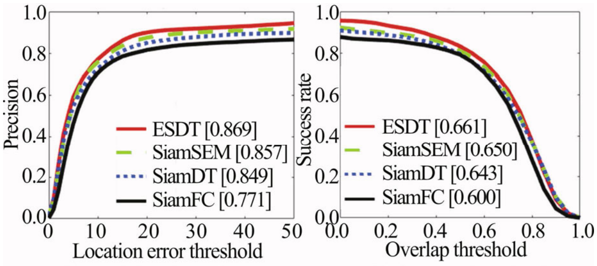
ESDT is compared with the top 9 trackers in the VOT2017[16]: C-COT[22], CSRDCF[23], SiamDCF[24], MCCT[25], ECO[25], CFCF[26], CFWCR[27], SiamFC[5], and SiamRPN[17]. Tab.2 shows that the proposed tracker achieves the highest robustness score while maintaining competitive *A* and *EAO* values. The *EAO* of ESDT is only 0.8% lower than that of CFWCR, which is first in rank. Although ESDT has a lower *EAO* score than CFWCR, ESDT is more robust than CFWCR due to the use of SEM. In addition, the accuracy score of EDST is lower than that of MCCT, which uses multi-cue correlation filters. However, the average overlap score of the proposed method is high. These results demonstrate that ESDT can achieve a balanced tracking performance in terms of reliability, accuracy, and robustness. SiamFC is our baseline tracker, the *A*, *R* and *EAO* criterion scores increased when SEM and template update strategy adopted.

Tab.2 Tracker performance comparison on VOT2017 (The bold number represents the best result, and the underlined one represents the second-best result.)

Trackers	*A*	*R*	*EAO*
C-COT	0.494	0.318	0.267
CSRDCF	0.491	0.356	0.256
SiamDCF	0.500	0.473	0.249
MCCT	**0.525**	0.323	0.270
ECO	0.483	0.276	0.280
CFCF	0.509	0.281	0.286
CFWCR	0.484	0.267	**0.303**
SiamRPN	0.490	0.460	0.244
SiamFC	0.500	0.591	0.188
ESDT	0.513	**0.264**	0.295

To verify the contributions of each component in our algorithm, the variations of our approach are implemented and evaluated. SiamSEM means using ESDT without template update module and SiamDT means ESDT without semantic enrichment module. SiamSEM, SiamDT and the baseline tracker SiamFC are evaluated on the benchmark of OTB-2015 and VOT-2017 as shown in Fig.6 and Tab.3.

Compared with SiamFC, SiamSEM use semantic enriched features for tracking, while the precision growth 5.6% measured by the AUC score as shown in Fig.6. The overall ESDT achieves a gain of 9.8% in AUC score compared with SiamFC. According to Tab.3, SiamSEM and SiamDT are compared with the baseline tracker SiamFC, improving respectively 2.7%, 7.1% in terms of *EAO*, which proves the effectiveness of the SEM and template update scheme in tracking.



(a) Precision plots of OPE (b) Success plots of OPE

Fig.6 Ablation experiments on the OTB-2015 benchmark

Tab.3 Ablation experiments on VOT2017

Trackers	A	R	EAO
SiamFC	0.500	0.591	0.188
SiamDT	0.506	0.415	0.215
SiamSEM	0.511	0.364	0.259
ESDT	0.513	0.264	0.295

A robust Siamese network with enriched semantics and dynamic templates is proposed in this study by introducing SEM. The SEM enriches the semantic information of features without using the deep networks and improves the classification capability of ESDT effectively. Furthermore, the target template is updated adaptively to cope with the change in target texture information caused by illumination and blur. Finally, experiments on mainstream public datasets are conducted to prove the effectiveness of the algorithm.

References

[1] Li X, Zha Y F, Zhang T Z, Cui Z, Zuo W M, Hou Z Q, Lu H C and Wang H Z, Journal of Image and Graphics **24**, 2057 (2019). (in Chinese)

[2] Li B, Wu W, Wang Q, Zhang F Y, Xing J L and Yan J J, SiamRPN++: Evolution of Siamese Visual Tracking with Very Deep Networks, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019.

[3] Zhang Z P and Peng H W, Deeper and Wider Siamese Networks for Real-Time Visual Tracking, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019.

[4] Fan H and Ling H B, Siamese Cascaded Region Proposal Networks for Real-Time Visual Tracking, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019.

[5] L. Bertinetto, J. Valmadre, J. F. Henriques, A. Vedaldi and P. H. Torr, Fully-convolutional Siamese Networks for Object Tracking, Proc. of European Conference on Computer Vision, 850 (2016).

[6] Krizhevsky A, Sutskever I and Hinton G.E, Imagenet Classification with Deep Convolutional Neural Networks, Advances in Neural Information Processing Systems, 1097 (2012).

[7] K. Simonyan and A. Zisserman, Very Deep Convolutional Networks for Large-Scale Image Recognition, Proc. of International Conference on Learning Representations, 2015.

[8] K. He, X. Zhang, S. Ren and J. Sun, Deep Residual Learning for Image Recognition, Proc. of IEEE Conference on Computer Vision and Pattern Recognition, 770 (2016).

[9] Anfeng He, Chong Luo, Xinmei Tian and Wenjun Zeng, A Twofold Siamese Network for Real-Time Object Tracking, Proc. of IEEE Conference on Computer Vision and Pattern Recognition, 2018.

[10] Zhu Z., Wu W., Zou W. and Yan J., End-to-end Flow Correlation Tracking with Spatial-Temporal Attention, Proc. of IEEE Conference on Computer Vision and Pattern Recognition, 2018.

[11] Zhang Z, S Qiao, C Xie, W Shen, B Wang and Alan Yuille, Single-Shot Object Detection with Enriched Semantics, Proc. of IEEE Conference on Computer Vision and Pattern Recognition, 2018.

[12] Wang M M, Liu Y and Huang Z Y, Large Margin Object Tracking with Circulant Feature Maps, Proc. of IEEE Conference on Computer Vision and Pattern Recognition, 4800 (2017).

[13] L.C. Chen, G. Papandreou, I. Kokkinos, K. Murphy and A. L. Yuille, Deeplab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs, IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017.

[14] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg and L. Fei-Fei, International Journal of Computer Vision **115**, 211 (2015).

[15] Y. Wu, J. Lim and M. Yang, IEEE Transactions on Pattern Analysis and Machine Intelligence **37**, 1834 (2015).

[16] M. Kristan, A. Leonardis, J. Matas, M. Felsberg, R. Pflugfelder, L. Cehovin Zajc, T. Vojir, G. Hager, A. Lukezic, A. Eldesokey and G. Fernandez, The Visual Object Tracking Vot2017 Challenge Results, Proc. of IEEE International Conference on Computer Vision, 1949 (2017).

[17] Bo Li, Junjie Yan, Wei Wu, Zheng Zhu and Xiaolin Hu, High Performance Visual Tracking with Siamese Region Proposal Network, Proc. of IEEE Conference on Computer Vision and Pattern Recognition, 8971 (2018).

[18] Zhu Z., Wang Q., Li B., Wu W., Yan J. and Hu W., Distractor-aware Siamese Networks for Visual Object Tracking, Proc. of European Conference on Computer Vision, 101 (2018).

[19] J. Valmadre, L. Bertinetto, J. F. Henriques, A. Vedaldi and P. H. Torr, End-to-end Representation Learning for Correlation Filter Based Tracking, Proc. of IEEE Conference on Computer Vision and Pattern Recognition, 2017.

[20] M. Danelljan, G. Bhat, F. S. Khan and M. Felsberg, Eco: Efficient Convolution Operators for Tracking, Proc. of IEEE Conference on Computer Vision and Pattern Recognition, 4800 (2017).

[21] Hamed Kiani Galoogahi, Ashton Fagg and Simon Lucey, Learning Background-Aware Correlation Filters for Visual Tracking, Proc. of IEEE Conference on Computer Vision and Pattern Recognition, 2017.

[22] M. Danelljan, A. Robinson, F. S. Khan and M. Felsberg, Beyond Correlation Filters: Learning Continuous Convolution Operators for Visual Tracking, Proc. of European Conference on Computer Vision Workshop, 850 (2016).

[23] A. Lukezic, T. Vojir, L. Cehovin Zajc, J. Matas and M. Kristan, Discriminative Correlation Filter with Channel and Spatial Reliability, Proc. of IEEE Conference on Computer Vision and Pattern Recognition, 4800 (2017).

[24] Q. Wang, J. Gao, J. Xing, M. Zhang and W. Hu, DCFNet: Discriminant Correlation Filters Network For Visual Tracking, arXiv preprint,arXiv:1704.04057, 2017.

[25] N Wang, W Zhou, Q Tian, R Hong, M Wang and H Li, Multi-Cue Correlation Filters for Robust Visual Tracking, Proc. of IEEE Conference on Computer Vision and Pattern Recognition, 2018.

[26] E. Gundogdu and A. Alatan, Good Features to Correlate for Visual Tracking, IEEE Transactions on Image Processing, 2526 (2018).

[27] H He, Y Fan, J Zhuang and H Bai, Correlation Filters with Weighted Convolution Responses, Proc. of IEEE International Conference on Computer Vision, 2017.