Zhu and Wang *EURASIP Journal on Advances in Signal Processing* (2019) 2019:41
https://doi.org/10.1186/s13634-019-0639-z

EURASIP Journal on Advances in Signal Processing

RESEARCH Open Access

Robust visual tracking via samples ranking

Heyan Zhu[1*] and Hui Wang[2]

Abstract

In recent years, deep convolutional neural networks (CNNs) have achieved great success in visual tracking. To learn discriminative representations, most of existing methods utilize information of image region category, namely target or background, and/or of target motion among consecutive frames. Although these methods demonstrated to be effective, they ignore the importance of the ranking relationship among samples, which is able to distinguish one positive sample better than another positive one or not. This is especially crucial for visual tracking because there is only one best target candidate among all positive candidates, which tightly bounds the target. In this paper, we propose to take advantage of the ranking relationship among positive samples to learn more discriminative features so as to distinguish closely similar target candidates. In addition, we also propose to make use of the normalized spatial location information to distinguish spatially neighboring candidates. Extensive experiments on challenging image sequences demonstrate the effectiveness of the proposed algorithm against several state-of-the-art methods.

Keywords: Visual tracking, Convolutional neural network, Ranking, Normalized spatial localization

1 Introduction

Visual tracking has been one of the most fundamental topics in computer vision due to its important roles in numerous applications such as surveillance, human-computer interaction, and automatic driving [1–20]. It aims to estimate the states (e.g., location, scale, rotation) of a target in a video after specifying the target in the first frame usually using a rectangle. While significant efforts have been made in the past decades, developing a robust tracking algorithm for complicated scenarios is still a challenging task due to interfering factors like heavy occlusion, pose changes, large scale variations, camera motion, and illumination variations.

In recent years, inspired by feature learning based on sparse coding [21], hierarchical features learned by CNNs have greatly boost the performance of visual tracking methods [22–27]. To learn discriminative representations, most of existing methods utilize information from image (region) category, namely target or background [24, 26, 28–30], and/or from target motion among consecutive frames [31, 32]. Choi et al [24] propose to utilize the category information when learning target/background classification. In [28], Dong and Shen employ the distance relationship among positive, negative, and target template to learning more discriminative features. A feature net, a temporal net, and a spatial net are designed in [31] to extract general feature representation, encode target trajectory, and refine tracking results using local spatial object information, respectively.

Although these methods demonstrated to be effective, they ignore the importance of the ranking relationship among samples, which is able to distinguish whether one positive sample is better than another positive one or not. Different from the image classification task, visual tracking is location-sensitive, which means a good visual tracking CNN model is able to not only tell positive samples from negative ones but also can distinguish the quality of positive samples and hence filter out the one having the largest overlap ratio with the ground truth. As shown in Fig. 1, the left panel shows the cases of most existing classification scores, which can correctly distinguish positive samples from negative ones. However, the order among positive samples cannot be guaranteed. As a result, the best target candidate may not obtain the highest classification score and hence it cannot be filtered out as the tracking result.

*Correspondence: 1980zhuheyan@163.com
[1]School of Opto-electronic Information, Yantai University, 30 Qinquan Road, Yantai, 264005 China
Full list of author information is available at the end of the article

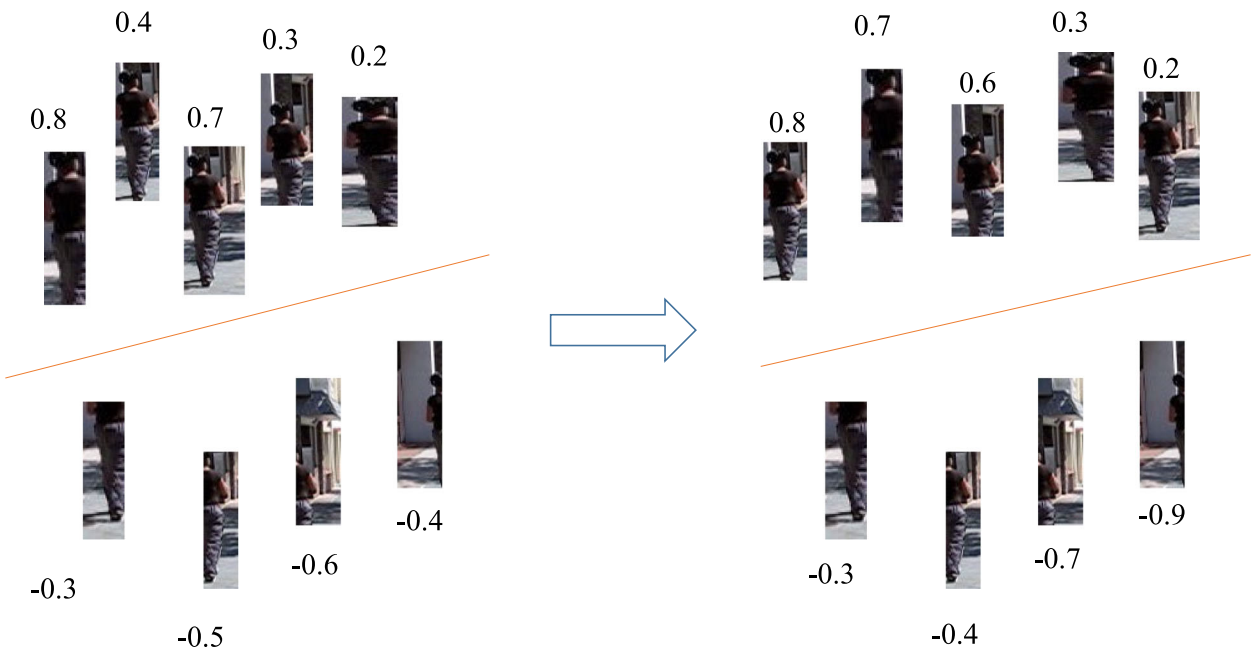

Fig. 1 Most of existing tracking methods only require to tell positive samples from negative ones as shown by the left panel. However, this is not enough for the visual tracking task as it needs to filter out the one having the largest IoU with the ground truth. To overcome this problem, we propose to predict scores that are consistent with the ranking order in term of IoU with the ground truth as shown by the right panel

To address the abovementioned problem, in this paper, we propose to take advantage of the distance ranking relationship among positive samples to learn more discriminative features. We require that the confidence score order should be consistent with the IoU order of samples with ground truth (IoU is the intersection over union of two bounding boxes). With such a constraint, the model not only is able to tell positive samples from negative ones, but also has the ability to assign the highest confidence to the candidate that is most similar to the target template. In Fig. 1, we show an illustration of the expected scoring scheme on the right panel. In addition, we observe that spatially close samples generally have the same CNN features due to resolution reduction after convolution or pooling operations. To overcome this problem, we also propose to make use of the normalized spatial location information to enhance the difference of spatially neighboring candidates. Figure 2 shows that the proposed approach is able to achieve better performance compared to several state-of-the-art tracking methods.

In summary, we make the following contributions in this paper:

1. We propose a tracking method to take the ranking relationship among samples into consideration, which is able to estimate samples' scores with the consistent ranking in terms of the IoU metric with the ground truth.
2. We propose to take advantage of the location information of samples to distinguish them from each other even in the case they are closely positioned.
3. Extensive experimental results on large object tracking datasets show the effectiveness of the proposed tracking methods in comparison with several state-of-the-art tracking methods.

2 Related work

In this section, we briefly review the closely related tracking methods.

Generally, most of existing tracking methods falls into either the non-CNN-based category or the CNN-based category according to whether CNN features are used. The non-CNN-based tracking methods usually employ the sparse coding framework to obtain effective image representations [17, 33–38]. In [17], spatial structure among selected local templates are enhanced to exclude distractors introduced by noisy templates. Lan et al. [33, 35, 36] propose to mine the common and specific patterns in sparse codings so as to discriminate positive and negative examples.

Wang et al. [39] first introduce the deep learning technology into the visual tracking task, where a denoising auto-encoder is employed to learn compact image representations in a self-supervised manner. After that, Hong et al. propose to make use of the gradient back-propagation algorithm to generate a saliency map for the

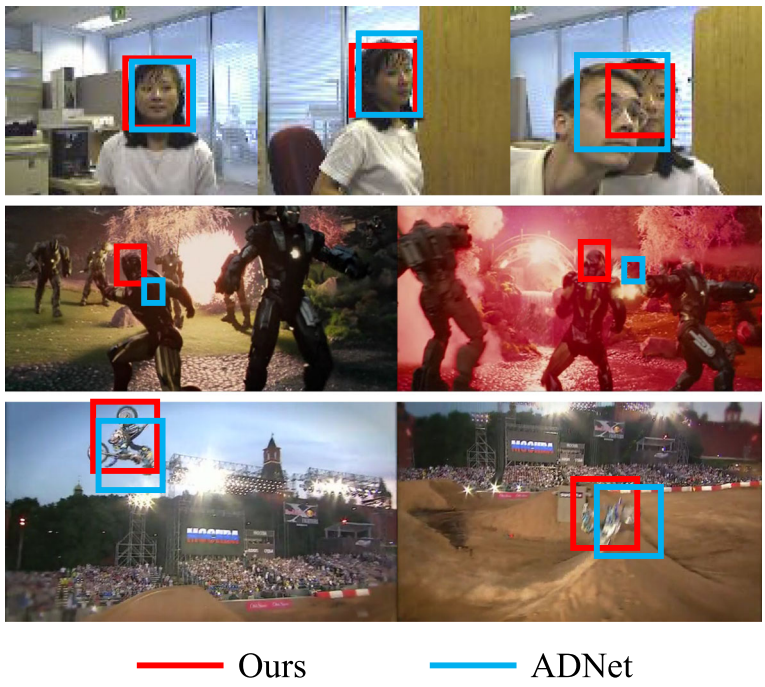

Fig. 2 Comparison of the proposed approach against state-of-the-art tracking methods ADNet on three example sequences challenged by large occlusion (top row), illumination variation (middle row), and fast motion (bottom row). Our approach performs more robustly than the ADNet tracking method

tracking target in order to facilitate to localize the target. These two methods only utilize CNN features extracted from one of the last fully connected (FC) layer. However, the deeper features are rich in semantic information, which benefits the tracker to distinguish target from background. But visual tracking is a location-sensitive task, and deeper features cannot provide spatial details as its low spatial resolution (1×1 resolution for FC features). To overcome this problem, [40] proposes to combine CNN features extracted from both shallow layers and deep layers, which shows to be effective on visual tracking benchmarks. One drawback of [40] is that the combination weights for different features are fixed for all frames and all videos. This is not feasible because different features perform best in different scenarios. To overcome this problem, Qi et al. [22] propose to adaptively generate combination weights via an improved Hedge algorithm. The aforementioned methods use pretrained CNN models for the image classification task. Due to the fundamental difference between these two tasks, directly adopting or simply fine-tuning image classification models limit the performance of CNNs. To better adapt to visual tracking task, a multi-domain CNN is designed in [41] to avoid category ambiguity that one class is the tracking target in one video while being background in another video. Nam and Han [41] also introduce hard negative sample mining and bounding box refinement to further improve tracking performance. Very recently, [42] propose to enhance tracking results via pixel-wise object segmentation. The advantage of segmentation based methods is that rotated minimum bounding rectangle instead of axis-aligned box can be determined.

The other line is to develop real-time CNN-based tracking methods. Tao et al. [43] propose the first real-time CNN-based tracking method. They design a siamese network to learn a similarity function and using ROI pooling to reduce repeated feature calculation. The price is to sacrifice tracking accuracy. Later, Bertinetto et al. [44, 45] propose to implement correlation filter learning within the end-to-end CNN training, which achieves a balance between the tracking accuracy and tracking speed. In [46], Yun et al. propose to determine the target state in a new frame via moving the tracking result in the previous frame in left/right/up/down four directions and zoom in of zoom out the bounding box until an stop action is generated. This method avoids selecting tracking target in hundreds of target candidates and hence improve the tracking speed. Very recently, [25] propose to quickly adapt pretrained CNN models to test image sequences via meta-learning, which usually accomplish adaptation to new videos within five iterations. Li et al. [47] propose to integrate region proposal network into siamese network to address the scale problem of siamese network.

Overall, most of existing CNN-based visual tracking methods just make use of image region category information and/or target motion information. They neglect the ranking relationship among samples, and hence, the positive target candidate with highest confidence may not be the best. To overcome this problem, we propose a tracking method to align the confidence ranking consistent with that of IoU score compared to ground truth.

3 Method

In this section, we first detail the proposed neural network and then we describe how to train the network.

3.1 Architecture

The proposed deep convolutional neural network is equipped with two branches in a siamese architecture as shown in Fig. 3. In each branch, the first three layers are used to learn common representations among all kinds of objects, such as corner points and edges. It can be implemented using pretrained CNN models originally designed for image classification, such as AlexNet [48], VGG [49], and ResNet [50]. Here, we adopt the first three layers of VGGM [51] due to its balance between computational cost and classification accuracy. The next three fully connected layers are used to learn high level embeddings of the input image. It is initialized randomly from a Gaussian distribution. Before classification, we concatenate the image embeddings and its spatial information $(x_i, y_i, w/W, h/H)$, where (x_i, y_i) denote the coordinate of the top-left point of the input image region, w, h denote the width and height of the image region, and W, H denote the width and height of the video frame.

We employ the softmax loss as the supervision for target/background classification:

$$l_{\text{cls}}(x_i, y_i) = -f(x_i)_{y_i} + \log\left(\sum_{j=0,1} e^{f(x_i)_j}\right) \tag{1}$$

where y_i denotes the class label of the input image region x_i, $f(x_i)_{y_i}$ denotes the y_ith element of the network output $f(\cdot)$, $j = 0, 1$ denote the class labels: 0 for background, 1 for target. To constrain the network predicted scores to be consistent with their ranking in terms of IoU with the ground truth, we also adopt the margin ranking loss:

$$l_{\text{rank}}(x_i, x_j) = \max\left(0, m - \left(f(x_i)_1 - f(x_j)_1\right)\right) \tag{2}$$

where x_i, x_j denote two input image regions and m denotes the least margin. If the training sample x_i has a larger IoU with the ground truth than x_j, it should rank before x_j which means its probability being the target $f(x_i)_1$ should be larger than $f(x_j)_1$. The overall loss for a training pair (x_i, y_i, x_j, y_j) is

$$\begin{aligned} L(x_i, y_i, x_j, y_j) =& l_{cls}(x_i, y_i) + l_{cls}(x_j, y_j) \\ &+ l_{\text{rank}}(x_i, x_j) \end{aligned} \tag{3}$$

3.2 Training

The network is trained in the end-to-end scheme using stochastic gradient descent (SGD) with moment 0.9. The training data is sampled according to [41]. In each frame 5500 samples are randomly extracted around the ground truth. The learning rate is fixed to $2e-4$. In each iteration, each mini-batch contains 32 positive and 32 negative samples, which have ≥ 0.7 and ≤ 0.5 IoU with the ground

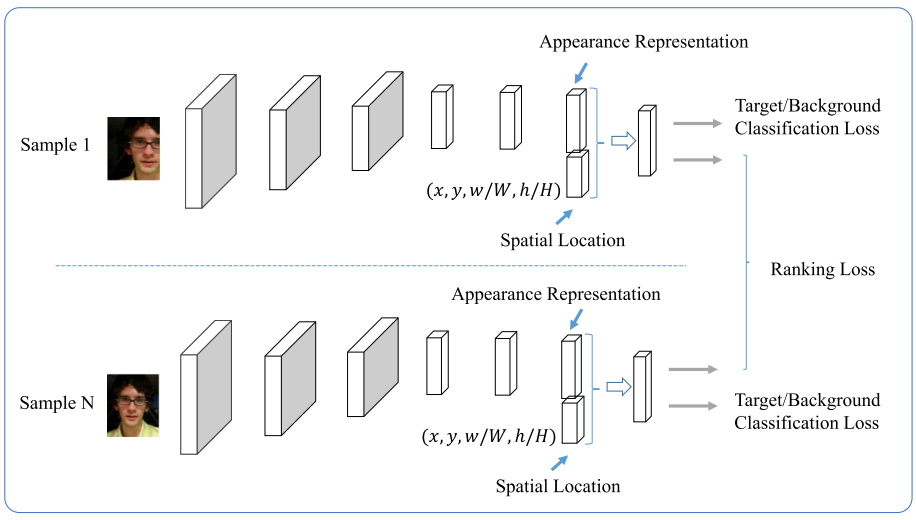

Fig. 3 The main architecture of the proposed neural network. In the training phase, it takes two image regions as input and output the target/background classification scores. Softmax loss and margin ranking loss are employed for training. In the test phase, only one branch is remained

Table 1 AUC score and precision at a threshold of 20 pixels on the OTB100 dataset for the ablation analysis on the ranking loss (denoted by RL) and spatial location feature (denoted by SLF)

	Ours	Ours w/o RL	Ours w/o SLF
AUC	0.668	0.641	0.653
Precision	0.895	0.873	0.882

w/o denotes "without"

truth bounding box, respectively. The network converges at about 200 iterations.

3.3 Tracking

Let x_i denotes the ith target candidate, the tracking result is the one with the largest target confidence:

$$x^* = \arg\max_{i=0,\cdots,N} f(x_i) \tag{4}$$

where N denotes the number of target candidate. According to [41], the model will be updated when the maximum target confidence is less than zero or after a fixed interval (short-term update interval is 20 frames and long-term update interval is 100 frames). The data for model update are sampled in each frame around the tracking result.

4 Experiments

In this section, we first introduce the evaluation protocols. Then, we examine the effectiveness of the proposed tracking method on large scale datasets compared to state-of-the-art tracking methods.

4.1 Evaluation protocols

We adopt the commonly used success plots and precision plots [52] as the main evaluation metrics, which avoid the drawback of using only one threshold to measure the success. The trackers in success plots are ranked in terms of the area under curve (AUC) and are ranked in precision plots in terms of success rate at a threshold of 20 pixels between center points of tracked results and ground truth.

The implementation is based on PyTorch. We sample 300 target candidates in each frame. The model is update every 20 frames or when the largest confidence is negative. The unoptimized code runs at 1 FPS on a machine with an i7-3.4 GHz CPU and a GeForce GTX 1080 GPU. Fine-tuning samples are collected as the video goes, where 50 positive and negative samples are extracted in each tracked frame around the tracking result.

We compare the proposed method with six state-of-the-art tracking approaches including ADNet [46], CFNet [45], HDT [22], MCPF [53], CREST [54], and MetaTracker [25].

4.2 Ablation analysis

In this section, we evaluate the effectiveness of the introduced ranking loss and the spatial location features, respectively. Table 1 presents the tracking performance on the OTB100 dataset in terms of AUC and precision scores. It shows that the tracking performance drops about 2% if the ranking loss is not employed. If the spatial location features are not used, the tacking accuracy drops about 1% in terms of both AUC and precision metrics. These demonstrate the effectiveness of both the ranking loss and the spatial location features.

4.3 Quantitative evaluation

In Fig. 4, we provide the overall performance on the OTB100 dataset. It shows that the proposed tracking method achieves favorable performance compared to

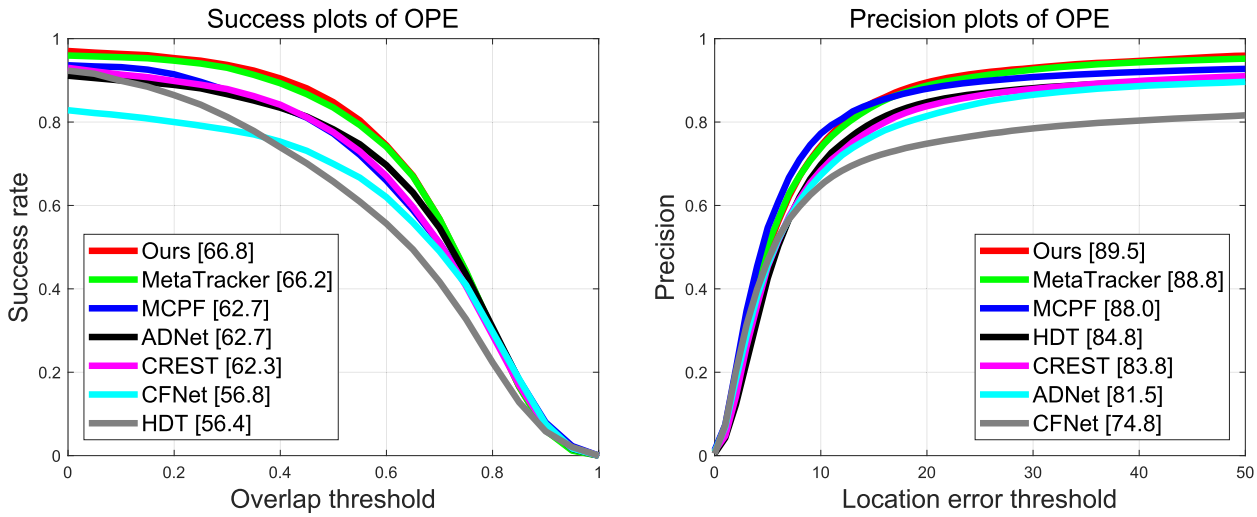

Fig. 4 Tracking results based on 11 attribute in terms of precision plots

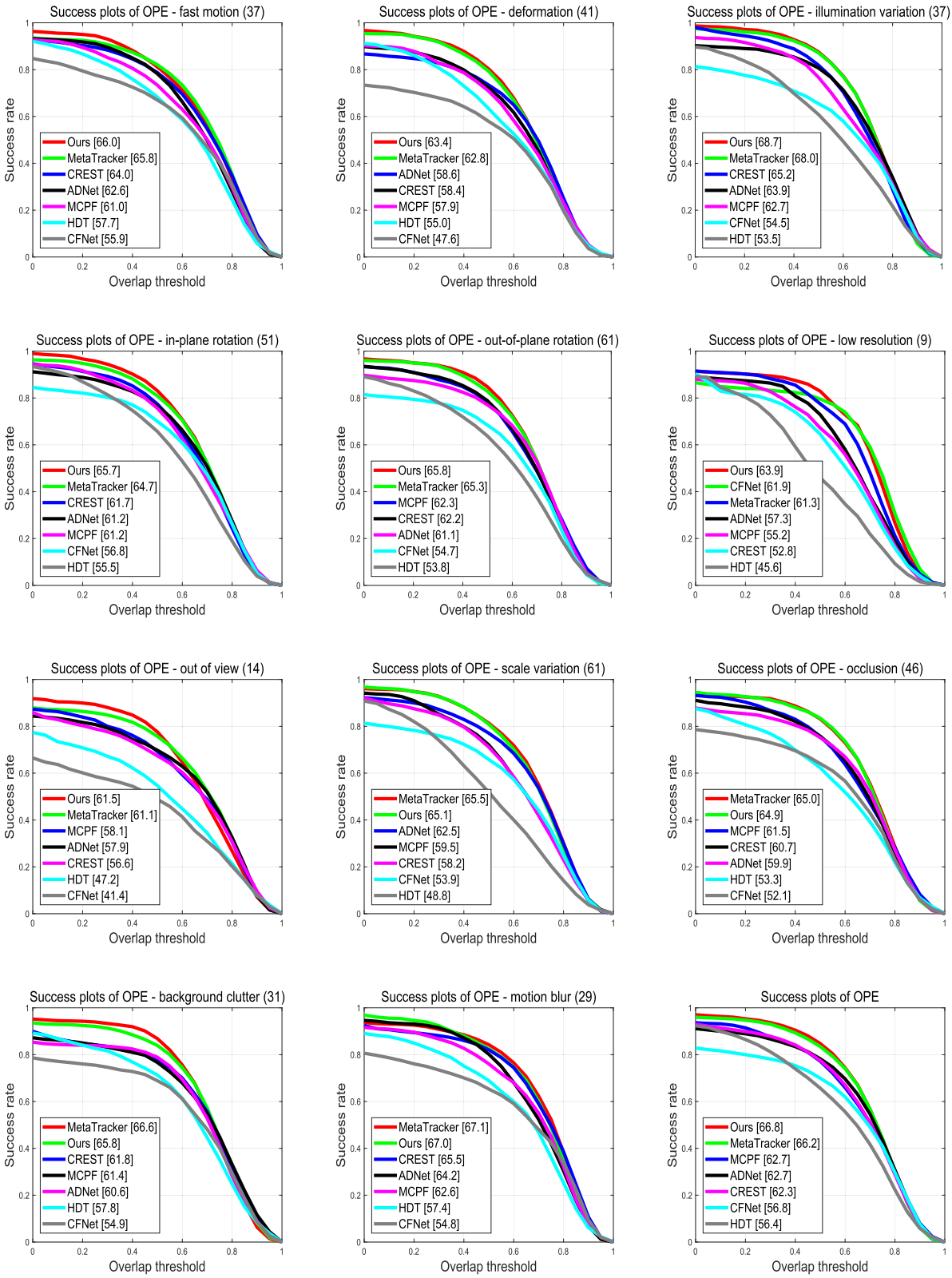

Fig. 5 Tracking results based on 11 attribute challenges in terms of success plots

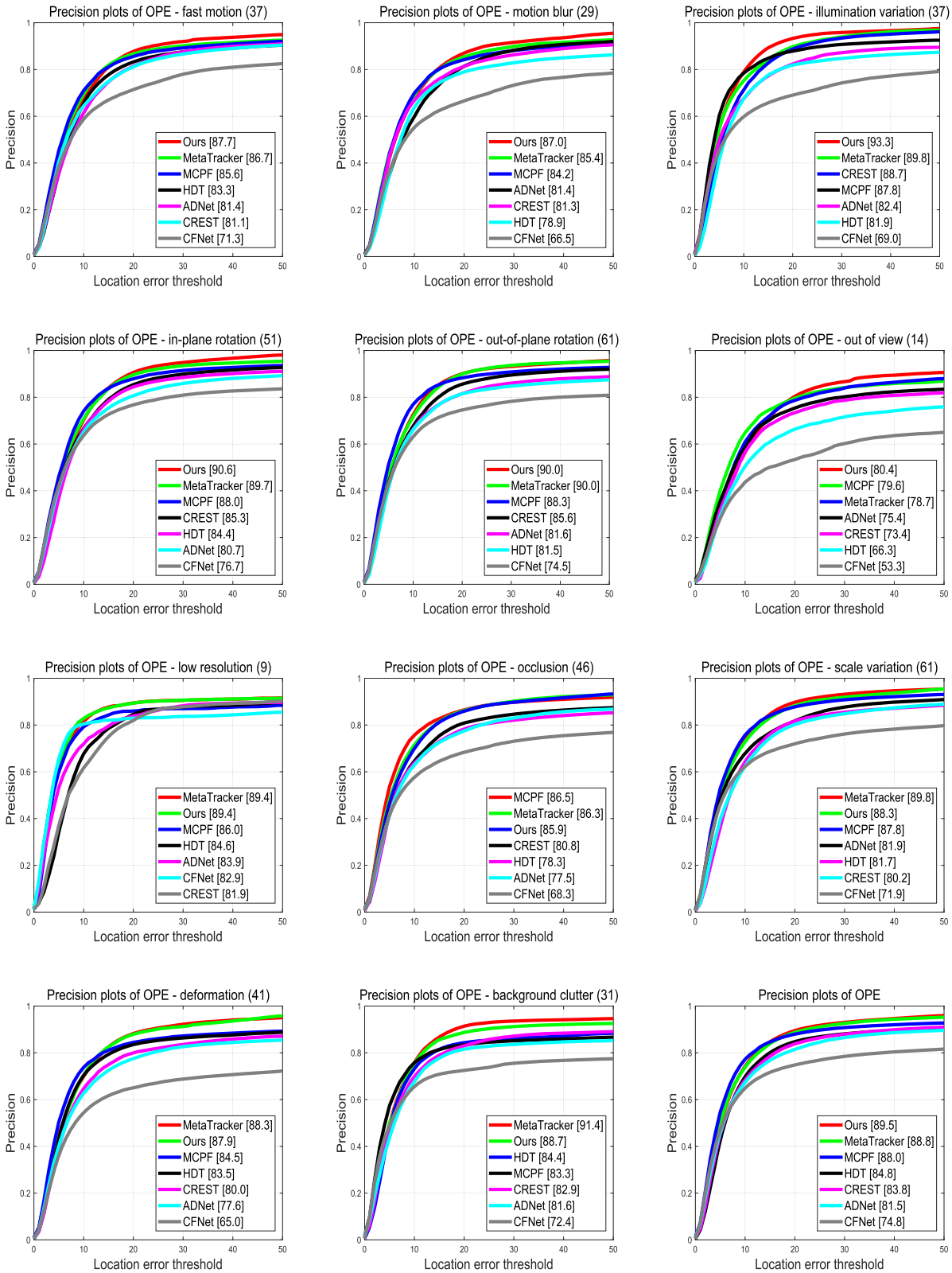

Fig. 6 Tracking results based on 11 attribute in terms of precision plots

state-of-the-art trackers such as MetaTracker, CREST, or ADNet. Specially, our tracking method performs about 4% better than CREST, which takes both appearance information and temporal motion information into consideration. In contrast, the only motion cues utilized in our method happens during the target candidate sampling, which adhere by a Gaussian distribution centered at the last tracking result.

To further evaluate the proposed method, we also conduct the attribute-based performance evaluation on the OTB100 dataset in terms of success plots and precision plots. The results are presented in Figs. 5 and 6. The results in Fig. 5 show that our tracking approach performs best on 7 out of 11 attributes, which includes faster motion, deformation, illumination variation, in-plane and out-of-plane rotations, low resolution, and out-of-view. In the of tracking precision, similar performance can be observed in Fig. 6.

For completeness, we also present the tracking results on the VOT 2016 dataset [55] in Table 2. The results show that our method performs favorably with an EAO of 0.320 compared against state-of-the-art tracking methods, such as CCOT [56] and SiamFC [44].

4.4 Qualitative evaluation

In Fig. 7, we present sample tracking results of the evaluated methods on both OTB100 [52] and UAVDT [57] datasets. For presentation clarity, only results of the top 7 performing trackers are shown.

Occlusion. The target in the UAV-Traffic sequence *S0103* undergoes occlusions caused by trees. Only the proposed method and MDNet are able to locate the target, while other trackers such as CREST and ADNet falsely locate on background out of the view as shown in frames 75 and 152. Similar performance happens in the OTB100 video *Girl2*, where the target girl gradually occluded by a man walking with a bicycle. Such success can be attributed to powerful deep CNN features regularized by both the classification loss and the ranking loss, as well as the spatial location loss.

Camera motion. The target object in the *BlurBody* image sequence is blurred due to camera shaking. For such an image sequence, the proposed tracking method and CCOT methods are still able to precisely locate the target. In contrast, other trackers such as ADNet and MCPF locate the target with much background as shown in frame 236. The effectiveness of the proposed algorithm benefits from the camera motion branch as evaluated in Table 1. In the UAV-Traffic image sequence *S0602*, the camera hovers over the crossroads, which leads to huge appearance variance of the target (the blue bus). The bounding boxes show that the proposed approach tracks the target more accurately than others during the hover, while ADNet falsely locates on the road and MCPF fails to identify the target from background as shown in frame 291.

Object motion. The target in the OTB100 sequence *DragonBaby* hits his opponent using the turn-around kick. As shown by the bounding boxes, both the proposed tracking method and MCPF methods are able to locate the target accurately in such a procedure, but other trackers lose the target, such as ADNet. With reference to ablation evaluations in Table 1, both the spatial location features and the ranking loss helps to capture discriminative information in such a scene.

Scale. In the UAV-Traffic sequence *S1701*, the size of the target bus changes intensively and the observation view changes from bird-view to side-view, which cause large appearance variations. In such a challenging scene, the proposed ANT method locates the target more accurately than others such as MDNet and HDT, as shown in frames 200 and 324. The performance gain of the proposed algorithm can be mainly attributed to the ranking loss, classification loss, and spatial location features, which learns robust representations under various challenges.

Illumination. The target in *Ironman* has drastic movements in a dark night with large illumination changes in the backgrounds. In such a poor lighting condition, the proposed algorithm accurately locates the target in most frames while other trackers drift far away as shown in frames 129 and 165. Similar performance can be observed in the UAV-Traffic sequences *S1301* and *S1303*. As evaluated in Table 1, the illumination branch contributes most in such situations.

Table 2 Tracking results on the VOT2016 dataset in terms of expected average overlap (EAO), accuracy rank (A), and robustness rank (R)

Trackers	Ours	CCOT	Staple	MDNet	EBT	SiamFC
EAO	0.320	0.331	0.295	0.257	0.291	0.277
A	1.47	1.98	1.87	1.72	3.62	1.30
R	2.10	1.95	3.23	2.8	2.13	3.17

5 Conclusion

In this paper, we propose a novel tracking method, which takes advantages of the ranking relationship among positive samples to learn more discriminative features so as to distinguish closely similar target candidates. To achieve this goal, we propose a sample ranking method to select discriminative samples. In addition, we also propose a spatial normalization method to make use of the normalized spatial location information to distinguish spatially neighboring candidates. Extensive experiments on challenging image sequences demonstrate the effectiveness of the proposed algorithm against several state-of-the-art methods.

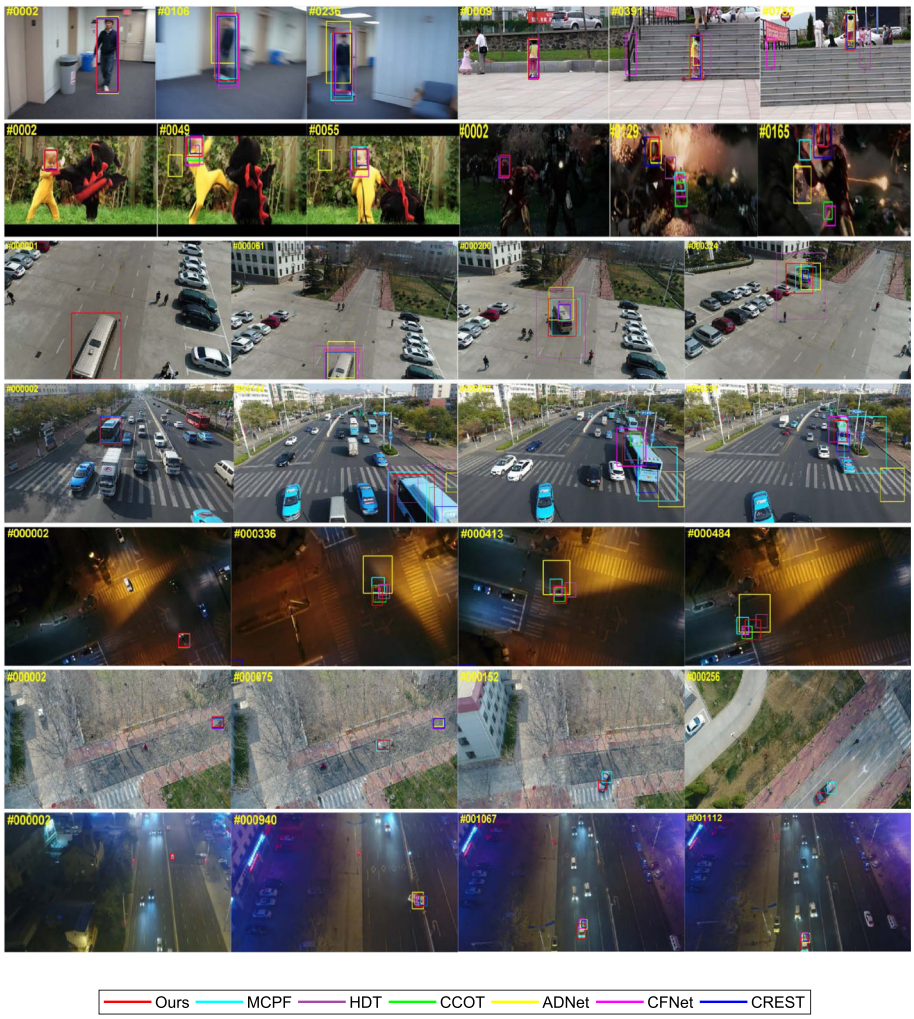

Fig. 7 Several samples of tracking results on both the OTB100 dataset and the UAV dataset (from top to bottom, left to right: BlurBody, Girl2, DragonBaby, Ironman, S1701, S0602, S1301, S0103, and S1303)

Acknowledgements
This work acknowledged the Editor and anonymous Reviewers.

Authors' contributions
Quanling and Xuefeng conceived the method and developed the algorithm. Weigang conceived the method, oversaw the project, and wrote the first draft. Lei assembled formulations and drafted the manuscript. All authors read and approved the final manuscript.

Funding
This work is partly supported by the National Natural Science Foundation of China (61703143).

Availability of data and materials
Data and source code are available from the corresponding author upon request.

Ethics approval and consent to participate
Not applicable.

Consent for publication
All the data (including individual details, images, or videos) in the paper are all from public datasets.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]School of Opto-electronic Information, Yantai University, 30 Qinquan Road, Yantai, 264005 China. [2]Lab, CNCERT/CC, Beijing, 100029 China.

Received: 7 April 2019 Accepted: 23 August 2019
Published online: 09 September 2019

References

1. S. Zhang, H. Yao, X. Sun, S. Liu, Robust visual tracking using an effective appearance model based on sparse coding. ACM Trans. Intell. Syst. Technol. **3**(3), 1–18 (2012)
2. S. Zhang, H. Yao, X. Sun, X. Lu, Sparse coding based visual tracking: review and experimental comparison. Pattern Recogn. **46**(7), 1772–1788 (2013)
3. S. Zhang, H. Yao, H. Zhou, X. Sun, S. Liu, Robust visual tracking based on online learning sparse representation. Neurocomputing. **100**(1), 31–40 (2013)
4. S. Zhang, H. Zhou, H. Yao, Y. Zhang, K. Wang, J. Zhang, Adaptive normal hedge for robust visual tracking. Signal Process. **110**, 132–142 (2015)

5. S. Yi, Z. He, X. You, Y. Cheung, Single object tracking via robust combination of particle filter and sparse representation. Signal Process. **110**, 178–187 (2015). https://doi.org/10.1016/j.sigpro.2014.09.020
6. S. Zhang, S. Kasiviswanathan, P. C. Yuen, M. Harandi, in *Twenty-ninth AAA I Conference on Artificial Intelligence*. Online dictionary learning on symmetric positive definite manifolds with vision applications (AAAI Press, 2015), pp. 3165–3173
7. L. Zhang, W. Wu, T. Chen, N. Strobel, D. Comaniciu, Robust object tracking using semi-supervised appearance dictionary learning. Pattern Recogn. Lett. **62**, 17–23 (2015)
8. S. Zhang, H. Zhou, F. Jiang, X. Li, Robust visual tracking using structurally random projection and weighted least squares. IEEE Trans. Circ. Syst. Video Technol. **25**(11), 1749–1760 (2015)
9. X. Ma, Q. Liu, Z. He, X. Zhang, W. Chen, Visual tracking via exemplar regression model. Knowl.-Based Syst. **106**, 26–37 (2016). https://doi.org/10.1016/j.knosys.2016.05.028
10. Z. He, S. Yi, Y. Cheung, X. You, Y. Y. Tang, Robust object tracking via key patch sparse representation. IEEE Trans. Cybern. **47**(2), 354–364 (2017). https://doi.org/10.1109/TCYB.2016.2514714
11. Q. Liu, X. Lu, Z. He, C. Zhang, W. Chen, Deep convolutional neural networks for thermal infrared object tracking. Knowl.-Based Syst. **134**, 189–198 (2017). https://doi.org/10.1016/j.knosys.2017.07.032
12. S. Zhang, X. Lan, Y. Qi, P. C. Yuen, Robust visual tracking via basis matching. IEEE Trans. Circ. Syst. Video Technol. **27**(3), 421–430 (2017)
13. Y. Yao, X. Wu, L. Zhang, S. Shan, W. Zuo, in *Proceedings of the European Conference on Computer Vision (ECCV)*. Joint representation and truncated inference learning for correlation filter based tracking. Lecture Notes in Computer Science, vol. 11213 (Springer, Cham, 2018), pp. 560–575
14. S. Zhang, X. Lan, H. Yao, H. Zhou, D. Tao, X. Li, A biologically inspired appearance model for robust visual tracking. IEEE Trans. Neural Netw. Learn. Syst. **28**(10), 2357–2370 (2017)
15. X. Lu, Y. Liang, Z. He, Discriminative collaborative representation-based tracking. J. Electron. Imaging. **27**(05), 053040 (2018). https://doi.org/10.1117/1.JEI.27.5.053040
16. S. Zhang, Y. Qi, F. Jiang, X. Lan, P. C. Yuen, H. Zhou, Point-to-set distance metric learning on deep representations for visual tracking. IEEE Trans. Intell. Transp. Syst. **19**(1), 187–198 (2018)
17. Y. Qi, L. Qin, J. Zhang, S. Zhang, Q. Huang, M.-H. Yang, Structure-aware local sparse coding for visual tracking. IEEE Trans. Image Proc. **27**(8), 3857–3869 (2018)
18. Y. Qi, S. Zhang, L. Qin, Q. Huang, H. Yao, J. Lim, M. Yang, Hedging deep features for visual tracking. IEEE Trans. Pattern Anal. Mach. Intell. **41**(5), 1116–1130 (2019)
19. X. Li, Q. Liu, N. Fan, Z. He, H. Wang, Hierarchical spatial-aware siamese network for thermal infrared object tracking. Knowl.-Based Syst. **166**, 71–81 (2019). https://doi.org/10.1016/j.knosys.2018.12.011
20. Y. Qi, H. Yao, X. Sun, X. Sun, Y. Zhang, Q. Huang, in *2014 IEEE International Conference on Image Processing (ICIP)*. Structure-aware multi-object discovery for weakly supervised tracking (Paris, 2014), pp. 466–470. https://doi.org/10.1109/ICIP.2014.7025093
21. P. Wilf, S. Zhang, S. Chikkerur, S. A. Little, S. L. Wing, T. Serre, Computer vision cracks the leaf code. Proc. Nat. Acad. Sci. U.S.A. **113**(12), 3305–3310 (2016)
22. Y. Qi, S. Zhang, L. Qin, H. Yao, Q. Huang, J. Lim, M. Yang, in *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016*. Hedged Deep Tracking (Las Vegas, 2016), pp. 4303–4311
23. M. Danelljan, G. Bhat, F. S. Khan, M. Felsberg, in *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*. ECO: efficient convolution operators for tracking, vol. 1 (Honolulu, 2017), pp. 6931–6939. https://doi.org/10.1109/CVPR.2017.733
24. J. Choi, H. J. Chang, T. Fischer, S. Yun, K. Lee, J. Jeong, Y. Demiris, J. Y. Choi, *Context-aware deep feature compression for high-speed visual tracking* (2018), pp. 479–488. https://doi.org/10.1109/CVPR.2018.00057
25. E. Park, A. C. Berg, in *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part III*. Meta-tracker: fast and robust online adaptation for visual object trackers (2018), pp. 587–604. arXiv:1801.03049
26. I. Jung, J. Son, M. Baek, B. Han, in *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part IV*. Real-time mdnet. Lecture Notes in Computer Science, vol. 11208 (Springer, Cham, 2018), pp. 89–104
27. Y. Qi, S. Zhang, L. Qin, H. Yao, Q. Huang, J. Lim, M. Yang, in *Proceedings of the Thirty-third AAAI Conference on Artificial Intelligence*. Learning attribute-specific representations for visual tracking, vol. 33 (2019), pp. 8835–8842. https://doi.org/10.1609/aaai.v33i01.33018835
28. X. Dong, J. Shen, in *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XIII*. Triplet loss in siamese network for object tracking (2018), pp. 472–488. https://doi.org/10.1007/978-3-030-01261-8_28
29. T. Yang, A. B. Chan, in *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part IX*. Learning dynamic memory networks for object tracking. Lecture Notes in Computer Science, vol. 11213 (Springer, Cham, 2018), pp. 153–169
30. Y. Qi, L. Qin, S. Zhang, Q. Huang, H. Yao, Robust visual tracking via scale-and-state-awareness. Neurocomputing. **329**, 75–85 (2019)
31. Z. Teng, J. Xing, Q. Wang, C. Lang, S. Feng, Y. Jin, in *2017 IEEE International Conference on Computer Vision (ICCV), Venice, 2017*. Robust Object Tracking Based on Temporal and Spatial Deep Networks (2017), pp. 1153–1162. https://doi.org/10.1109/ICCV.2017.130
32. F. Li, C. Tian, W. Zuo, L. Zhang, M. Yang, in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Learning Spatial-Temporal Regularized Correlation Filters for Visual Tracking (Salt Lake City, 2018), pp. 4904–4913. https://doi.org/10.1109/CVPR.2018.00515
33. X. Lan, S. Zhang, P. C. Yuen, R. Chellappa, Learning common and feature-specific patterns: a novel multiple-sparse-representation-based tracker. IEEE Trans. Image Proc. **27**(4), 2022–2037 (2018)
34. X. Lan, A. J. Ma, P. C. Yuen, R. Chellappa, Joint sparse representation and robust feature-level fusion for multi-cue visual tracking. IEEE Trans. Image Proc. **24**(12), 5826–5841 (2015)
35. X. Lan, A. J. Ma, P. C. Yuen, in *2014 IEEE Conference on Computer Vision and Pattern Recognition*. Multi-cue visual tracking using robust feature-level fusion based on joint sparse representation (Columbus, 2014), pp. 1194–1201. https://doi.org/10.1109/CVPR.2014.156
36. X. Lan, S. Zhang, P. C. Yuen, in *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016*. Robust joint discriminative feature learning for visual tracking (AAAI Press, New York, 2016), pp. 3403–3410
37. X. Lan, M. Ye, R. Shao, B. Zhong, P. C. Yuen, H. Zhou, Learning modality-consistency feature templates: a robust RGB-infrared tracking system. Trans. Ind. Electron. IEEE. **66**(12), 9887–9897 (2019). https://doi.org/10.1109/TIE.2019.2898618
38. X. Lan, M. Ye, S. Zhang, H. Zhou, P. C. Yuen, Modality-correlation-aware sparse representation for RGB-infrared object tracking. Pattern Recog. Lett. (2018). https://doi.org/https://doi.org/10.1016/j.patrec.2018.10.002
39. N. Wang, D. Yeung, in *NIPS'13 Proceedings of the 26th International Conference on Neural Information Processing Systems, Curran Associates Inc., USA*. Learning a deep compact image representation for visual tracking, vol. 1 (2013), pp. 809–817
40. C. Ma, J. Huang, X. Yang, M. Yang, in *2015 IEEE International Conference on Computer Vision (ICCV)*. Hierarchical Convolutional Features for Visual Tracking (Santiago, 2015), pp. 3074–3082. https://doi.org/10.1109/ICCV.2015.352
41. H. Nam, B. Han, in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Learning multi-domain convolutional neural networks for visual tracking (Las Vegas, 2016), pp. 4293–4302. https://doi.org/10.1109/CVPR.2016.465
42. Q. Wang, L. Zhang, L. Bertinetto, W. Hu, P. H. S. Torr, Fast online object tracking and segmentation: a unifying approach (2018). https://doi.org/http://arxiv.org/abs/1812.05050. arXiv preprint
43. R. Tao, E. Gavves, A. W. M. Smeulders, in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Siamese Instance Search for Tracking (Las Vegas, 2016), pp. 1420–1429. https://doi.org/10.1109/CVPR.2016.158
44. L. Bertinetto, J. Valmadre, J. F. Henriques, A. Vedaldi, P. H. S. Torr, in *Computer Vision - ECCV 2016 Workshops. ECCV 2016. Lecture Notes in Computer Science, vol. 9914*, ed. by G. Hua, H. Jégou. Fully-convolutional siamese networks for object tracking (Springer, Cham, 2016), pp. 850–865
45. J. Valmadre, L. Bertinetto, J. F. Henriques, A. Vedaldi, P. H. S. Torr, in *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. End-to-end representation learning for correlation filter based tracking, vol. 1 (2017), pp. 5000–5008
46. S. Yun, J. Choi, Y. Yoo, K. Yun, J. Y. Choi, in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Action-decision networks

for visual tracking with deep reinforcement learning (Honolulu, 2017), pp. 1349–1358. https://doi.org/10.1109/CVPR.2017.148
47. B. Li, J. Yan, W. Wu, Z. Zhu, X. Hu, in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. High performance visual tracking with siamese region proposal network (Salt Lake City, 2018), pp. 8971–8980. https://doi.org/10.1109/CVPR.2018.00935
48. A. Krizhevsky, I. Sutskever, G. E. Hinton, in *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a Meeting Held December 3-6, 2012, Lake Tahoe, Nevada, United States*. ImageNet Classification with Deep Convolutional Neural Networks (NIPS. Curran Associates Inc, Nevada, 2012), pp. 1106–1114
49. K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition (2014). arXiv preprint arXiv:1409.1556
50. K. He, X. Zhang, S. Ren, J. Sun, in *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. Deep residual learning for image recognition (2016), pp. 770–778
51. K. Chatfield, K. Simonyan, A. Vedaldi, A. Zisserman, Return of the devil in the details: delving deep into convolutional nets, CoRR (2014). http://arxiv.org/abs/1405.3531. arXiv preprint
52. Y. Wu, J. Lim, M. Yang, Object tracking benchmark. IEEE Trans. Pattern Anal. Mach. Intell. **37**(9), 1834–1848 (2015)
53. T. Zhang, C. Xu, M. Yang, T. Zhang, C. Xu, M. Yang, in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Multi-task correlation particle filter for robust object tracking (Honolulu, 2017), pp. 4819–4827. https://doi.org/10.1109/CVPR.2017.512
54. Y. Song, C. Ma, L. Gong, J. Zhang, R. W. H. Lau, M. Yang, in *2017 IEEE International Conference on Computer Vision (ICCV)*. CREST: convolutional residual learning for visual tracking (Venice, 2017), pp. 2574–2583. https://doi.org/10.1109/ICCV.2017.279
55. M. Kristan, A. Leonardis, J. Matas, M. Felsberg, R. P. Pflugfelder, L. Cehovin, *et al*, in *Computer Workshops. in Computer Vision - ECCV 2016 Workshops - Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part II*, ed. by G. Hua, H. Jégou. The visual object tracking VOT 2016 challenge results, vol. 9914 (Springer, Cham, 2016), pp. 777–823
56. M. Danelljan, A. Robinson, F. S. Khan, M. Felsberg, in *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V*. Beyond correlation filters: learning continuous convolution operators for visual tracking. Lecture Notes in Computer Science, vol. 9909 (Springer, Cham, 2016), pp. 472–488
57. D. Du, Y. Qi, H. Yu, Y. Yang, K. Duan, G. Li, W. Zhang, Q. Huang, Q. Tian, in *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part X*. The unmanned aerial vehicle benchmark: object detection and tracking. Lecture Notes in Computer Science, vol. 11214 (Springer, Cham, 2018), pp. 375–391

Publisher's Note