



Hierarchical Representations with Discriminative Meta-filters in Dual Path Network for Tracking

Fei Xie¹, Ning Wang², Yuncong Yao¹, Wankou Yang¹(✉), Kaihua Zhang²,
and Bo Liu³

¹ School of Automation, Southeast University, Nanjing 210096, China
{220191672,wkyang}@seu.edu.cn

² School of Automation, Nanjing University of Information Science and Technology,
Nanjing 210044, China

³ JD.com, Beijing, China

Abstract. In visual tracking task, accuracy and robustness are critical issues for achieving remarkable performance. In this paper, we propose a novel dual path network with discriminative meta-filters and hierarchical representations to solve these issues. We first design geometrically sensitivity pathway (GESP) and geographical sensitivity pathway (GASP) as two subtasks for target classification and scale estimation. GASP mainly includes powerful discriminative meta-filters to find coarse location of target and GESP can refine region of interests online while adapt the appearance model to the target swiftly. Then, a dual path network is developed in an online and offline framework. Specifically, meta-filters are trained offline in order to gain meta-knowledge of similar tracking scenes. Finally, we present three suggestions on deigning modern tracker. Extensive experiments on VOT2018 datasets verify the superior performance of proposed method compared with other state-of-the-arts, achieving expected average overlap (EAO) of 0.467.

Keywords: Object tracking · Meta-filter · Dual path network · Online learning · Hierarchical deep features

1 Introduction

Generic visual tracking is a crucial task in computer vision aiming at locating the specific continuous object in video. Limited information, usually the first

This work is supported in part by National Major Project of China for New Generation of AI (No. 2018AAA0100400), in part by the Natural Science Foundation of China under Grant nos. 61773117, 61876088, the Primary Research & Development Plan of Jiangsu Province - Industry Prospects and Common Key Technologies under Grant No. BE2017157.

F. Xie—He is currently working toward the Master degree in the School of Automation, Southeast University.

annotation is provided during visual object tracking. One unique characteristic of generic object tracking is that no prior knowledge (e.g., the object class) about the object, as well as its surrounding environment, is allowed [1]. The quality of localization and scale estimation of target are the most influential factor of the performance.

Recently, localization and scale estimation tend to be two subtasks of the tracking problem [2]. Before the deep learning methods, trackers based on Discriminative Correlation Filter (DCF) [3, 4] framework took dominant positions in tracking method. Traditional correlation trackers suffer from inefficiency and low accuracy due to its inherent flaw. It is natural that the deep learning ways are applied to other computer vision tasks, such as object detection, semantic segmentation and visual tracking. SiamFC [5], introduces siamese learning paradigm into visual object tracking, though it employs brutal multi-scale test which is inaccurate and inefficiency [2]. Then, SiamRPN tracker family [6–8] perform an accurate and efficient target scale estimation by introducing the Region Proposal Network (RPN) [9]. However, the pre-defined anchor settings not only introduce ambiguous similarity, but also demand huge prior-knowledge about target. SiamFC++ [10] adopts the anchor-free regression and classification style based on Siamese learning paradigm, it still heavily rely on the sufficient prior-knowledge about target. Motivated by the aforementioned analysis, we propose three suggestions on designing modern visual object trackers:

- **Balance between online learning and offline training:** The breakthroughs on object detection provide a better way to replace multi-scale estimation in object tracking. For example, RPN [9] structure achieves astonishing accuracy in SiamRPN [6]. Because siamese formulation does not provide a powerful discriminative model, we highly recommend that online learning needs to be well-designed. The Correlation based trackers [3, 4, 11, 12] are able to tackle with online model updation. However, the problems of model drift and insufficient training of online model result in low accuracy.
- **Fully utilization of Multi-level deep convolutional features:** Deep model should be trained for robustness, while the shallow model should emphasize accurate target localization [13]. We highly recommend that deep and shallow models should be emphasized equally in order to have better robustness performance. Even though the high quality training data is crucial for the success of end-to-end representation [7], we argue that models designed for both deep and shallow features can reduce the burden of offline training.
- **Online searching strategies are highly recommended in scale estimation branch:** Both the RPN structure from Faster-RCNN [9] or one-stage anchor-free detection from FCOS [14] output the coordinates of target directly without online searching strategy. We strongly consist that it cannot tackle with severe appearance deformation and complex scenes. In our work, we choose the IoU-Net [15] prediction proposed by Atom [2] as our scale estimation branch. It can perform online searching strategy when the coarse location of target is determined.

2 Related Work

Generic object tracking can be divided into two frameworks: Tracking framework and detection framework. Generally, tracking framework trackers are mainly based on correlation filters. MOSSE [4] proposes a CF tracker by learning a minimum output sum of squared error for target appearance and calculate in Fourier domain. KCF [3] adopts ridge regression and circulant matrix to facilitate the speed of calculation in Fourier domain. C-COT [16] converts feature maps of different resolutions into a continuous spatial domain to achieve better accuracy. The subsequent ECO [17] has better efficiency by removing the redundant correlation filters.

ATOM [2] tracker adopts IoU-Net [15] and online learning to classify the target and estimate the scale. Online learning and offline training are combined together. ATOM achieves better robustness performance than Siamese-based trackers. However, it still lack of multi-level deep convolutional features fusion and its online learning is totally independent of offline training which can be further improved. DiMP [18] combines online training and offline training together.

SiamRPN and its succeeding works [3, 4, 11, 12] modifies a Region Proposal Network after a siamese network. They have direct bounding box regression ability thanks to extensive offline training. However, the robustness still suffers from the weak discriminative ability of siameses-based detection networks. The pre-defined anchors of Region Proposal Network (RPN) [9] also need to be well-designed. Even though the SiamFC++ [10] adopts an anchor-free style for bounding box regression, its performance still heavily rely on extensive offline training and robustness cannot be improved as much as accuracy.

3 Proposed Method

Two meta-filters in Geographical Sensitivity Pathway (GASP) are trained to have more discriminative power between foreground and background. The geometrically sensitivity pathway (GESP) focus more on the appearance model of the object in order to estimate the scale accurately.

3.1 Dual Path Network

The whole pipeline of our tracker consists of two meta-filters and a Box Fast Adaption Module. Hierarchical feature representations are used for two meta-filters in order to achieve better performance on localizations. Similar to the object segmentation in [19], the Box Fast Adaption Module can have accurate object outline estimation after the localization process (Fig. 1).

3.2 Multi-hierarchical Independent Discriminative Filters in Online Learning

Inspired by discriminative correlation filter (DCF) approaches, we formulate our learning objective based on L^2 classification error. Each sample x_k contains D

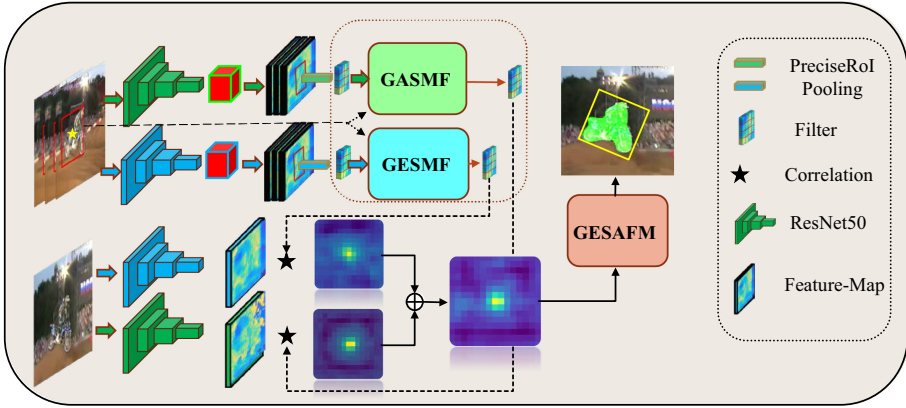


Fig. 1. Pipeline of Dual Pathway Network. GASMF stands for meta-filters in Geographical Sensitivity Pathway. GESMF stands for meta-filters in Geometrically Sensitivity Pathway. GESAFM is Appearance Fast Adaption Module in Geometrically Sensitivity Pathway

feature channels $x_j^1, x_j^2, \dots, x_j^D$, extracted from the same image patch, where k is the index of the samples. Assume that $f = \{f_d\}_{d=1:D}$ is a set of D channel features. The correlation filters algorithm can be formulated as:

$$\arg \min_f \sum_{k=1}^K \|\phi(x_k, f) - y_k\|_{L^2}^2 + \lambda \sum_{d=1}^D \|f\|_{L^2}^2 \quad (1)$$

where x_k is the cyclic shift sample of the x_k and y_k is the Gaussian response label. The optimization problem in Eq. (1) can be solved efficiently in the Fourier domain.

$$\phi(x_k, f) = \sum_{d=1}^D f_d * x_k^d \quad (2)$$

In our work, we try to combine the online optimization with offline training, thus we approximate the loss with a quadratic function and optimize it by backward propagation instead of Fast Fourier Transform (FFT).

In this section, the discriminative learning loss is described in details. The input to our model predictor D consists of a training set $S_{\text{train}} = \{(x_j)\}_{j=1}^n$ of deep feature maps $x_j \in \mathcal{X}$ generated by the backbone network F . During online tracking, correlation filter is optimized to generate a target model $f = D(S_{\text{train}})$. The model f is defined as the filter weights of a convolutional layer. The maximum value of the model output should localize the center of target.

$$L(f) = \frac{1}{|S_{\text{train}}|} \sum_{(x) \in S_{\text{wain}}} \|r(x * f, c)\|^2 + \|\lambda f\|^2 \quad (3)$$

Here, $*$ denotes convolution and λ is a regularization factor. The function $r(s, c)$ computes the residual at every spatial location based on the target confidence scores $s = x * f$ and the ground-truth target center coordinate c . In Eq. (1), $r(s, c) = s - y_c$, traditional correlation filter trackers optimize the residuals between response and the Gaussian target scores. Thus, the difference of target and distractor response usually represents the discriminative ability of the correlation filters. However, during online tracking, background noise and distractors are far more abundant than our target resulting in imbalance of the positive and negative samples.

In order to learn a more discriminative filter, it is common to have a weight matrix in the learning loss. In our work, We employ a hinge-like loss in r , clipping the scores at zero as $\max(0, s)$ in the background region. Thus, the filter is more focus on the hard negative distractors instead of easy negative samples. We believe that it could contribute to a more discriminative filter and efficiency online optimization.

$$r(s, c) = v_c \cdot (m_c s + (1 - m_c) \max(0, s) - y_c) \quad (4)$$

The mask m_c modifies the spatial weight of scores, having values in the interval $m_c(t) \in [0, 1]$ at each spatial location $t \in \mathbb{R}^2$.

In our work, we use convolutional layers D to generate the filter $f = D(S_{\text{train}})$ by implicitly minimizing the error (3).

$$f^{(i+1)} = f^{(i)} - \alpha \nabla L(f^{(i)}) \quad (5)$$

Instead of minimizing the error (3) in Fourier domain, we approximate the error with a quadratic function and directly employ gradient descent optimization using a step length α .

$$L(f) \approx \tilde{L}(f) = \frac{1}{2} (f - f^{(i)})^T Q^{(i)} (f - f^{(i)}) + (f - f^{(i)})^T \nabla L(f^{(i)}) + L(f^{(i)}) \quad (6)$$

Here, the filter variables f and $f^{(i)}$ are seen as vectors and $Q^{(i)}$ is positive definite square matrix. The steepest descent is adopted in order to achieve a fast convergence performance. By solving $\frac{d}{d\alpha} \tilde{L}(f^{(i)} - \alpha \nabla L(f^{(i)})) = 0$, we could find the step length α .

$$\alpha = \frac{\nabla L(f^{(i)})^T \nabla L(f^{(i)})}{\nabla L(f^{(i)})^T Q^{(i)} \nabla L(f^{(i)})} \quad (7)$$

In this work, We set $Q^{(i)} = (J^{(i)})^T J^{(i)}$, where $J^{(i)}$ is the Jacobian of the residuals at $f^{(i)}$. This design of positive definite square matrix $Q^{(i)}$ involves with second-order gradient descent of residuals at $f^{(i)}$ which can contribute to a fast and efficient convergence.

Compared to the traditional correlation filter (CF) algorithms, We treat the hierarchical features differently. Because the shallow and deep features are both critical to the localization and classification, we train a set of independent filters for each feature. The decomposition of the function of two filters are beneficial to the overall performance. Conventional CF algorithms with one single filter is usually difficult to tackle with both classification and localization tasks during online tracking leading to model drift and insufficient online learning.

3.3 Filter Generations in Meta-learning Style

The motivation of our learning algorithm is that discriminative filters for similar visual objects in arbitrary background have amounts of sharing weights. Filters for objects with the same high-level semantic information should be robust towards changes, motion blur, scale variations, etc. To extract useful sharing filter weights in similar tracking scenes, we separate scene-independent information through offline training (Fig. 2).

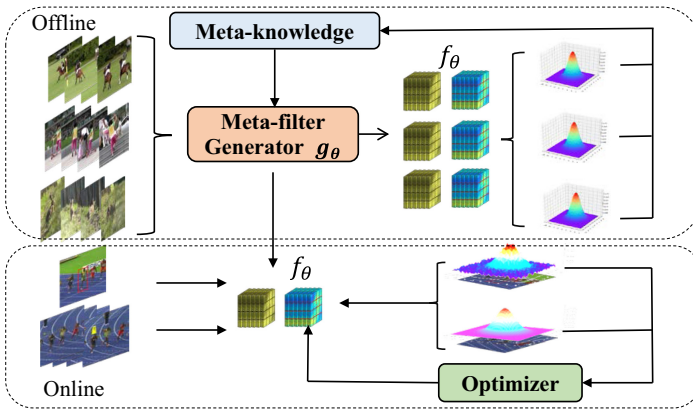


Fig. 2. Multi-hierarchical independent discriminative filters combined with online learning and offline training framework.

Algorithm 1. Meta-filters $f^{(i)}$ in offline training

Training samples $S_{\text{train}} = \{(x_j, c_j)\}_{j=1}^n$, test samples $S_{\text{test}} = \{(x_j, c_j)\}_{j=1}^m$

repeat

$f^{(i)} \leftarrow \text{FilterGen}(S_{\text{train}})$ # Generate filters

$\nabla L(f^{(i)}) \leftarrow \text{FilterGrad}(f^{(i)}, S_{\text{test}})$ # Apply filters

$\text{FilterGen} \leftarrow \text{BackProp}(\nabla L(f^{(i)}))$ # Update FilterGen

until N_{iter}

With these sharing weights stored in convolutional networks to generate meta-filters, our online discriminative model for classification can be adapted to the specific objects fastly. We introduce a network module called filter generation network g_θ . It consists of two convolutional layer and a precise ROI pooling. During offline training, the $S_{\text{train}} = \{(x_j, c_j)\}_{j=1}^n$, composed of several tracklets, are used to generate meta-filters through averaging the pooled feature maps. And then, the test samples $S_{\text{test}} = \{(x_j, c_j)\}_{j=1}^m$ are applied with generated filters to optimize the filter generation network.

Details of our meta-filters in Geographical Sensitivity Pathway (GASP) and Geometrically Sensitivity Pathway (GESp) are show in Fig. 3. ResNet-50 Block3 features in different stage are passed to a convolutional block (Cls). Regions defined by the input bounding boxes are then pooled to a fixed size using Precise Pooling layers. After a convolutional block, the weights of filter are generated to perform as convolutional block for features of searching image. Online optimizers optimize weights of filters during online tracking while offline optimizers try to learn meta-knowledges of filter-generation.

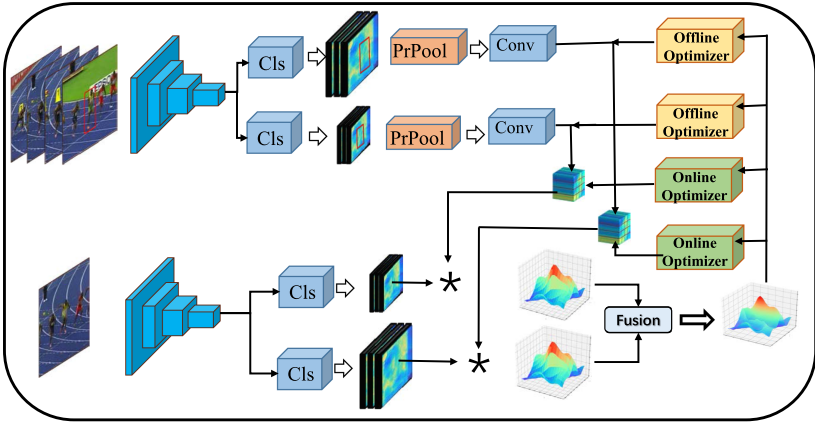


Fig. 3. Full architecture of our discriminative meta-filters. Pseudo-siamese network is not shown here for simplicity.

3.4 Appearance Fast Adaption Module

After the coarse spatial location of target is figured out, we need a subnetwork to acquire the accurate localization of target. In this work, we adopt an independent IoU-Nets [15] with template feature modulation. We train our independent IoU-Net [15] with template feature modulation for measuring the differences between proposals and ground truth. Full architecture can be viewed in Fig. 4.

The template features x_0 and searching area features x are extracted by modulation branch and test branch. The bounding box annotation A_0 is as extral

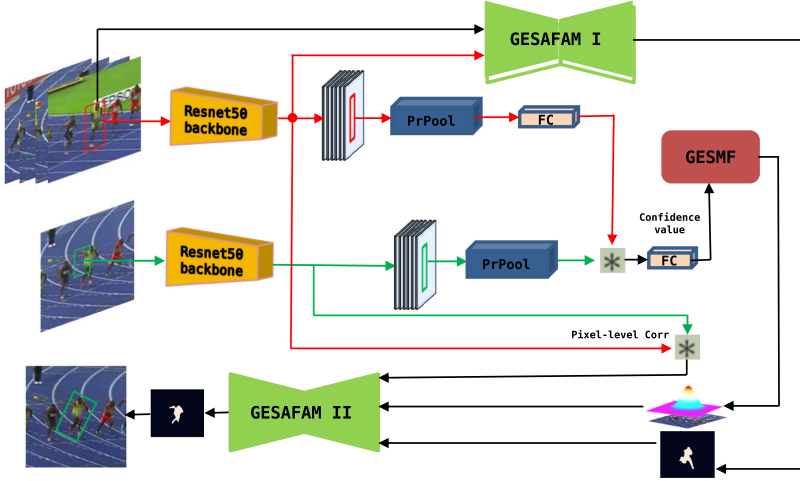


Fig. 4. Full architecture of our Appearance Fast Adaption Module (AFAM I) and Appearance Fast Adaption Module II (AFAM II).

Algorithm 2. Online searching strategy in Box Fast Adaption Module

Random Proposals Generation $P_i = \text{GaussianRandom}(A_{i-1}) , \sum_1^k i = 8$
 $N_{\text{iter}} = 0, \text{ step length} = \alpha$
repeat
 $f^{(i)} \leftarrow \text{testBranFexactor}(P_i)$ # Feature extraction for proposal
 $M_{\text{vec}} \leftarrow \text{ModuBran}(x_0, A_i)$ # Modulation convolutional kernels
 $\text{BoxConVal}^{(i)} = g(c(x_0, A_0) \cdot z(x, A))$ # Confidence value of proposal
 $p \leftarrow \text{BackProp}(\nabla(\text{BoxConVal}^{(i)}))$ # Optimization through Backward Propagation
 $P_i \leftarrow P_i + \alpha p$ # Proposal updation
 $N_{\text{iter}} = N_{\text{iter}} + 1$
until $N_{\text{iter}} = 10$

modulation information for generating box confidence value. The modulation information $c(x_0, A_0)$ is added to the test branch as convolution kernel. The feature representation of search area $z(x, A)$ has strong spatial correspondence with the searching frame. Thus it could reflect the spatial coordinate difference between template and test frame.

$$\text{BoxConVal}(A) = g(c(x_0, A_0) \cdot z(x, A)) \tag{8}$$

During online tracking, we apply another online searching strategy to maximum the confidence value with bounding box optimization. We use Gaussian distribution and previous position of target to generate initial proposals. For each proposal, we obtain the confidence value through the Box Fast Adaption Module. By backward propagation to obtain the gradient of confidence value,

we optimize the length and center position of current proposal directly. Details are shown in Algorithm 2.

Appearance Fast Adaption Module II (AFAM II) provides pixel-level target information. We use the features extracted from ResNet-50. For the first frame and ground truth, we obtain the pseudo-mask for the target from the AFAM II. Then the extra information from Appearance Fast Adaption Module I (AFAM I) and pseudo-mask are concatenated together. The refinement network will output the final appearance estimation. Although the AFAM II are pretrained on training segmentation sequences from Youtube-VOS, yet it is not design for the segmentation task. During training, we use bounding box labels as inputs to predict target mask. So it should be considered as target appearance estimation, not instance segmentation.

4 Experiments

Meta-filters in Geographical Sensitivity Pathway (GASMF) and Box Fast Adaption Module in Geometrically Sensitivity Pathway (GESBFAM) are firstly trained jointly with ImageNet pretrained weights. Because ImageNet pretrained models are for classification task which may not suitable for tracking, we firstly train the GASMP and GESBFAM 40 epochs in the training splits of TrackingNet [20], LaSOT [21], GOT10k [1] and COCO [22] datasets to adapt backbone to tracking task. Then, we add the meta-filters in Geometrically Sensitivity Pathway (GESMF) to train another 30 epochs for a more discriminative power model. We train our model by sampling 26,000 frame-pairs per epoch. We use ADAM [23] with learning rate decay of 0.2 every 10th epoch. We use features extracted from the third block from Resnet. We set the kernel size of the meta-filters to $64 * 4 * 4$. Appearance Fast Adaption Module (AFAM) in Geometrically Sensitivity Pathway are pre-trained on 3471 training segmentation sequences from Youtube-VOS [24].

4.1 Ablation Studies

We compared the performance of different combinations in Resnet50. ResNet-50 Block3 features in different stage. If we select adjacent layers, more redundancy and interference will be introduced into our tracking framework, thus causing the performance degradation. From the Table 2 the best performance achieved is from the layer3a and layer3e. When using two meta-filters, the EAO comes to 0.455, which demonstrates the effectiveness of two filters. The Box Fast Adaption Module improves accuracy a lot which is 0.652 comparing to 0.597 (Table 1).

4.2 Results on Several Benchmarks

VOT2018 [25] datasets consist of 60 test sequences. With no training dataset provided, VOT is the most challenging benchmark for tracking which has topics including fast motion, occlusion, etc. We tested our tracker on this benchmark

Table 1. VOT2018-comparison with different settings.

Tracker	Settings0	Settings1	Settings2	Settings3	Settings4	Settings5
Filters	Single	Single	Two	Two	Two	Two
Features	Block2e	Block3e	Block2e+3e	Block3e+3f	Block3a+3e	Block3a+3e
Box adaption	–	–	–	–	–	Yes
A ↑	0.595	0.596	0.596	0.597	0.596	0.652
R ↓	0.172	0.168	0.157	0.165	0.155	0.155
EAO ↑	0.421	0.435	0.455	0.448	0.457	0.467

and present the results in Table 2 and Table 3. To the best of our knowledge, we achieves an EAO of 0.467 on VOT2018 (Kristanetal, 2018) and EAO of 0.334 on VOT2019 benchmark which is the new state-of-the-art performance. Our tracker also can run at 30 FPS in Nvidia GeForce 1080ti which is still very competitive (Tables 4 and 5).

Table 2. VOT2018-comparison with state-of-the-art trackers. The top three results are in **red**, **blue** and **green** fonts. Best viewed in color display.

Tracker	Ours	SiamFC++ [10]	ATOM [2]	SiamRPN++ [8]	DaSiamRPN [7]	ECO [17]
Where	–	AAAI20	CVPR19	CVPR19	ECCV18	CVPR17
A ↑	0.652	0.587	0.590	0.600	0.586	0.484
R ↓	0.155	0.183	0.204	0.234	0.276	0.276
EAO ↑	0.467	0.426	0.401	0.414	0.383	0.280

Table 3. VOT2019 realtime-comparison with state-of-the-art trackers. The top three results are in **red**, **blue** and **green** fonts. Best viewed in color display.

Tracker	Ours	SiamBAN [26]	DiMP [18]	SiamRPN++ [8]	SiamMask [27]	SiamCRF [25]
Where	–	CVPR20	CVPR19	CVPR19	ECCV18	CVPR17
A ↑	0.636	0.602	0.582	0.599	0.594	0.549
R ↓	0.276	0.396	0.371	0.482	0.461	0.346
EAO ↑	0.334	0.327	0.321	0.285	0.287	0.262

Table 4. GOT-10K-comparison with state-of-the-art trackers. The top three results are in **red**, **blue** and **green** fonts. Best viewed in color display.

Tracker	Ours	SiamFC++	ATOM [2]	SiamRPN++ [8]	SiamFCv2 [5]	ECO [17]
Where	–	AAAI20	CVPR19	CVPR19	ICCV19	CVPR17
AO ↑	60.0	59.5	55.6	51.8	37.4	31.6
SR(0.5) ↑	71.6	69.5	63.4	61.8	40.4	30.9
SR(0.75) ↑	46.0	47.9	40.2	32.5	14.4	11.1

Table 5. OTB-15-The top three results are in **red**, **blue** and **green** fonts. Best viewed in color display.

Tracker	Ours	ATOM [2]	SiamRPN++ [8]	DaSiamRPN [7]	ECO [17]
Where	–	CVPR19	CVPR19	ECCV18	CVPR17
Success \uparrow	67.7	66.9	69.6	65.8	69.1

5 Conclusions

In this paper, we propose three suggestions on designing modern visual object trackers. We combine offline training and online learning of discriminative filters together. The meta-learning ways are stressed and successfully applied in object tracking. The meta-knowledge of the filter generations on similar tracking scenes are learned through convolutional network. Gradient descent optimization is carefully designed to adapt our filters to unseen objects efficiently. Moreover, a pseudo-siamese network structure empowers the discriminative ability of our meta-filters. Our tracker can perform online searching strategies to find the best object bounding box. The balance of online searching and offline training helps us to achieve better results with less training resource.

References

1. Huang, L., Zhao, X., Huang, K.: Got-10k: a large high-diversity benchmark for generic object tracking in the wild. In: IEEE Transactions on Pattern Analysis and Machine Intelligence (2019)
2. Danelljan, M., Bhat, G., Khan, F.S., Felsberg, M.: Atom: accurate tracking by overlap maximization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4660–4669 (2019)
3. Henriques, J.F., Caseiro, R., Martins, P., Batista, J.: High-speed tracking with kernelized correlation filters. IEEE Trans. Pattern Anal. Mach. Intell. **37**(3), 583–596 (2014)
4. Bolme, D.S., Beveridge, J.R., Draper, B.A., Lui, Y.M.: Visual object tracking using adaptive correlation filters. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 2544–2550. IEEE (2010)
5. Bertinetto, L., Valmadre, J., Henriques, J.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
6. Li, B., Yan, J., Wu, W., Zhu, Z., Hu, X.: High performance visual tracking with siamese region proposal network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8971–8980 (2018)
7. Zhu, Z., Wang, Q., Li, B., Wu, W., Yan, J., Hu, W.: Distractor-aware siamese networks for visual object tracking. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 101–117 (2018)
8. Li, B., Wu, W., Wang, Q., Zhang, F., Xing, J., Yan, J.: Siamrpn++: evolution of siamese visual tracking with very deep networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4282–4291 (2019)

9. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. In: *Advances in Neural Information Processing Systems*, pp. 91–99 (2015)
10. Xu, Y., Wang, Z., Li, Z., Ye, Y., Yu, G.: Siamfc++: towards robust and accurate visual tracking with target estimation guidelines. arXiv preprint [arXiv:1911.06188](https://arxiv.org/abs/1911.06188) (2019)
11. Sun, C., Wang, D., Lu, H., Yang, M.-H.: Correlation tracking via joint discrimination and reliability learning. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 489–497 (2018)
12. Valmadre, J., Bertinetto, L., Henriques, J., Vedaldi, A., Torr, P.H.S.: End-to-end representation learning for correlation filter based tracking. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2805–2813 (2017)
13. Bhat, G., Johnander, J., Danelljan, M., Khan, F.S., Felsberg, M.: Unveiling the power of deep tracking. In: *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 483–498 (2018)
14. Tian, Z., Shen, C., Chen, H., He, T.: FCOS: fully convolutional one-stage object detection. In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 9627–9636 (2019)
15. Jiang, B., Luo, R., Mao, J., Xiao, T., Jiang, Y.: Acquisition of localization confidence for accurate object detection. In: *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 784–799 (2018)
16. 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.) *European Conference on Computer Vision*, vol. 9909, pp. 472–488. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46454-1_29
17. Danelljan, M., Robinson, A., Shahbaz Khan, F., Felsberg M.: Eco: efficient convolution operators for tracking. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6638–6646 (2017)
18. Bhat, G., Danelljan, M., Van Gool, L., Timofte, R.: Learning discriminative model prediction for tracking. In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 6182–6191 (2019)
19. Lukezic, A., Matas, J., Kristan, M.: D3s-a discriminative single shot segmentation tracker. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7133–7142 (2020)
20. Muller, M., Bibi, A., Giancola, S., Alsubaihi, S., Ghanem, B.: Trackingnet: a large-scale dataset and benchmark for object tracking in the wild. In: *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 300–317 (2018)
21. Fan, H., 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*, pp. 5374–5383 (2019)
22. Lin, T.-Y., et al.: Microsoft coco: common objects in context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) *European Conference on Computer Vision*, vol. 8693, pp. 740–755. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10602-1_48
23. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint [arXiv:1412.6980](https://arxiv.org/abs/1412.6980) (2014)
24. Xu, N., et al.: Youtube-vos: sequence-to-sequence video object segmentation. In: *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 585–601 (2018)

25. Matej Kristan, et al.: The seventh visual object tracking vot2019 challenge results. In: Proceedings of the IEEE International Conference on Computer Vision Workshops (2019)
26. Chen, Z., Zhong, B., Li, G., Zhang, S., Ji, R.: Siamese box adaptive network for visual tracking. arXiv preprint [arXiv:2003.06761](https://arxiv.org/abs/2003.06761) (2020)
27. Wang, Q., Zhang, L., Bertinetto, L., Hu, W., Torr, P.H.S.: Fast online object tracking and segmentation: a unifying approach. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1328–1338 (2019)