Accurate Visual Tracking with Attention Feature Fusion

Shuo HU, Linna SUN, Hui YU*, *Senior Member, IEEE*

***Abstract*— Modern tracker demands to perform efficiently robust classification and accurate object state estimation. Recently, the feature fusion plays a vital role in term of accuracy and robustness for a visual tracking system. The traditional feature fusion methods are generally performed via direct summation or concatenation operation, which are entirely unaware of importance of assigning appropriate weights to different levels of features for a robust model. To tackle this issue, a novel deep-leaning based tracker with attention fusion is proposed in this paper. We propose a improved network structure based on ResNet, which is more conducive for the fusion of hierarchical features. The proposed tracker adopts a nonlinear method for feature fusion in backbone network, and introduces an iterative multi-scale attention module to learn different weights of the hierarchical features. In the classification network which is learned online, the third and fourth layer features extracted from the backbone network are used to obtain the coarse location. The extracted features are assigned different weights by an attention mechanism and fed into the estimation network to perform a iterative refinement for the accurate bounding box estimation. The experiments show the proposed tracker's efficiency and effectiveness.**

I. INTRODUCTION

Visual object tracking is a very important branch of computer vision, and has been widely used in many fields, such as video intelligent traffic monitoring, robotics, surveillance, and human–computer interactions [[1]–[5]]. The generic object tracker intends to learn an appearance model of the target based on less prior knowledge, often a single initial frame in the image sequences or video. Many State-of-the-art trackers adopt the framework combining an estimation module and a classification module [[6], [7]]. The classification module aims at performing a robust coarse location of the target. In recent years, many tracking methods with different robust classifiers have been proposed, which can be roughly divided into two categories: correlation filters based methods [[8], [9]] and deep feature representation methods [[10],[11]]. The estimation task is intent to obtain an accurate bounding box. For this task, the trackers [[12]–[14]] based on the Region Proposal Network (RPN) [[15]] provides an accurate and efficient target state estimation. However, the pre-defined anchor settings not only introduce ambiguous similarity scoring that severely hinders the robustness but also need access to prior information of data distribution [[7]]. ATOM [[6]] iteratively refines the coarse initial bounding boxes via gradient ascending for a higher overlap between the predicted bounding box and ground truth [[16]], and achieves a significant promotion in term of accuracy. However, Atom is not suitable for occlusion, deformation , and scale variation in some scenarios. To alleviate the problem mentioned above, we propose a novel ATOM-based tracking algorithm with feature fusion. A nonlinear fusion method is introduced in the backbone network of the original algorithm for robust feature representation, and an iterative multi-scale attention module named MS-CAM [[17]] is adopted to obtain different weight layers learned from hierarchical features. Contrasted to the traditional feature fusion method of direct summation, the MS-CAM can learn a better robust model. The fused features are fed into the estimation and classification network. It enables the feature extraction network to focus on the semantic features with high resolutions and suppress the disturbances. In the classification module which is learned online, the third and fourth layer features extracted from the backbone network are fused and sent into the classier to calculate the response map for obtaining the coarse target location, which exploits effectively the low-level information and high-level information for robustness. In the estimation module, a iterative refinement is performed for the accurate bounding box. Our contribution can be summarized as follow:

- To tackle the problem of the insufficient image features extracted by traditional convolutional neural network, an improved network structure based on the ResNet [18] is proposed. The proposed structure is more conducive for the fusion of hierarchical features. Different levels of features are assigned appropriate weights for robustness.
- Furthermore, an iterative multi-scale attention module is introduced to learn hierarchical weights of the features, which is efficient comparing to the original feature fusion of residual blocks.

II. RELATED WORKS

The scale variation of objects is one of the key challenges in object tracking. Some modern trackers including DCF [[19]] and SiamFC [[20]] , utilize multi-scale pyramids to estimate the target scale. The tracker mentioned above calculates the scale corresponding and assumed the highest classification score as the predicted target scale in the current frame. However, the bounding box estimation requires a high-level understanding of the state of targets. Inspired by DCF [[19]] and IoU-Net [[16]], ATOM [[6]] proposes a tracking architecture, consisting of dedicated target estimation and classification. The tracker iteratively refines the coarse initial bounding boxes via gradient ascending for a higher overlap between the predicted bounding box and ground truth [[16]], and achieves a significant promotion in term of accuracy. This approach not only makes a significant

*This project is supported by National Natural Science Foundation of China under Grant 62073279.

Shuo HU and Linna SUN are with the School of Electrical Engineering, Yanshan University, Qinhuangdao 066004, China; e-mail: hus@ysu.edu.cn

Hui YU is with School of Creative Technologies, University of Portsmouth, Winston Churchill Avenue, Portsmouth, PO1 2DJ, U K. (corresponding author, phone: +44(0)239 2845470; e-mail: hui.yu@port.ac.uk).

improvement on accuracy but also brings a heavy computation burden.

Recently, some researchers [[17]][[21]-[23]] utilize attention mechanisms to resolve the scale issue. Generally, in deep networks, the attention mechanism which mimics the human visual attention mechanism [[24]], is originally developed on a global scale. Inspired by ParseNet [[25]], MS-CAM [[17]], which combine local and global features in CNNs and follow the idea of spatial attention with aggregating multi-scale feature contexts inside the attention module, outperforms state-of-the-art networks with fewer layers or parameters per network.

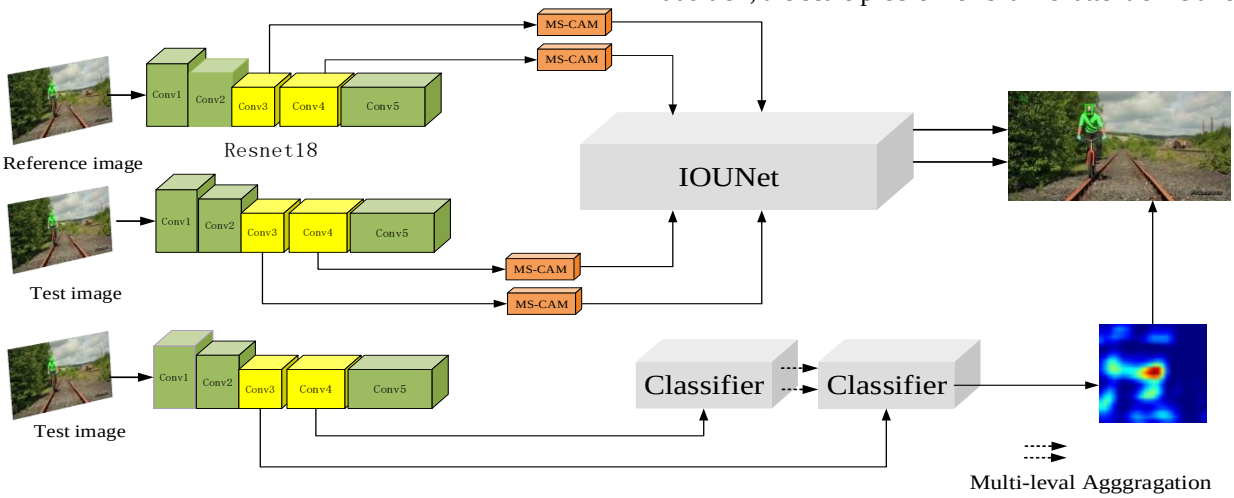


Figure 1 the framework of the proposed network

III. ATTENTIONAL FEATURE FUSION TRACKER FOR OBJECT TRACKING

In this section, we describe our attentional feature fusion tracker in detail. Based on the structure of ATOM [[6]], we introduce the an attentional feature fusion module in the backbone network. As shown in Figure 1, the proposed tracker consists of three components: classification, estimation which are similar to ATOM, and backbone network with an attentional fusion module.

The framework of the proposed algorithm is shown in Figure 1. Similar to ATOM [[6]], in the classification component, the correlation filtering method is utilized as the classifier to obtain the rough location of the target. And in the estimation component, the IOU-net [[16]] trained off-line is adopted for the accurate location of the target.

For the robustness in different scenarios, we replace the residual blocks in the third and fourth layers of the original backbone network resnet18 with the improved residual blocks. In general, the tracking process is divided into two stages: classification and estimation. In the on-line classification branch, we send the third and fourth layer features into the classifier respectively, and then obtain the weight-fusion response maps. The fusion features of the third and fourth layers which are calculated by the multi-scale attention module, are fed to the estimation network to perform a iterative refinement for an accurate box estimation.

A. Multi-scale Attention Mechanism

In this paper, a nonlinear feature fusion method is introduced to improve the feature extraction network, which aggregates multi-scale features into the attention module for promoting performance. A multi-scale attention (MS-CAM) model [[17]] is adopted in the attention module, which exploits the local and global features in CNNs, and aggregates multi-scale feature contexts in the attention module. In addition, the scale problem of channel attention is alleviate by dot convolution operation.

MS-CAM [[17]] aggregates the local and global feature context information within the channel attenuation. By changing the size of the space pool, channel attention can be performed on multiple scales. Meanwhile, aim at keeping the backbone network as lightweight as possible, only the local context information is squeezed into the global context inside the attention module.

MS-CAM[17] adopts point-wise convolution as the local channel context information aggregator, and utilizes point-wise channel interaction for each spatial location. In order to reduce the parameters, the local channel context information $L(X) \in R^{C\times H\times W}$ is calculated by the bottleneck structure as follows[17]

$$L(X) = \beta(PWConv_2(\sigma(\beta(PWConv_1(Z'))))) \quad (1)$$

The kernel sizes of $PWConv_1$ and $PWConv_2$ are $\frac{C}{r}\times C\times 1\times 1$ and $C\times\frac{C}{r}\times 1\times 1$ respectively. Generally, $L(x)$ has the same shape as the input features, which can retain and highlight the important details existed in the low-level features. Furthermore, the refined features $X' \in R^{C\times H\times W}$ can be obtained by MS-CAM as follows [[17]]

$$X' = X \otimes M(X) = X \otimes \sigma(L(X) \oplus g(X)) \quad (2)$$

Where $M(X) \in R^{C\times H\times W}$ is the attention weights, $L(x)$ is local channel context information, $g(x)$ is global channel context

information,⊕ is broadcasting addition, and ⊗ represents the matrix multiplication.

B. *Attention feature fusion module*

Aiming to solve the problems of feature fusion scene unification, feature context aggregation and initial integration, MS-CAM[17] adopts an attention feature fusion module (AFF) and iterative attention feature fusion module (IAFF). In our work, only IAFF is adopted for features fusion. The model is shown in Figure 2.

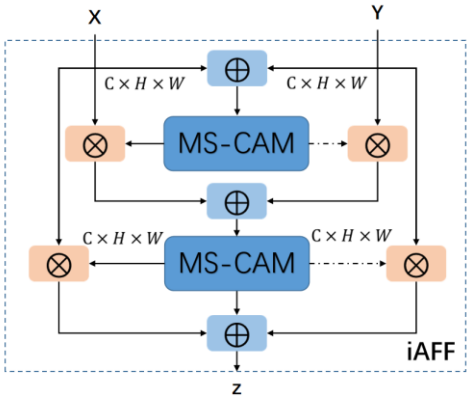


Figure 2 Illustration of iAFF

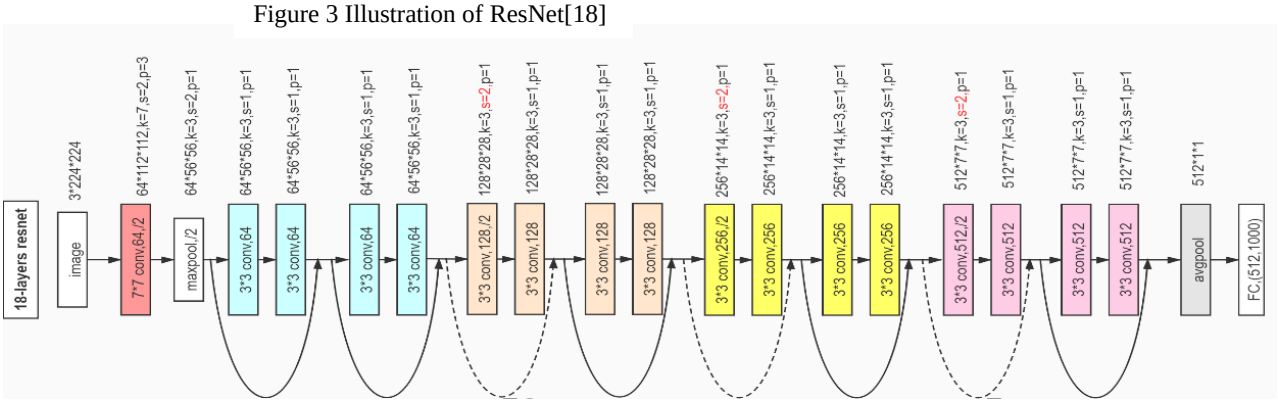
Where, $Z \in R^{C\times H\times W}$ is the fused feature. For simplicity, the element-wise summation is adopted as the initial integral. The AFF shown in Figure 3 can be expressed as follows

$$Z = M(X+Y)\otimes X + (1-M(X+Y))\otimes Y \quad (3)$$

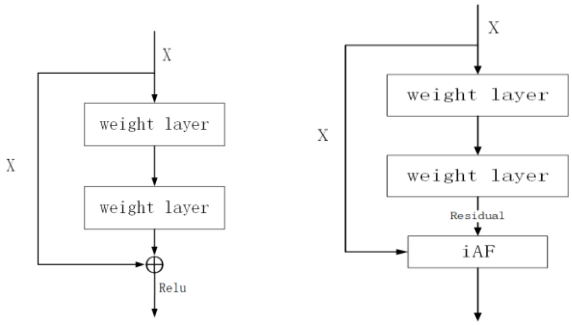
The dotted line denotes $1-M(X\oplus Y)$. For the capacity of soft selection or weighted averaging between X and Y. The fusion weight $M(X\oplus Y)$ is composed of real numbers between 0 and 1. And $1-M(X\oplus Y)$ also consists of real numbers between 0 and 1.

C. *the Improved Network Structure with Attention Feature Fusion*

Figure 3 Illustration of ResNet[18]



The backbone network of the proposed algorithm adopts ResNet [[18]] with shallow network layers. The structure of ResNet18 is shown in Figure 3. It can be seen from the figure that ResNet18 is composed of four basic blocks. The figure 5-a illustrates the basic structure of the residual block. From figure 5-a, we can known that the feature fusion adopts the method of adding the original features and the remaining features directly. We replace the fusion method with the nonlinear IAFF method, as shown in figure 5-b. The residual blocks 3 and 4 of the original feature extraction network are replaced by the improved residual block 3 and block 4.



(a) the original residual block (b) the improved residual block

Figure 4 Illustration of the proposed residual block

As shown in Figure 3, the yellow sections are the residual blocks with attention feature fusion. In the proposed tracker, the output features of the residual block mentioned above are sent to MS-CAM. Sequentially, the fused features are fed into the estimation network to predict the final regression box.

D. *Multi-layer Feature Fusion*

In general, the state-of-the-art tracker intends to use not only shallow features such as color and shape, but also deep features which contain semantic information. Shallow features are beneficial to the location of target tracking, however shallow features can not effectively improve the location accuracy of regression branches due to the lack of semantic information, meanwhile deep features can effectively improve the tracking performance in scenario change and other attributes. Inspired by RPT [26] and siamrpn++ [[14]], the proposed tracker uses the third and fourth layer features of ResNet as the output. The weighted fusion is performed on the response of these two layer features obtained by the filter to get the response maps of the target location. The procedure is shown in Figure 5.

We use $f(x)$ to denotes the final response map, $F_1(x)$ and $F_2(x)$ to denote the response map of the third and fourth layer features extracted by ResNet18 after filtering, and use

γ_1 and γ_2 to denote the weighted fusion coefficients. The expression of $f(x)$ is as follows:

$$f(x) = \sum_{i=1}^{2} \gamma_i f_i(x), \quad (4)$$

Where, $\gamma_1 = \frac{S_1}{S}, \gamma_2 = \frac{S_2}{S}$, S_1 and S_2 are the location confidence scores of the response map of target classification, and S is the sum of them.

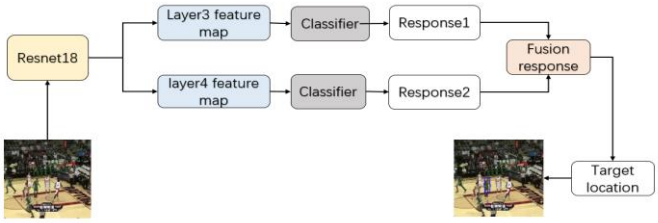


Figure 5 Procedure of the proposed tracker

IV. EXPERIMENTS

The experimental dataset is OTB2015 containing 100 video sequences. The platform mainly evaluates the algorithm from two indicators: accuracy and success rate. On the benchmark platform, the data set can make a detailed assessment of the tracking results of the tracking algorithm for each challenge and draw a graph.

Compared with the original algorithm, the training dataset we used is consistent with the original algorithm, using three large data sets trackingnet, lasot and coco2015. As more network parameters are trained, the original 40 epochs are increased to 50. The training is about 60 hours. The improved network, through attention feature fusion, can better integrate the scale and context information of the target, and gives more weights to the expressive features making the extracted features more robust.

DATASET

We evaluate the proposed tracker on OTB2015, and draw a comparison curve of the success rate and accuracy rate for the overall tracking effect and eleven indicators. The specific evaluation indicators of OTB include Low Resolution (LR) and Occlusion (OCC).,Deformation (DEF), Scale Variation (SV), Motion Blur (MB), Fast Motion (FM), In-Plane Rotation (IPR), Out-of-Plane Rotation (OPR), Out of View (OV), Background Clutters (BC) and Low Resolution (LR).

We observe that the original algorithm has poor performance against Occlusion (Occ), Deformation (DEF), Scale Variation (SV), and target Out of View (OV). We select five sequences that are challenging for these evaluation criteria to visualize the algorithm. They are Basketball, Bolt, Board, Matrix, Liquer, the visualization results are as follows

Figure 6(a) Basketball

Figure 6(b) Sequence.Bolt

Figure 6(c) Sequence Liquer

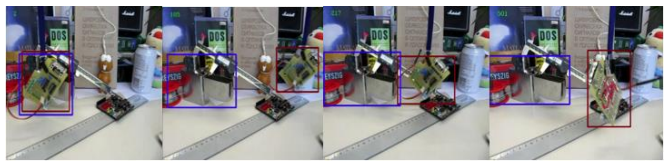


Figure 6(d) Sequence Matrix

Figure 6(e) Sequence Board

In the OTB dataset, the figure 6-a, 6-c, 6-d, and 6-e sequences, the target is occluded. The improved algorithm has a better tracking effect on these sequences than the original algorithm. In 6-a, 6-b, the target is deformed. In the 6-b, 6-c sequence, the target has a challenging attribute out of view. The proposed algorithm is more effective and has better tracking performance.

The proposed tracker is compared with the state-of-the-art trackers in terms of success and precision. The visualization results are as follows:

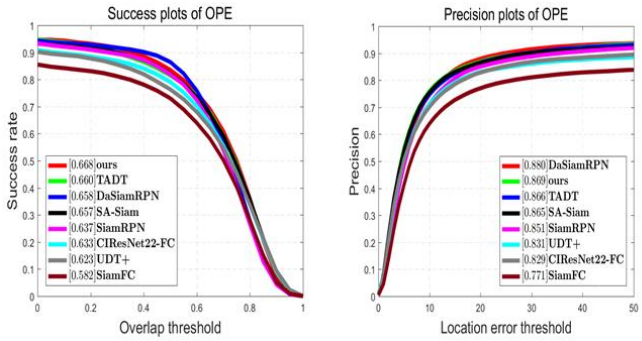


Figure 7-1 Success and Precision plot of OPE

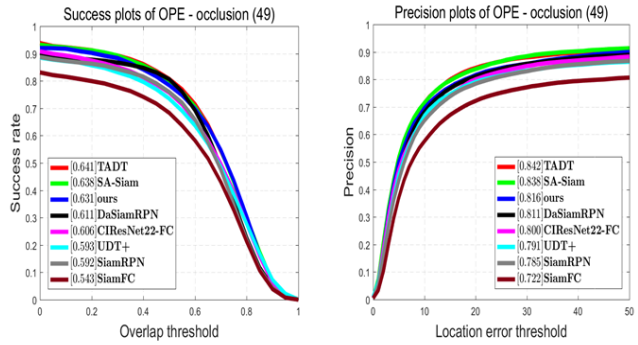


Figure 7-2 Success and Precision plot of OPE-Occ

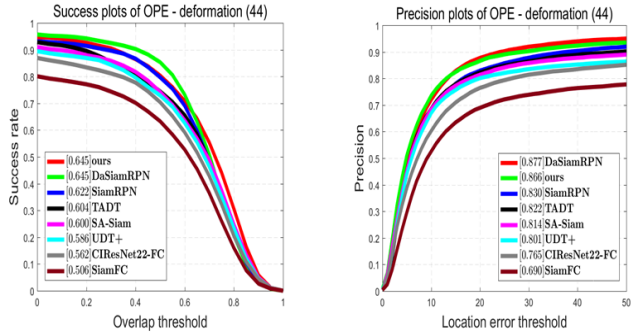


Figure 7-3 Success and Precision plot of OPE-DEF

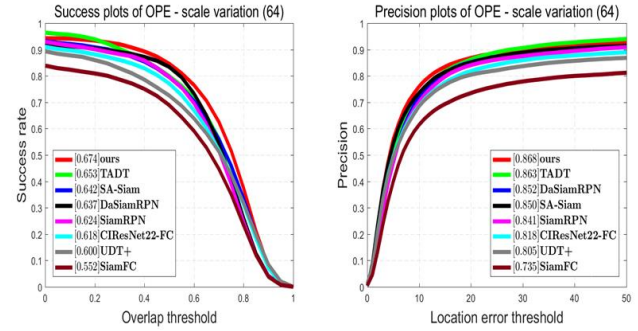


Figure 7-4 Success and Precision plot of OPE-SV

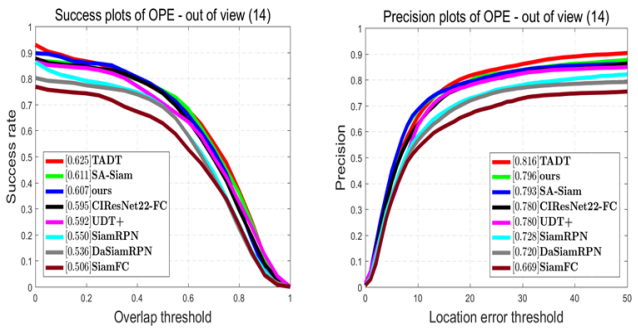


Figure 7-5 Success and Precision plot of OPE-OV

In Figure 7, we can see that for the Occlution (Occ) challenge in the target tracking task, the success rate of the proposed algorithm increases by 1.1%, and the accuracy rate increases by 1.6%. For the Deformation (DEF) challenge in the target tracking task, the success rate of the proposed algorithm is increased by 3.0%, and the accuracy rate is increased by 4.5%; for the Scale Variation (SV) challenge in the target tracking task, the success rate of the improved algorithm is increased by 0.9%, and the accuracy rate is increased by 0.3%; for the target tracking the Out of View (OV) challenge in the task, the improved algorithm's success rate increased by 3.7%, the accuracy rate increases by 2.1%, and the overall performance of the proposed algorithm increases by about 1%.

In Figure 7, we have also drawn the curves of several current mainstream algorithms for comparison, including TADT [27], SA-Siam [28], CIResNet22-FC [29], SiamFC [20], SiamsRPN [12]. DASiamsRPN [30] It can be observed that the proposed tracker outperforms other algorithms except DASiamRPN [30]. Among them, the CIResNet22-FC [29] algorithm is obtained by modifying the backbone network based on the SiamFC [20]. Through the comparison of CIResNet22-FC and SiamFC [20], we can also observe the importance of feature extraction in term of performance. The final result of the experiment proves the efficiency of the proposed feature fusion strategy.

V. Conclusion

In this paper, we propose an improved target tracking algorithm with attentional feature fusion. The algorithm uses a two-stage tracking strategy, which is target estimation for offline training and target positioning for online training. The features extracted by the backbone are sent to the classification and estimation module. The online classifier roughly estimates the target position, and then the estimation network obtains the final bounding box. We use an iterative attention mechanism for feature fusion in the backbone network, and use a channel attention mechanism and response graph fusion for feature fusion in the target position and target estimation. Comprehensive experiments have been conducted on the dataset OTB2015. Compared with the original algorithm, our method achieves a higher accuracy and success rate and is more robust to the object interference issue in the scene.

.

References

[1] Smeulders, A.W., Chu, D.M., Cucchiara, R., Calderara, S.; Dehghan, A.; Shah, M. Visual Tracking: An Experimental Survey. IEEE Trans. Pattern Anal. Mach. Intell. 2013, 36, 1442 – 1468.

[2] Trucco, E.; Plakas, K. Video Tracking: A Concise Survey. IEEE J. Ocean. Eng. 2006, 31, 520 – 529.

[3] Tsagkatakis, G.; Savakis, A. Online Distance Metric Learning for Object Tracking. IEEE Trans. Circuits Syst. Video Technol. 2011, 21, 1810 – 1821.

[4] Ming Y , Meng X , Fan C , et al. Deep Learning for Monocular Depth Estimation: A Review[J]. Neurocomputing, 2021.

[5] Zhang, X.; Yu, Q.; Yu, H. Physics Inspired Methods for Crowd Video Surveillance and Analysis: A Survey. IEEE Access 2018, 6, 66816 – 66830. doi:10.1109/ACCESS.2018.2878733.

[6] Danelljan, M.; Bhat, G.; Khan, F. S.; and Felsberg, M. Atom: Accurate tracking by overlap maximization. In Proceedings of the CVPR, 2019，4660–4669

[7] Xu Y , Wang Z , Li Z , et al. SiamFC++: Towards Robust and Accurate Visual Tracking with Target Estimation Guidelines[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(7):12549-12556.

[8] A. Lukezic, T. Voj´ır, L. C. Zajc, J. Matas, and M. Kristan. Discriminative correlation fifilter tracker with channel and spatial reliability. IJCV, 126(7):671–688, 2018.

[9] C. Ma, J.-B. Huang, X. Yang, and M.-H. Yang. Hierarchical convolutional features for visual tracking. In ICCV, 2015.
[10] G. Bhat, J. Johnander, M. Danelljan, F. S. Khan, and M. Felsberg. Unveiling the power of deep tracking. In ECCV, 2018.
[11] J. Valmadre, L. Bertinetto, J. F. Henriques, A. Vedaldi, and P. H. S. Torr. End-to-end representation learning for correlation fifilter based tracking. In CVPR, 2017
[12] Li, B.; Yan, J.; Wu, W.; Zhu, Z.; and Hu, X. High performance visual tracking with siamese region proposal network. In Proceedings of the CVPR, 2018, 8971–8980.
[13] Zhu, Z.; Wang, Q.; Li, B.; Wu, W.; Yan, J.; and Hu, W. Distractor-aware siamese networks for visual object tracking. In Proceedings of the European Conference on Computer Vision (ECCV), 2018, 101–117.
[14] Li, B.; Wu, W.; Wang, Q.; Zhang, F.; Xing, J.; and Yan, J. Siamrpn++: Evolution of siamese visual tracking with very deep networks. In Proceedings of the CVPR, 2019, 4282–4291.
[15] Ren, S.; He, K.; Girshick, R.; and Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. In Advances in neural information processing systems, 2015, 91–99.
[16] Jiang, B.; Luo, R.; Mao, J.; Xiao, T.; and Jiang, Y. Acquisition of localization confifidence for accurate object detection. In Proceedings of ECCV, 2018,784–799.
[17] Yimian Dai, Fabian Gieseke, Stefan Oehmcke, Yiquan Wu, and Kobus Barnard. Attentional feature fusion. arXiv e-prints, page arXiv:2009.14082, Sep. 2020
[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In CVPR, 2016, 770–778.
[19] Henriques J F , Caseiro R , Martins P , et al. High-Speed Tracking with Kernelized Correlation Filters[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2015, 37(3):583-596.
[20] Bertinetto L , Valmadre J , Henriques J F , et al. Fully-Convolutional Siamese Networks for Object Tracking[C]// European Conference on Computer Vision. Springer, Cham, 2016.
[21] Irwan Bello, Barret Zoph, Ashish Vaswani, Jonathon Shlens, and Quoc V. Le. Attention augmented convolutional networks. In IEEE International Conference on Computer Vision (ICCV), pages 3286–3295, October 2019.
[22] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 3146–3154, 2019.
[23] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 7132–7141, 2018.
[24] Keren Fu, Deng-Ping Fan, Ge-Peng Ji, and Qijun Zhao. JLDCF: joint learning and densely-cooperative fusion framework for RGB-D salient object detection. In CVPR, 2020, pages 3049–3059.
[25] H. Zhang, C. Wu, Z. Zhang, and etc. ResNeSt: SplitAttention Networks. arXiv e-prints, page arXiv:2004.08955, Apr. 2020
[26] Z. Ma, L. Wang, H. Zhang, W. Lu, J. Yin. RPT: Learning Point Set Representation for Siamese Visual Tracking. arXiv e-prints, page arXiv:2008.03467, Sep. 2020
[27] He, Anfeng, et al. "A twofold siamese network for real-time object tracking." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018.
[28] Li, Xin, et al. "Target-aware deep tracking." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019.
[29] Zhang, Zhipeng, and Houwen Peng. "Deeper and wider siamese networks for real-time visual tracking." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019.
[30] Zhu, Zheng, et al. "Distractor-aware siamese networks for visual object tracking." Proceedings of the European Conference on Computer Vision (ECCV). 2018.