VisDrone-SOT2019: The Vision Meets Drone Single Object Tracking Challenge Results

Dawei Du[1], Pengfei Zhu[2], Longyin Wen[3], Xiao Bian[4], Haibin Ling[5], Qinghua Hu[2], Jiayu Zheng[2], Tao Peng[2], Xinyao Wang[3], Yue Zhang[3], Liefeng Bo[3], Hailin Shi[26], Rui Zhu[26], Bo Han[6], Chunhui Zhang[8,24], Guizhong Liu[16], Han Wu[16], Hao Wen[25], Haoran Wang[7], Jiaqing Fan[9], Jie Chen[7], Jie Gao[7], Jie Zhang[7], Jinghao Zhou[10], Jinliu Zhou[7], Jinwang Wang[11], Jiuqing Wan[12], Josef Kittler[13], Kaihua Zhang[9], Kaiqi Huang[14], Kang Yang[9], Kangkai Zhang[8,24], Lianghua Huang[14], Lijun Zhou[15], Lingling Shi[7], Lu Ding[17], Ning Wang[9], Peng Wang[10], Qintao Hu[15], Robert Laganière[18], Ruiyan Ma[7], Ruohan Zhang[7], Shanrong Zou[10], Shengwei Zhao[8,24], Shengyang Li[19], Shengyin Zhu[20], Shikun Li[8,24], Shiming Ge[8,24], Shiyu Xuan[19,24], Tianyang Xu[21,13], Ting He[6], Wei Shi[22], Wei Song[7], Weiming Hu[14], Wenhua Zhang[7], Wenjun Zhu[6], Xi Yu[6], Xianhai Wang[9], Xiaojun Wu[21], Xiaotong Li[7], Xiaoxue Li[7], Xiaoyue Yin[10], Xin Zhang[7], Xin Zhao[14], Xizhe Xue[10], Xu Lei[11], Xueyuan Yang[16], Yanjie Gao[7], Yanyun Zhao[20]. Yinda Xu[6], Ying Li[10], Yong Wang[18], Yong Yang[16], Yuting Yang[7], Yuxuan Li[7], Zeyu Wang, Zhenhua Feng[13], Zhipeng Zhang[14], Zhiyong Yu[6], Zhizhao Duan[6], Zhuojin Sun[23]

[1]University at Albany, SUNY, Albany, NY, USA.
[2]Tianjin University, Tianjin, China.
[3]JD Digits, Mountain View, CA, USA.
[4]GE Global Research, Niskayuna, NY, USA.
[5]Stony Brook University, New York, NY, USA.
[6]Zhejiang University, Hangzhou, China.
[7]Xidian University, Xi'an, China.
[8]Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China.
[9]Nanjing University of Information Science and Technology, Nanjing, China.
[10]Northwestern Polytechnical University, Xi'an, China.
[11]Wuhan University, Wuhan, China.
[12]Beihang University, Beijing, China.
[13]University of Surrey, Guildford, UK.
[14]Institute of Automation, Chinese Academy of Sciences, Beijing, China.
[15]Institute of Optics and Electronics, Chinese Academy of Sciences, Chengdu, China.
[16]Xi'an Jiaotong University, Xi'an, China.
[17]Shanghai Jiao Tong University, Shanghai, China.
[18]University of Ottawa, Ottawa, Canada.
[19]Technology and Engineering Center for Space Utilization, Chinese Academy of Sciences, Beijing, China.
[20]Beijing University of Posts and Telecommunications, Beijing, China.
[21]Jiangnan University, Wuxi, China.
[22]INSKY Lab, Leotail Intelligent Tech, Shanghai, China.
[23]Yuneec Aviation Technology, Shanghai, China.
[24]University of Chinese Academy of Sciences, Beijing, China.
[25]CloudWalk Technology Co. Ltd., Guangzhou, China.
[26]JD AI Research, Beijing, China.

Abstract

The Vision Meets Drone (VisDrone2019) Single Object Tracking challenge is the second annual research activity focusing on evaluating single-object tracking algorithms on drones, held in conjunction with the International Conference on Computer Vision (ICCV 2019). The VisDrone-SOT2019 Challenge goes beyond its VisDrone-SOT2018 predecessor by introducing 25 more challenging sequences for long-term tracking. We evaluate and discuss the results of 22 participating algorithms and 19 state-of-the-art trackers on the collected dataset. The challenge results are publicly available at the website: http://www.aiskyeye.com/. We expect the VisDrone-SOT challenge to boost the research in single object tracking field.

1. Introduction

Single Object Tracking (SOT), or visual tracking, is one of the fundamental techniques of computer vision and the base of many high-level applications such as video surveillance and human-computer interaction. A large amount of state-of-the-art tracking methods are proposed to deal with various challenging factors in visual tracking including occlusion and deformation.

Recently, drones equipped with cameras have been applied in a wide range of applications because of its flexibility. Compared to the traditional cameras, drones bring new challenges to the tracking methods, such as abrupt camera motion, small target, and view point change. To deal with these problems, there is high demanding for new drone based tracking algorithms and datasets [14, 35]. However, the studies are seriously restricted by the lack of publicly available large-scale drone based datasets.

In 2018, The VisDrone team is established to advance the developments in detection and tracking algorithms for drone based scenes [51, 58, 59]. Specifically, the challenge for single-object tracking has been carried out in conjunction with the 15-th European Conference on Computer Vision (ECCV 2018), where 17 submitted trackers and 5 state-of-the-art methods are evaluated on the proposed VisDrone2018-SOT dataset. However, the previous challenge focuses on short-term tracking. In this year, we expand the dataset with more challenging sequences in terms of long-term tracking. Moreover, we conduct comprehensive evaluation for 41 tracking methods including 22 submissions and 19 state-of-the-art trackers for both short-term and long-term tracking. This paper summarizes the VisDrone-SOT2019 Challenge organized in conjunction with the 26-th International Conference on Computer Vision (ICCV2019) Drone Meets Drone: A Challenge workshop. All the results can be found at the website: http://www.aiskyeye.com/.

2. Related Work

In this section, we first describe the related training and evaluation datasets for visual tracking. Then we introduce state-of-the-art tracking algorithms, especially the Siamese network based trackers.

2.1. Training and Evaluation Datasets

In recent years, single-object tracking is dominated by deep learning based methods due to its discriminative representation. However, further improvement of tracking performance is restricted by existing small-scale benchmarks, such as OTB [52], NFS [18], UAVDT [14], UAV123 [35], and VOT2018 [26].

To solve this problem, more large-scale training datasets [37, 16, 24] are proposed in the community, which fully represents various appearance and motion patterns of objects in the wild. Based on the data of YoutubeBB [39], TrackingNet [37] annotates more than $30K$ videos with more than 14 million bounding boxes. LaSOT [16] collects $1,400$ challenging sequences with average $2,512$ frames per sequence. Moreover, every frame is carefully and manually annotated with a bounding box. GOT-10k [24] contains more than $10,000$ video clips with over 1.5 million manually labeled bounding boxes. It includes a majority of $560+$ classes of real-world moving objects and $80+$ classes of motion patterns.

Except the aforementioned visual tracking datasets, object detection datasets [41, 39, 30] are also introduced to facilitate the training of tracking networks. For ImageNet VID [41], 30 different classes of animals and vehicles are provided with almost 4500 videos and a total of more than one million annotated frames. YoutubeBB [39] is a large-scale object detection dataset with approximately $380,000$ video segments, which is annotated every second with upright bounding boxes. COCO [30] is a large-scale object detection, segmentation, and captioning dataset with $330K$ images and 80 object categories.

2.2. Siamese Network based Trackers

Compared with traditional trackers [22, 10, 8, 13, 15], deep learning based methods achieve comparably or better performance. However, it is difficult for realtime practical applications because of high computational complexity of neural networks. Recently, the Siamese network based trackers [44, 3, 21, 49, 28, 60, 27, 50] become popular for both high tracking accuracy and efficiency.

Tao *et al.* [44] tracks the target, simply by matching the initial target in the first frame with candidates in a new frame by a learned matching function of Siamese network. Similarly, Bertinetto *et al.* [3] train a novel end-to-end fully-convolutional Siamese network on the ILSVRC15 dataset for object detection in video. Held *et al.* [21] learn offline a generic relationship between an object's appearance

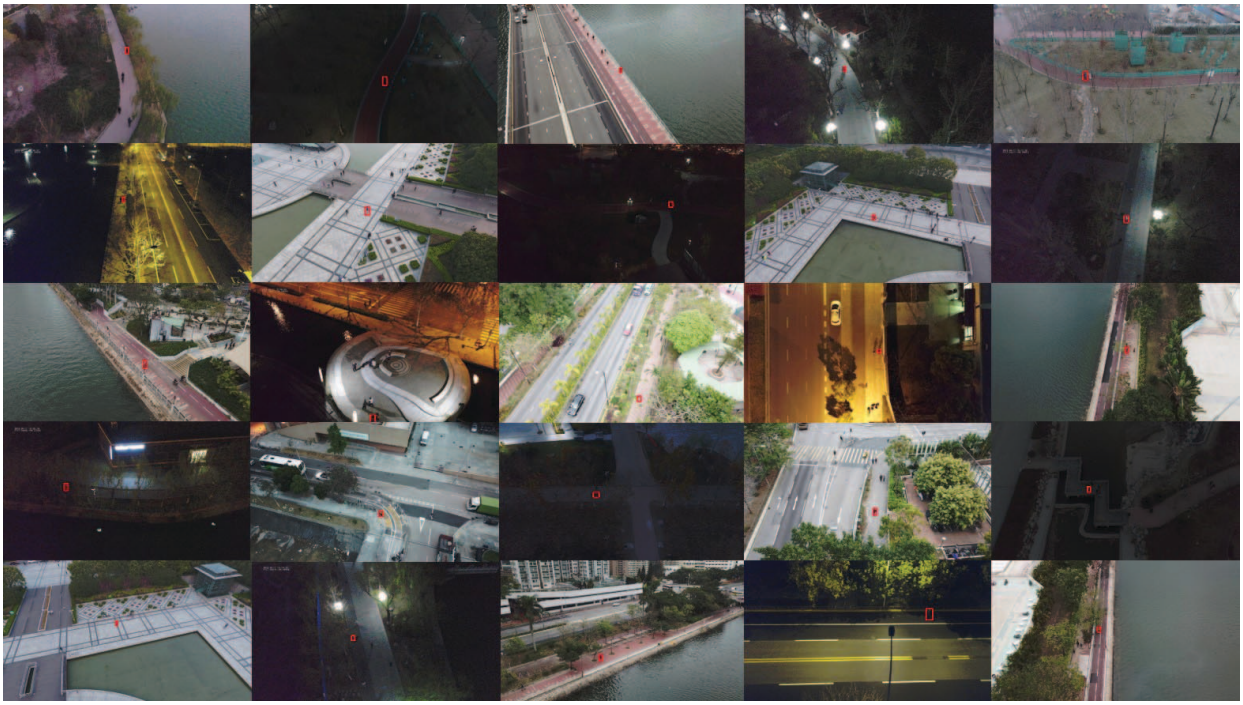

Figure 1. Screen-shots of new added sequences for long-term tracking. The first frame with the bounding box of the target is shown for each sequence.

and its motion. Inspired from correlation filters, Wang *et al.* [49] learn the convolutional features and perform the correlation tracking process simultaneously. Inspired from object detection, current state-of-the-art Siamese trackers [28, 60, 27] introduce the region proposal network after the Siamese network, resulting in promising tracking performance. Moreover, Wang *et al.* [50] improves the Siamese tracker by augmenting its loss with a binary segmentation task.

3. The VisDrone-SOT2019 Challenge

Following VisDrone-SOT2018 Challenge [51], The authors participating in the VisDrone-SOT2019 Challenge are required to submit no more than three tracking results for VisDrone-SOT2019 testing set. Then the best result among three submissions are selected as the performance of this tracker. We encourage the participants to use the provided training data, but also allow them to use additional training data such as UAVDT [14], UAV123 [35], TrackingNet [37], LaSOT [16], GOT-10k [24], and YoutubeBB [39].

3.1. The VisDrone-SOT2019 Dataset

Although the VisDrone-SOT2018 dataset [51] is not saturated, it lacks of sequences with long-term tracking. Therefore, we introduce the VisDrone-SOT2019L dataset in this challenge as the additional testing set. It includes 25 new collected challenging sequences ($82,644$ frames in total), which consists of 12 clips in the daylight and 13 clips at night. The scale of targets in the VisDrone-SOT2019L dataset is much smaller than that in the VisDrone-SOT2018 dataset, *i.e.*, 25.5 vs. 62.5. Besides, more distractors are introduced in this dataset (*e.g.*, occlusion, camera motion, and similar object). The screen-shots of new sequences are shown in Figure 1. That is, the VisDrone-SOT2019 dataset consists of VisDrone-SOT2018 and VisDrone-SOT2019L datasets. In summary, it includes 167 sequences with $188,998$ frames, which is divided into three non-overlapping subsets, *i.e.*, training set (86 sequences with $69,941$ frames), validation set (11 sequences with $7,046$ frames), and testing set (60 sequences with $112,011$ frames).

Similar to the VisDrone-SOT2018 Challenge [51], all sequences are sequence-level annotated by the following visual attributes: *aspect ratio change*, *background clutter*, *camera motion*, *fast motion*, *full occlusion*, *illumination variation*, *low resolution*, *out-of-view*, *partial occlusion*, *similar object*, *scale variation* and *viewpoint change*. On the other hand, two primary measures are used to analyze tracking performance: success and precision scores. Success score calculates the area under the curve based on the percentage of successfully tracked frames vs. the bounding box overlap threshold; while precision score denots the percentage of frames where the centers of the tracked object are within 20 pixels to the groundtruth. We refer to [52] for

further details.

3.2. Submitted Trackers

We have received 22 trackers from 19 different institutes in the VisDrone-SOT2019 Challenge. Many of them are improved from major computer vision conferences in very recent years. We briefly overview the submitted trackers and provide their descriptions in the Appendix A.

Among in the submitted algorithms, 9 trackers are inspired based on the very recent ATOM tracker [9], including ACNT (A.1), AST (A.2), ATOMFR (A.3), ATOMv2 (A.4), DATOM_AC (A.5), ED-ATOM (A.8), MATOM (A.11), SSRR (A.19) and TIOM (A.22). 5 trackers are variations of Siamese networks, *i.e.*, DC-Siam (A.6), DR-V-LT (A.7), SiamDW-FC (A.14), SiamFCOT (A.15) and SOT-SiamRPN++ (A.18). Two trackers are based on correlation filters, HCF (A.10) and TDE (A.21). Two trackers are based on convolutional neural networks, MDNet flow_MDNet_RPN (A.9) and Stable-DL (A.20). PTF (A.12) and SE-RPN (A.13) combine correlation filters and ATOM [9], while Siam-OM (A.16) and SMILE (A.17) are based on Siamese networks and ATOM [9].

Then, we evaluate 19 state-of-the-art tracking methods for comprehensive evaluation, including BACF [19], CSRDCF [32], DSiam [20], DSST [10], ECO [8], fDSST [12], HCFT [33], KCF [22], LCT [34], MDNet [38], PTAV [17], SCT [7], SRDCF [11], Staple [2], Staple_CA [36], STRCF [29], TRACA [6], SiameseFC [3], and CFNet [46]. Thus, we have in total 41 tracking methods in the VisDrone2019-SOT Challenge.

4. Results and Analysis

In this section, we first evaluate all the tracking methods on the overall VisDrone-SOT2019 dataset, and then discuss several representative trackers in terms of short-term tracking, long-term tracking and different visual attributes, respectively. Finally, several potential research directions are concluded.

4.1. Overall Performance

The overall tracking results are summarized in success and precision plots in Figure 2. Meanwhile, we also report the success and precision scores, tracking speed, implementation details, pre-trained dataset, and the references of each method in Table 1.

Compared with the VisDrone2018 Challenge [51], more submissions are based on convolutional neural networks (*e.g.*, ResNet and Siamese Network) due to the impressive performance, except HCF (A.10) and TDE (A.21). ED-ATOM (A.8) employs the IOU-predictor network to estimate the target, which is trained on several large-scale additional tracking datasets (*i.e.*, ImageNet, COCO, Got10k and LaSOT). Besides, a low-light image enhancement module [54] is applied to improve robustness. It not only achieves better performance than the other submissions, but has near real-time running speed of 20 FPS. Without pre-training on external tracking datasets, ATOMFR (A.3) (rank 2) embeds the SENet [23] into IoUNet in ATOM [9], which captures the interdependencies within feature channels. SMILE (A.17) (rank 3) combines two state-of-the-art tracking methods including ATOM [9] and SiamRPN++ [27]. Siam-OM (A.16) deals with short-term tracking by ATOM [9] and long-term tracking by DaSiam [60], respectively. To deal with blurred scenes, both SMILE (A.17) and Siam-OM (A.16) use the SIFT method [31] to match features. DR-V-LT (A.7) is improved based on SiamRPN++ [27], and distinguishes similar objects by the additional verification network (MDNet [38]). Following DR-V-LT (A.7), SOT-SiamRPN++ (A.18) magnifies the original images twice to improve the performance for small targets. On the other hand, SiamDW-FC (A.14) runs in the higher speed of 75 FPS than SiamFCOT (A.15), but achieves relatively lower success and precision scores.

In terms of the precision scores, the best two trackers stay the same, *i.e.*, ED-ATOM (A.8) and ATOMFR (A.3). Besides, it is worth mentioning that the state-of-the-art trackers (*i.e.*, ECO [8] and MDNet [38]) submitted by the VisDrone Team rank at the middle level of all the 41 tracking methods based on the success and precision scores. That means many submitted methods achieve considerable improvements from existing methods.

Table 1. Comparison of all submissions in the VisDrone-SOT2019 Challenge. The success score, precision score, tracking speed (in FPS), backbone, and pre-trained dataset (C indicates COCO [30], G indicates Got-10k [24], I indicates ImageNet DET/VID [41], L indicates LaSOT [16], T indicates TrackingNet [37], V indicates VOT [26], Y indicates YoutubeBB [39], and $\times$ indicates that the methods do not use the pre-trained datasets) are reported.

Method	Success	Precision	Speed	Backbone	Pre-trained
ACNT (A.1)	53.2	69.8	5	ResNet-50	C,I,T
AST (A.2)	51.9	69.5	40	ResNet-50	C,G,I,L
ATOMFR (A.3)	61.7	84.2	7	ResNet-18	$\times$
ATOMv2 (A.4)	46.8	60.8	25	ResNet-18	L
DATOM_AC (A.5)	54.1	74.1	20	ResNet-50	I
DC-Siam (A.6)	46.3	63.6	2	ResNet-50	I,Y
DR-V-LT (A.7)	57.9	76.8	3	ResNet-50	$\times$
ED-ATOM (A.8)	**63.5**	**90.0**	20	ResNet-18	C,G,I,L
flow_MDNet_RPN (A.9)	52.6	75.0	1	VGG-M	$\times$
HCF (A.10)	36.1	50.7	10	$\times$	$\times$
MATOM (A.11)	40.9	57.2	30	ResNet-18	L,T
PTF (A.12)	54.4	76.1	2	ResNet-50	I
SE-RPN (A.13)	41.9	56.3	40	ResNet-50	I,Y
SiamDW-FC (A.14)	38.3	52.9	**75**	CIResNet22W	G
SiamFCOT (A.15)	47.2	59.2	48	GoogLeNet	C,G,I,L,Y
Siam-OM (A.16)	59.3	83.3	15	ResNet	C,I,V
SMILE (A.17)	59.4	81.6	1.5	ResNet	C,I
SOT-SiamRPN++ (A.18)	56.8	79.3	3.2	ResNet-50	C,I,Y
SSRR (A.19)	44.7	58.8	40	ResNet-34	C,L
Stable-DL (A.20)	38.2	54.6	0.8	VGG-19	$\times$
TDE (A.21)	37.2	56.5	0.3	ResNet-50	I
TIOM (A.22)	49.7	76.5	1	ResNet-18	L

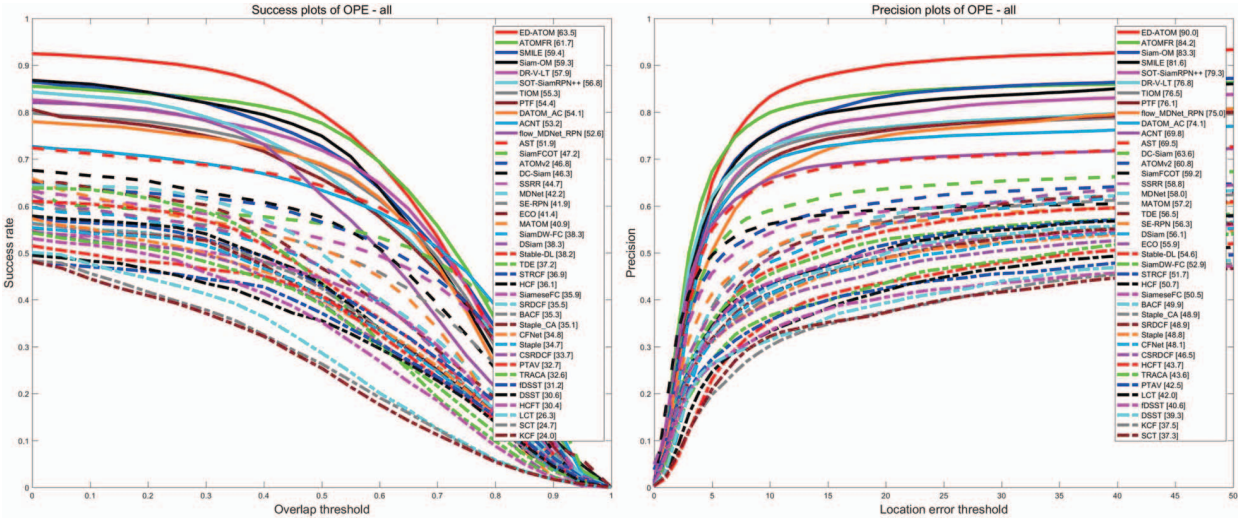

Figure 2. The success and precision plots of the submitted trackers. The success and precision scores for each tracker are presented in the legend.

4.2. Short-term Tracking

Compared to the trackers in VisDrone-SOT2018 Challenge [51], the submissions in this year further improve the tracking performance in short-term tracking. As shown in Figure 3, all the top 5 trackers in this year achieve larger than 71.0 success score. Besides, three of the top 5 trackers (*i.e.*, ED-ATOM (A.8) and ATOMFR (A.3) and Siam-OM (A.16)) achieve larger than 93.0 precision score. They perform better than the best tracker of VisDrone-SOT2018 Challenge [51] LZZ-ECO with 68.0 success score and 92.9 precision score. This is attributed to more accurate target scale estimation in the baseline ATOM [9]. Different from the correlation filters based tracking methods with a simple multi-scale search of target scales, ATOM [9] predicts the overlap between the ground-truth and an estimated bounding box directly.

4.3. Long-term Tracking

Generally speaking, a practical tracker in real-world applications should be able to track the object in a relative long period. However, the majority of current datasets includes the sequences with less than 600 frames [16]. Compared with VisDrone-SOT2018 challenge focusing on short-term tracking, we add 25 long-term tracking sequences (3,300 frames in average) in this year, *i.e.*, VisDrone-SOT2019L dataset. The corresponding tracking results are shown in Figure 4. Compared to the tracking performances in short-term tracking, the performances on the VisDrone-SOT2019L dataset are severely degraded because of a large mount of target drifting. ED-ATOM (A.8) achieves the best performance with 48.9 success score and 81.9 precision score, followed by DR-V-LT (A.7). There are more significant drop in performance for the other submitted trackers. This phenomenon demonstrates the effectiveness of the target verification module in visual tracking, especially in complex scenarios.

4.4. Performance Analysis by Attributes

We show the performance of each tracker in terms of 12 attributes in Figure 5 and Figure 6. ED-ATOM (A.8) achieves the best performance in 5 attribute subsets including *background clutter*, *fast motion*, *illumination variation*, *low resolution*, and *similar object*. ATOMFR (A.3) ranks at the first place in the remain 7 attributes including *aspect ratio change*, *camera motion*, *full occlusion*, *out-of-view*, *partial occlusion*, *scale variation* and *viewpoint change*.

In terms of *illumination variation* and *low resolution*, SMILE (A.17) and DR-V-LT (A.7) rank the second place respectively. By contrast, ATOMFR (A.3) ranks the 9-th and 5-th places. We speculate that it has no corresponding solution to deal with appearance variations especially for small objects.

4.5. Discussion

To design more effective tracker in drone based scenarios, there are several directions worth to explore based on the submitted algorithms.

- **Data Argumentation.** Data argumentation is important in network training with limited training data. The following data argumentation modules can be used: rescale (SOT-SiamRPN++ (A.18)), horizontal flip, rotation, shift, image contrast by Gamma correction (ED-ATOM (A.8) and Sima-OM (A.16)) and Laplacian operator (SOT-SiamRPN++ (A.18)).

- **Key-points Estimation.** In drone based scenes, there

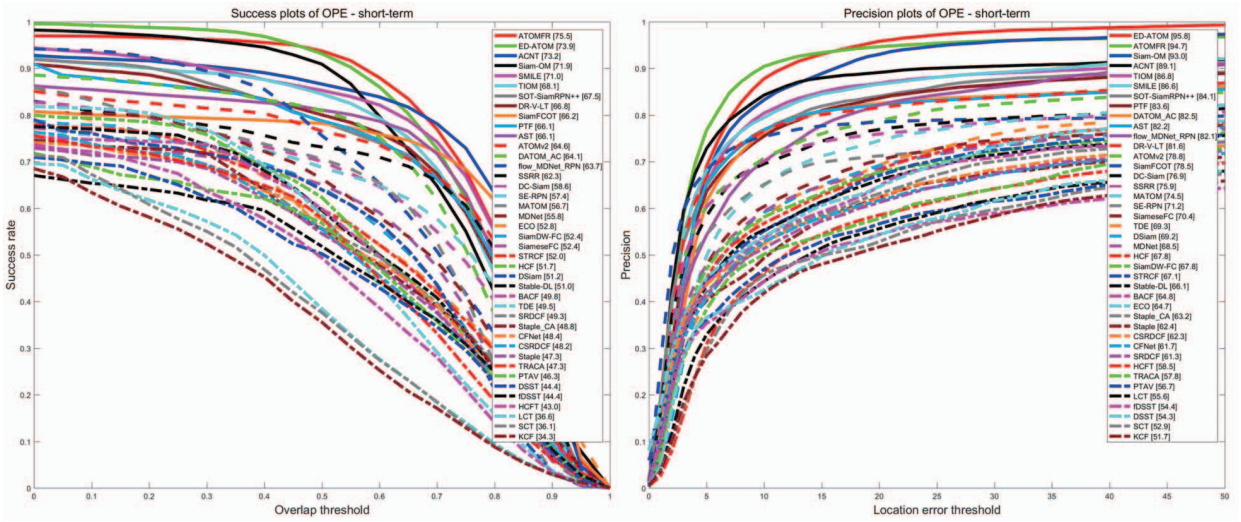

Figure 3. The success and precision plots of the submitted trackers in term of short-term tracking. The success and precision scores for each tracker are presented in the legend.

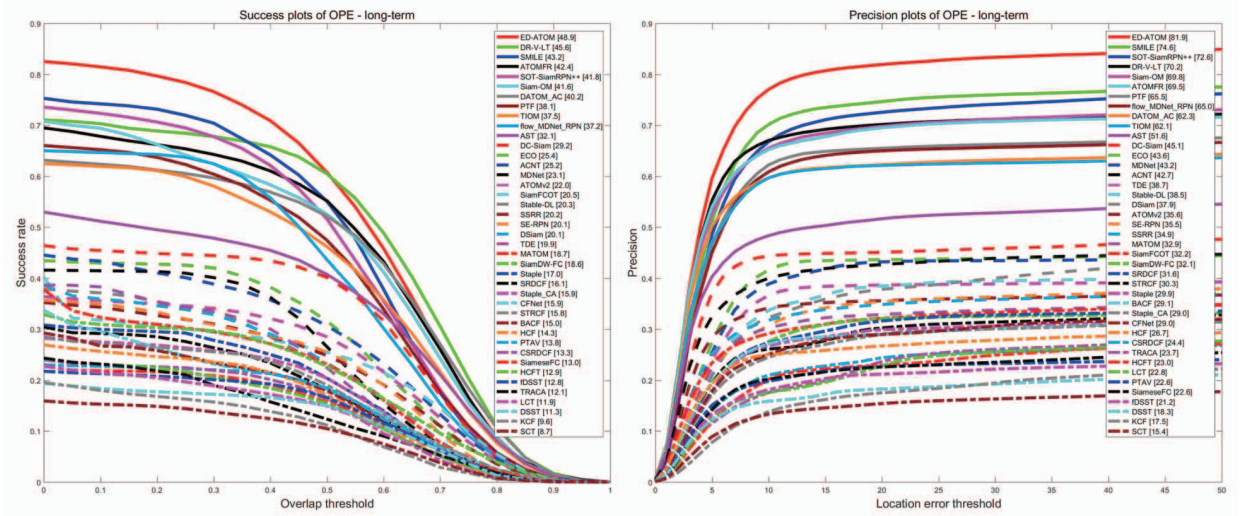

Figure 4. The success and precision plots of the submitted trackers in term of long-term tracking. The success and precision scores for each tracker are presented in the legend.

exist challenging factors such as camera motion, fast motion and object blur. The feature matching methods (SIFT [31] and SURF [1]) and affine transformation are effective to capture the targets in two consecutive frames, see ATOMv2 (A.4), HCF (A.10), PTF (A.12), SMILE (A.17) and TIOM (A.22).

- **Long-term Tracking.** Compared with short-term tracking, we should pay more attention on object verification to reduce target drifting. For example, both DC-Siam (A.6) and DR-V-LT (A.7) design two-stage network to learn the discriminative representation of the target. Besides, the object detector can be used to re-detect the target after partial/full occlusion (see TDE (A.21)).

5. Conclusion

In this paper, we review the VisDrone-SOT2019 challenge, which is second workshop in conjunction with ICCV 2019, to discuss state-of-the-art tracking performance evaluation in drone based scenes, following the successful VisDrone-SOT2018. The testing set of the VisDrone-SOT2019 dataset is expanded from that of the VisDrone-SOT2018 dataset by adding more challenge sequences in long-term tracking. 22 tracking algorithms are submitted to this challenge, the majority of which are improved from exiting trackers such as ATOM [9] and SiamRPN++ [27]. The

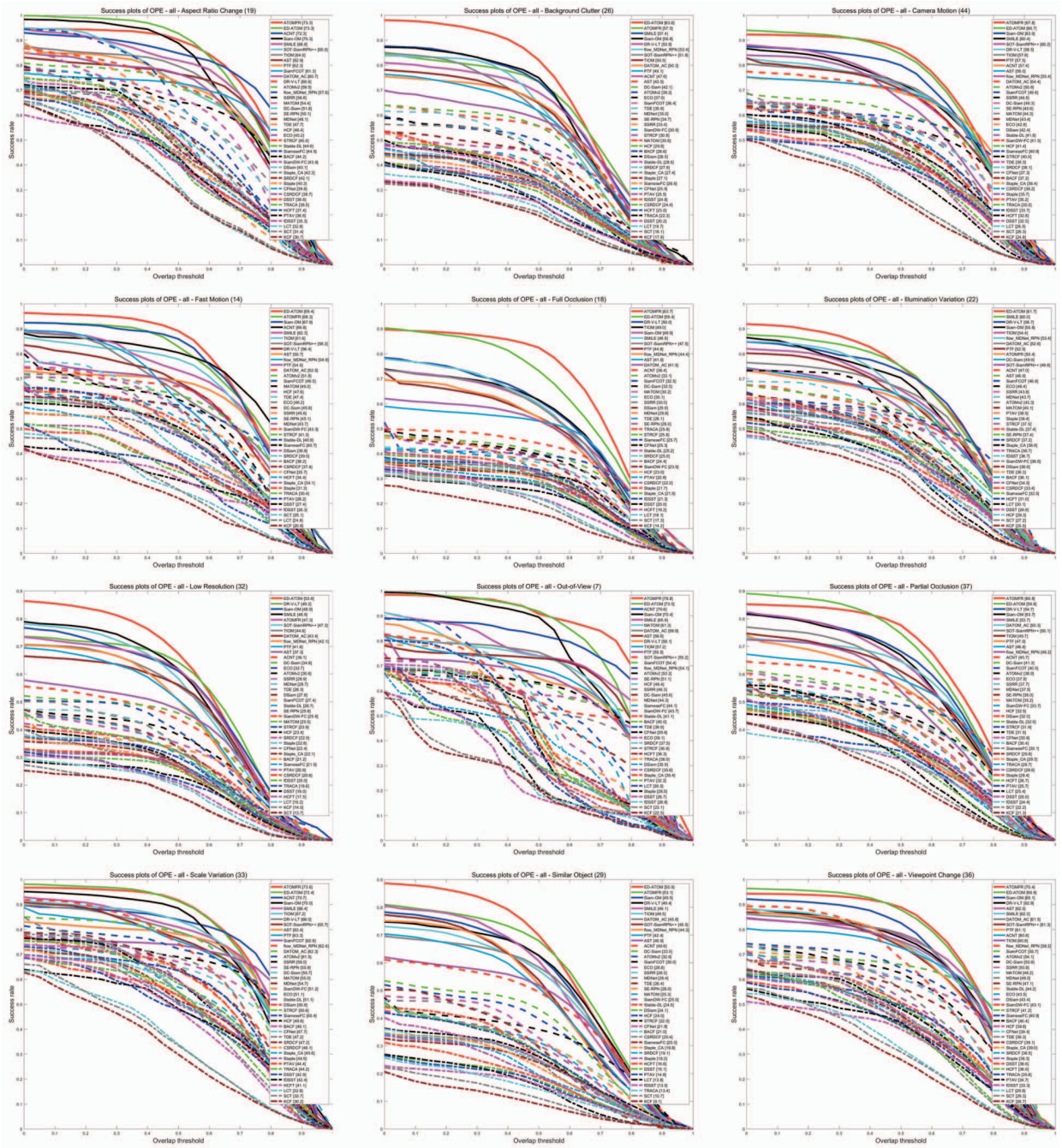

Figure 5. The success plots for the submitted trackers in different attributes. The number presented in the title indicates the number of sequences with that attribute.

top three performers are ED-ATOM (A.8), ATOMFR (A.3), and SMILE (A.17), achieving 63.5, 61.7 and 59.4 success score respectively. We believe VisDrone has become a comprehensive platform for study of object detection and tracking in drones. For future work, we will expand both the dataset for more complex scenarios and the real-time evaluation for practical applications.

Acknowledgements

This work was supported by the National Natural Science Foundation of China under Grants 61502332, 61876127 and 61732011, Natural Science Foundation of Tianjin Under Grants 17JCZDJC30800, Key Scientific and Technological Support Projects of Tianjin Key R&D Pro-

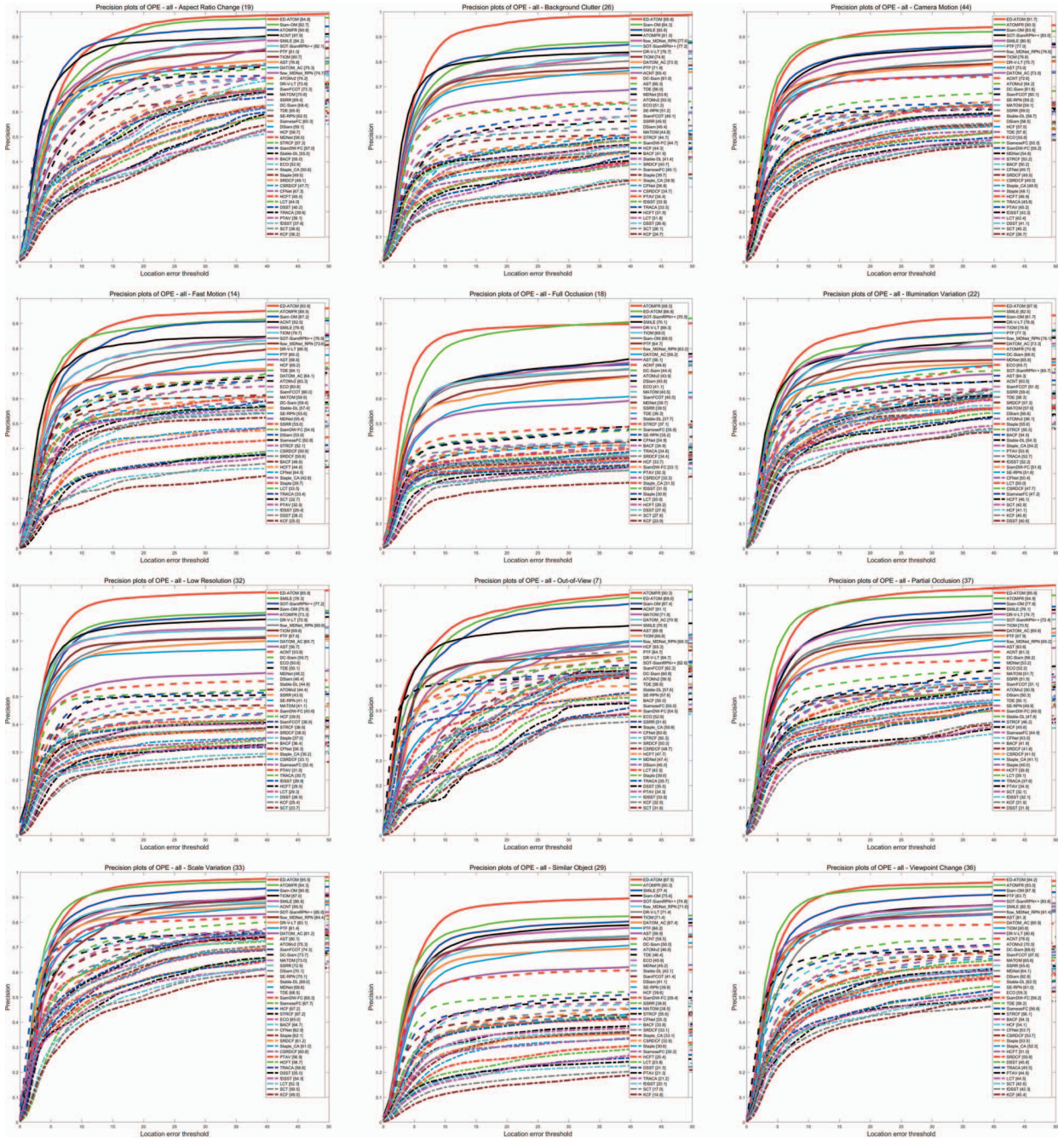

Figure 6. The precision plots for the submitted trackers in different attributes. The number presented in the title indicates the number of sequences with that attribute.

gram 18YFZCGX00390 and 18YFZCGX00680 and JD Digits.

A. Submitted Trackers

In the appendix, we summarize 22 trackers submitted in the VisDrone2019-SOT Challenge, which are ordered alphabetically.

A.1. Adaptive Correction Network based tracker (ACNT)

Tianyang Xu, Xiaojun Wu, Zhenhua Feng and Josef Kittler
tianyang_xu@163.com

ACNT is a two-stage adaptive correction network based

correlation filter tracker. First, the correlation filter is employed to predict the centre location with fixed Resnet-50 (imagenet-classification-trained) feature. Second, an IoU-Distance net (pairwise-trained) is used to optimise the bounding box. The training stage is similar as ATOM [9], except we reformulate the loss function by simultaneously considering Bounding Box Overlap and Centre Distance (IoU-Distance net).

A.2. More Accurate and stable for tracking (AST)

Kang Yang, Xianhai Wang, Ning Wang, Jiaqing Fan and Kaihua Zhang
{779760348, 1719256598, 1098069058, 1296870572, 360572857}@qq.com

AST is improved from ATOM [9]. During the experiment, we find that the results of each evaluation fluctuate to a certain extent because of Gaussian distribution of the bounding box. Therefore we add the novel attention model for robust representation. At the same time, we use the deformable convolution to align the tracking frame with the first frame. Through the experiment we find this alignment strategy is better than the Spatial Transformer Networks [25].

A.3. Accurate Tracking by Overlap Maximization and Feature Recalibration (ATOMFR)

Wenhua Zhang, Haoran Wang and Jinliu Zhou
zhangwenhua_nuc@163.com,
{wanghaoran, zhoujinliu}@stu.xidian.edu.cn

ATOMFR enhances the performance of ATOM [9] by embeding the Squeeze-and-Excitation blocks [23] into IoUNet in ATOM. Our motivation is to explicitly model the interdependencies within feature channels. In addition, we do not intend to introduce a new spatial dimension for the fusion of feature channels but a new Feature Recalibration strategy. Specifically, it is learned to automatically acquire the importance of each feature channel, and then according to this importance to enhance useful features and suppress features that are of little use to the current task. Compared to ATOM, our method is better in overall performance when the tracking environment is complex.

A.4. Accurate Tracking with Overlap Minimization, version 2 (ATOMv2)

Lianghua Huang, Xin Zhao and Kaiqi Huang
huanglianghua2017@ia.ac.cn,
{xzhao, kaiqi.huang}@nlpr.ia.ac.cn

ATOMv2 is improved from the ATOM tracker [9]. Our modifications are: 1) We estimate camera motion across frames using features matching (SURF features [1]) and homography transformation matrix estimation. In this way, the searching area is more robust to severe camera motion. 2)We use thresholding to determine whether to update the model or not.

A.5. Deeper spatio-temporal context aware ATOM with channel attention (DATOM_AC)

Xizhe Xue, Xiaoyue Yin, Shanrong Zou and Ying Li
{xuexizhe, 2015302412, shanrongzou}@mail.nwpu.edu.cn, lybyp@nwpu.edu.cn

DATOM_AC is based on the ATOM tracker [9], channel attention and spatio-temporal context information. The proposed tracker introduces channel attention into the network design to enhance feature representation learning. Specifically, a Squeeze-and-Excitation block [23] is coupled to each convolutional layer to generate channel attention. Channel attention reflects the channel-wise importance of each feature channel and is used for feature weighting in online tracking. Besides, to make our algorithm adapt to appearance changes of target, we decide to capture the spatio-temporal context information. We propose the spatio-temporal branch to learn the feature of previous frame, which can also be utilized to estimate the location of target in current frame. By fusing the results from spatio-temporal and reference branches, we obtain our final result. Both the channel attention and spatio-temporal information are computationally lightweight and impose only a slight increase in model complexity. Our tracker is pre-trained on the ILSVRC2015 dataset [41] and fine-tuned on the VisDrone2018 train set.

A.6. Learning Discriminative Classification for Siamese Visual Tracking (DC-Siam)

Jinghao Zhou and Peng Wang
jensen.zhoujh@gmail.com, peng.wang@nwpu.edu.cn

DC-Siam formulates the visual tracking problem as a regression task by the Siamese network based tracking algorithms. Specifically, it is regarded as a cross-correlation problem by learning a similarity map from deep models with embedding space, where one branch for learning the feature representation of the target, and the other one for the search area. We complement our siamese regression network with a classification module which is a 2-layer fully convolutional neural network. As for the online updating technique, instead of using the brute-force stantdard gradient descent (SGD), we use a more sophisticated optimization strategy that is better suited for online learning considering a quick convergence speed. Detailed information can be found in [9]. We train our network on the training set of COCO [30], ImageNet DET/VID [41], YouTubeBB Dataset [39] and VisDrone2019.

A.7. Distractor-Reduction and Verification Networks for Long-term Visual Object Tracking (DR-V-LT)

Shiyu Xuan and Shengyang Li
{xuanshiyu17, shyli}@csu.ac.cn

DR-V-LT adds the distractor-aware and veritification network based on the SiamRPN++ [27]. SiamRPN++ correctly distinguishs the foreground and background but lacks the ability to distinguish similar objects. In the long-term situations, when the tracked object is lost, the algorithm is very easy to drift to the similar objects. The verification network is used to avoid the algorithm tracking the wrong similar object. Moreover, we propose a two-stage distractor-reduction method. In the distractor-aware stage, we use the score of SiamRPN++ to find the similar objects. In the distractor-suppressive stage, the similar objects are used as the negative sample to update the verification network online. The score of the similar object is suppressive through this way. The architecture of network is the same as SiamRPN++, and the verification network is the same as MDNet [38].

A.8. Accurate target state estimation for drone tracking (ED-ATOM)

Chunhui Zhang, Shengwei Zhao, Kangkai Zhang, Shikun Li, Hao Wen and Shiming Ge
{zhangchunhui,zhaoshengwei,zhangkangkai, lishikun}@iie.ac.cn, wenhao@cloudwalk.cn, geshiming@iie.ac.cn

Our method is following [9], which contains two modules: target estimation and object classification. The target estimation network is performed by the IOU-predictor network. We also use the ResNet-18 trained on ImageNet as backbone. We use the pre-trained target estimation model from [9]. The solution pipeline is as following: 1) Train the IOU-predictor network on Imagenet-VID/DET [41], COCO [30], GOT-10k [24] and LaSOT [16]. 2) Use a low-light image enhancement algorithm [54] to process the VisDrone2019 dataset. 3) Fine-tune offline target estimation model on VisDrone2019 dataset and processed VisDrone2019 dataset. 4) An online data augmentation scheme [5, 55] is conducted on the tracking module to facilitate the model generalizability. A simple search strategy [60] based on tracking state is also used to improve robustness. We advise readers to read [9, 4] for more details.

A.9. Flow Guided MDNet with Redetection (flow_MDNet_RPN)

Han Wu, Xueyuan Yang, Yong Yang and Guizhong Liu
{xjtuwh, yxy1995, yy1996}@stu.xjtu.edu.cn, liugz@xjtu.edu.cn

The flow_MDNet_RPN tracker is inspried from MDNet [38] with consideration of temporal coherency. First, PWC-Net [42] extracts flow information and predicts global motion compensation on the pre-frame object state. Then, MDNet generates the tracking bounding box, which is refined by bounding box regression to find tight bounding box enclosing the target. Next, a one-dimensional correlation filter is used to adapt to the target scale changes, and semantic proposal generated by GA-RPN [47] is selected to adapt to the target aspect ratio change so as to improve target localization accuracy. The similarity between the bounding box and the initial target is calculate by the Siamese network [27].

A.10. Hard negative mining for correlation filters in visual tracking (HCF)

Zhuojin Sun, Yong Wang and Chunhui Zhang
harvards@gmail.com, ywang6@uottawa.ca, zhangchunhui@iie.ac.cn

HCF is a robust tracking method in which a hard negative mining scheme is employed in each frame. In addition, a target verification strategy is developed by introducing a peak signal-to-noise ratio (PSNR) criterion [43, 19]. For this challenge, to overcome camera movement scenes, we predict the position of the object by the affine transformation between frames, and then track the object base on the prediction position.

A.11. More Accurate Tracking by Overlap Maximization (MATOM)

Lijun Zhou and Qintao Hu
{zhoulijun16,hqt0099}@mails.ucas.edu.cn

MATOM is an improved version of the popular tracker ATOM [9] with Kalman Filter and YOLOv3 [40] object detection algorithm. Compared with original ATOM tracker, our algorithm can increase tracking robustness by predicting trajectory with Kalman filter. Simultaneously, if both Kalman filter and ATOM tracker fail to track the object, the tracking results are corrected by YOLOv3 results.

A.12. Preferred Tracking Framework for Large-Scale Dataset with Shaking and Occlusion (PTF)

Ruohan Zhang, Jie Chen, Jie Gao, Xiaoxue Li and Lingling Shi
{ruohan950427,chenjie818826}@163.com, gaojie_jiangsu@126.com, xxli_3@stu.xidian.edu.cn, llshi_1@stu.xidian.edu.cn

PTF is the preferred tracking framework for large-scale dataset with shaking and occlusion to solve the single object tracking problem. Specifically, the main part of our framework comes from MobileNet-based tracking by detection algorithm (MBMD) [56]. By analyzing the whole processing of MBMD object tracking, it is clear that there are vigorous changing of object boxes' location. In order to solve this problem, ECO [8] and ATOM are used for later adjustment. We find that the results of ECO cannot deal with the shaking problem, so use the SIFT features and regard the process of ECO and SIFT as ECOO. Then, the results of ECOO and ATOM are used to revise the MBMD. After that, a preferred processing, which is based on an exact threshold, is performed to get the final results.

A.13. Learning Equivariance: Siamese Equivariant Region Proposal Network for Accurate Online Object Tracking (SE-RPN)

Xu Lei and Jinwang Wang
{leixuchn,jwwangchn}@whu.edu.cn

SE-RPN is the Siamese equivariant region proposal network for accurate online object tracking. Specifically, we leverage the equivariant property to guide the anchoring, and learn equivariance in the correlation mechanism of SiamRPN [28]. By reformulating the anchoring mechanism within the SiamRPN tracking framework, our algorithm not only provides better initialization for region proposal, but also mitigates the misalignment problems in the correlation.

A.14. Deeper and Wider Siamese Networks for Real-Time Visual Tracking (SiamDW-FC)

Zhipeng Zhang and Weiming Hu
zhangzhipeng2017@ia.ac.cn, wmhu@nlpr.ia.ac.cn

SiamDW-FC improves the original SiamFC-based model [57] by leveraging deeper and wider convolutional neural networks to enhance tracking robustness and accuracy. Direct replacement of backbones with existing powerful architectures, such as ResNet and Inception, does not bring improvements. The main reasons are that 1) large increases in the receptive field of neurons lead to reduced feature discriminability and localization precision; and 2) the network padding for convolutions induces a positional bias in learning. To address these issues, we propose new residual modules to eliminate the negative impact of padding, and further design new architectures using these modules with controlled receptive field size and network stride. The designed architectures are lightweight and guarantee real-time tracking speed when applied to SiamFC [57] and SiamRPN [28].

A.15. Fully Convolutional method for Object Tracking (SiamFCOT)

Yinda Xu and Zeyu Wang
yinda_xu@zju.edu.cn, wangzeyu0408@outlook.com

SiamFCOT is capable to perform efficient tracking while reaches a high accuracy. The proposed tracker consists of the following components: feature extraction, feature matching, head network, mask refinement and post-processing. Inspired by the recently emerging anchor-free detection technique, we adopt the head architecture and box encoding/decoding protocol of FCOS detector [45]. A post-processing procedure of SiamRPN-style [28] is performed on the regressed candidate boxes with their correspondence scores to generate the unique final box as the tracking result on the current frame. We train the tracker based on Imagenet-VID/DET [41], COCO [30], YoutubeBB [39], LaSOT [16], and GOT-10k [24]. Note that the pipeline and model has not been specifically fine-tuned for VisDrone task, so there still remains potential to exploit for future work.

A.16. Combination of DaSiam and ATOM (Siam-OM)

Xin Zhang, Xiaotong Li and Jie Zhang
{xinzhang1,lixiaotong}@stu.xidian.edu.cn,
1437614843@qq.com

Siam-OM deals with the video tracking sequences in two cases based on the number of video frames. If one sequence has more than $3,000$ frames, we classify it as long sequence, otherwise, short. For short sequences, we use the ATOM framework [9]. To improve the recognizability of the target object, we enhance the original input image using Gamma non-linear correction, which improves the tracking performance greatly in poor lighting conditions. At the same time, we reduce the rate of hard negative learning, which makes the tracker more robust in the case of short-term occlusion. For long sequences, we use the DaSiam [60] framework with ResNet structure [27]. To solve the problem that the target always switches back and forth, we use the sift feature matching algorithm [31] to calculate the offset of the target between the current frame and the previous frame.

A.17. Strategy and Motion Integrated Long-term Experts (SMILE)

Ruiyan Ma, Yanjie Gao, Yuting Yang, Wei Song and Yuxuan Li
3028408083@qq.com, yjgao@stu.xidian.edu.cn,
ytyang_1@stu.xidian.edu.cn, 522545707@qq.com,
liyuxuan_xidian@126.com

SMILE combines two state-of-the-art trackers including ATOM [9] and SiamRPN++ [27]. Our method makes the systems more reliable respectively as different features play different role in the process of tracking based on their reliability. In addition, we improve the prediction of blurred scenes by using the SIFT algorithm [31] to match features. By estimating motion, the regression boxes can continue tracking the target in case of occlusion. When encountering dark or low-resolution scenes, we use threshold judgement and image brightness enhancement processing.

A.18. An improved SiamRPN++ algorithm based on sift matching algorithm, OpticalFlow-PyrLK and Template Self-calibration (SOT-SiamRPN++)

Zhizhao Duan, Wenjun Zhu, Xi Yu, Bo Han, Zhiyong Yu and Ting He
{21825106,21810114,21810157,21810207}@zju.edu.cn, yuk1062@163.com, the@zju.edu.cn

SOT-SiamRPN++ is improved from SiamRPN++ [27]. In order to improve the tracking performance for small targets, the original images are twice magnified. At training stage, the magnified images are cut as the same size as the original image based on the center of bounding boxes, with some basic data augmentation. At testing stage, the cut work are based on the center of bounding boxes in last frame. To overcome the camera shaking, we use the idea of template self-calibration, we modify the the box in last frame and fix it in the center of the target. The bounding box will not change in the template self-calibration algorithm to avoid the box labeling in the wrong target. To solve slight occlusion, we adopt the method of optical flow and tracker running simultaneously. The input of optical flow is the picture of the previous frame and the current frame, and the location of optical flow feature points extracted from the tracker box of the previous frame. The output is to predict the position of the current frame optical flow feature points.When encountering occlusion, the position of the tracker becomes inaccurate, and the position marked by optical flow feature points is taken as the output. For night scene, we first use Laplacian operator to augment the contrast of images, and then the SiamRPN++ tracker can track the target by combining optical flow. We also use the SIFT to match the target when the SiamRPN++ fails to track the target. Our training sets are ImageNet VID/DET [41], YoutubeBB [39], COCO [30], and the official VisDrone2018 training set [51].

A.19. A Self-adaptive Search Region and Re-ID object tracking method (SSRR)

Ning Wang and Kaihua Zhang
20181222016@nuist.edu.cn

SSRR is improved from the ATOM algorithm [9]. First, we design a self-adaptive searching region based on the motion speed of the target. Then, we add a Re-ID tracking module to recognize when the target is lost. Specifically, we use ResNet-34 to calculate appearance feature, as well as discriminative correlation filters and IOUNet to determine the location and bounding box of the target. The network is fine-tuned on LaSOT [16], COCO [30] and VisDrone2019 train-set.

A.20. Stable responsibility based deep learning tracker (Stable-DL)

Yong Wang, Lu Ding, Robert Laganière, Jiuqing Wan and Wei Shi
ywang6@uottawa.ca, dinglu@sjtu.edu.cn, laganier@eecs.uottawa.ca, wanjiuqing@buaa.edu.cn, weishi_insky@126.com

Stable-DL is a novel stable responsibility based tracking method that uses two deep layers of VGG-19 backbone to extract features. These two tracking results from two layers are used to compute a stable responsibility which is a metric to evaluate the quality of tracking results. The final result is fused by the two tracking results and stable responsibility.

A.21. Tracking and detection: a unified approach (TDE)

Chunhui Zhang, Shengwei Zhao, Zhuojin Sun, Yong Wang and Shiming Ge
{zhangchunhui, zhaoshengwei, geshiming}@iie.ac.cn, harvards@gmail.com, ywang6@uottawa.ca

The TDE tracker unifies tracking and detection in a simple way, achieving high performance general object tracking. It mainly consists of two parts: the discriminative correlation filter and detection module. Specifically, we use the LADCF tracker [53] as the tracking module and the Yolov3 detector [40] as the detection module. Besides, we use a tracking failure measurement method like [48] to decide when and how to refine tracking result according to detecting result. If the refinement conditions are met, we conduct adaptive tracking by weighting both the tracking and detection results.

A.22. Tracking by Improved Overlap Maximization (TIOM)

Shengyin Zhu and Yanyun Zhao
{pansiyang, zyy}@bupt.edu.cn

TIOM is built upon Accurate Tracking by Overlap Maximization (ATOM) [9]. We have made some modifi-

cations to ATOM. First, we use generalized intersection over union (GIoU) to replace traditional IoU in the target estimation network. Second, to prevent camera shake, we use the surf feature matching method [1] to calculate the offset of the target between two consecutive frames to locate the target position in the current frame correctly. In addition, to deal with long-term occlusion, we enlarge search area gradually and use Kalman Filter to predict motion trajectory of the target if occlusion is detected. We use ResNet-18 pretrained on ImageNet as our backbone network and fine-tune it on the VisDrone2018 train set [51] and LaSOT [16].

References

[1] H. Bay, T. Tuytelaars, and L. V. Gool. SURF: speeded up robust features. In *ECCV*, pages 404–417, 2006.

[2] L. Bertinetto, J. Valmadre, S. Golodetz, O. Miksik, and P. H. S. Torr. Staple: Complementary learners for real-time tracking. In *CVPR*, pages 1401–1409, 2016.

[3] L. Bertinetto, J. Valmadre, J. F. Henriques, A. Vedaldi, and P. H. S. Torr. Fully-convolutional siamese networks for object tracking. In *ECCVW*, pages 850–865, 2016.

[4] G. Bhat, M. Danelljan, L. V. Gool, and R. Timofte. Learning discriminative model prediction for tracking. *CoRR*, abs/1904.07220, 2019.

[5] G. Bhat, J. Johnander, M. Danelljan, F. S. Khan, and M. Felsberg. Unveiling the power of deep tracking. In *ECCV*, pages 493–509, 2018.

[6] J. Choi, H. J. Chang, T. Fischer, S. Yun, K. Lee, J. Jeong, Y. Demiris, and J. Y. Choi. Context-aware deep feature compression for high-speed visual tracking. In *CVPR*, pages 479–488, 2018.

[7] J. Choi, H. J. Chang, J. Jeong, Y. Demiris, and J. Y. Choi. Visual tracking using attention-modulated disintegration and integration. In *CVPR*, pages 4321–4330, 2016.

[8] M. Danelljan, G. Bhat, F. S. Khan, and M. Felsberg. ECO: efficient convolution operators for tracking. In *CVPR*, pages 6931–6939, 2017.

[9] M. Danelljan, G. Bhat, F. S. Khan, and M. Felsberg. ATOM: accurate tracking by overlap maximization. *CoRR*, abs/1811.07628, 2018.

[10] M. Danelljan, G. Häger, F. S. Khan, and M. Felsberg. Accurate scale estimation for robust visual tracking. In *BMVC*, 2014.

[11] M. Danelljan, G. Häger, F. S. Khan, and M. Felsberg. Learning spatially regularized correlation filters for visual tracking. In *ICCV*, pages 4310–4318, 2015.

[12] M. Danelljan, G. Häger, F. S. Khan, and M. Felsberg. Discriminative scale space tracking. *TPAMI*, 39(8):1561–1575, 2017.

[13] D. Du, H. Qi, L. Wen, Q. Tian, Q. Huang, and S. Lyu. Geometric hypergraph learning for visual tracking. *TCYB*, 47(12):4182–4195, 2017.

[14] D. Du, Y. Qi, H. Yu, Y. Yang, K. Duan, G. Li, W. Zhang, Q. Huang, and Q. Tian. The unmanned aerial vehicle benchmark: Object detection and tracking. In *ECCV*, pages 375–391, 2018.

[15] D. Du, L. Wen, H. Qi, Q. Huang, Q. Tian, and S. Lyu. Iterative graph seeking for object tracking. *TIP*, 27(4):1809–1821, 2018.

[16] H. Fan, L. Lin, F. Yang, P. Chu, G. Deng, S. Yu, H. Bai, Y. Xu, C. Liao, and H. Ling. Lasot: A high-quality benchmark for large-scale single object tracking. *CoRR*, abs/1809.07845, 2018.

[17] H. Fan and H. Ling. Parallel tracking and verifying: A framework for real-time and high accuracy visual tracking. In *ICCV*, pages 5487–5495, 2017.

[18] H. K. Galoogahi, A. Fagg, C. Huang, D. Ramanan, and S. Lucey. Need for speed: A benchmark for higher frame rate object tracking. In *ICCV*, pages 1134–1143, 2017.

[19] H. K. Galoogahi, A. Fagg, and S. Lucey. Learning background-aware correlation filters for visual tracking. In *ICCV*, pages 1144–1152, 2017.

[20] Q. Guo, W. Feng, C. Zhou, R. Huang, L. Wan, and S. Wang. Learning dynamic siamese network for visual object tracking. In *ICCV*, pages 1781–1789, 2017.

[21] D. Held, S. Thrun, and S. Savarese. Learning to track at 100 FPS with deep regression networks. In *ECCV*, pages 749–765, 2016.

[22] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista. High-speed tracking with kernelized correlation filters. *TPAMI*, 37(3):583–596, 2015.

[23] J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks. In *CVPR*, 2018.

[24] L. Huang, X. Zhao, and K. Huang. Got-10k: A large high-diversity benchmark for generic object tracking in the wild. *CoRR*, abs/1810.11981, 2018.

[25] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu. Spatial transformer networks. In *NeurIPS*, pages 2017–2025, 2015.

[26] M. Kristan, A. Leonardis, J. Matas, M. Felsberg, R. P. Pflugfelder, L. C. Zajc, T. Vojír, G. Bhat, A. Lukezic, A. Eldesokey, G. Fernández, and *et al.* The sixth visual object tracking VOT2018 challenge results. In *ECCV Workshops*, pages 3–53, 2018.

[27] B. Li, W. Wu, Q. Wang, F. Zhang, J. Xing, and J. Yan. Siamrpn++: Evolution of siamese visual tracking with very deep networks. *CoRR*, abs/1812.11703, 2018.

[28] B. Li, J. Yan, W. Wu, Z. Zhu, and X. Hu. High performance visual tracking with siamese region proposal network. In *CVPR*, pages 8971–8980, 2018.

[29] F. Li, C. Tian, W. Zuo, L. Zhang, and M. Yang. Learning spatial-temporal regularized correlation filters for visual tracking. In *CVPR*, pages 4904–4913, 2018.

[30] T. Lin, M. Maire, S. J. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: common objects in context. In *ECCV*, pages 740–755, 2014.

[31] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004.

[32] A. Lukezic, T. Vojir, L. C. Zajc, J. Matas, and M. Kristan. Discriminative correlation filter with channel and spatial reliability. In *CVPR*, pages 4847–4856, 2017.

[33] C. Ma, J. Huang, X. Yang, and M. Yang. Hierarchical convolutional features for visual tracking. In *ICCV*, pages 3074–3082, 2015.

[34] C. Ma, X. Yang, C. Zhang, and M. Yang. Long-term correlation tracking. In *CVPR*, pages 5388–5396, 2015.

[35] M. Mueller, N. Smith, and B. Ghanem. A benchmark and simulator for UAV tracking. In *ECCV*, pages 445–461, 2016.

[36] M. Mueller, N. Smith, and B. Ghanem. Context-aware correlation filter tracking. In *CVPR*, pages 1387–1395, 2017.

[37] M. Müller, A. Bibi, S. Giancola, S. Al-Subaihi, and B. Ghanem. Trackingnet: A large-scale dataset and benchmark for object tracking in the wild. In *ECCV*, pages 310–327, 2018.

[38] H. Nam and B. Han. Learning multi-domain convolutional neural networks for visual tracking. In *CVPR*, pages 4293–4302, 2016.

[39] E. Real, J. Shlens, S. Mazzocchi, X. Pan, and V. Vanhoucke. Youtube-boundingboxes: A large high-precision human-annotated data set for object detection in video. In *CVPR*, pages 7464–7473, 2017.

[40] J. Redmon and A. Farhadi. Yolov3: An incremental improvement. *CoRR*, abs/1804.02767, 2018.

[41] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. S. Bernstein, A. C. Berg, and F. Li. Imagenet large scale visual recognition challenge. *IJCV*, 115(3):211–252, 2015.

[42] D. Sun, X. Yang, M. Liu, and J. Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *CVPR*, pages 8934–8943, 2018.

[43] Z. Sun, Y. Wang, and R. Laganière. Hard negative mining for correlation filters in visual tracking. *MVA*, 30(3):487–506, 2019.

[44] R. Tao, E. Gavves, and A. W. M. Smeulders. Siamese instance search for tracking. In *CVPR*, pages 1420–1429, 2016.

[45] Z. Tian, C. Shen, H. Chen, and T. He. FCOS: fully convolutional one-stage object detection. *CoRR*, abs/1904.01355, 2019.

[46] J. Valmadre, L. Bertinetto, J. F. Henriques, A. Vedaldi, and P. H. S. Torr. End-to-end representation learning for correlation filter based tracking. In *CVPR*, pages 5000–5008, 2017.

[47] J. Wang, K. Chen, S. Yang, C. C. Loy, and D. Lin. Region proposal by guided anchoring. *CoRR*, abs/1901.03278, 2019.

[48] M. Wang, Y. Liu, and Z. Huang. Large margin object tracking with circulant feature maps. In *CVPR*, pages 4800–4808, 2017.

[49] Q. Wang, J. Gao, J. Xing, M. Zhang, and W. Hu. Dcfnet: Discriminant correlation filters network for visual tracking. *CoRR*, abs/1704.04057, 2017.

[50] Q. Wang, L. Zhang, L. Bertinetto, W. Hu, and P. H. S. Torr. Fast online object tracking and segmentation: A unifying approach. *CoRR*, abs/1812.05050, 2018.

[51] L. Wen, P. Zhu, D. Du, X. Bian, H. Ling, Q. Hu, and *et al.* Visdrone-sot2018: The vision meets drone single-object tracking challenge results. In *ECCVW*, pages 469–495, 2018.

[52] Y. Wu, J. Lim, and M. Yang. Object tracking benchmark. *TPAMI*, 37(9):1834–1848, 2015.

[53] T. Xu, Z. Feng, X. Wu, and J. Kittler. Learning adaptive discriminative correlation filters via temporal consistency preserving spatial feature selection for robust visual tracking. *CoRR*, abs/1807.11348, 2018.

[54] Z. Ying, G. Li, Y. Ren, R. Wang, and W. Wang. A new low-light image enhancement algorithm using camera response model. In *ICCV Workshops*, pages 3015–3022, 2017.

[55] C. Zhang, S. Ge, Y. Hua, and D. Zeng. Robust deep tracking with two-step augmentation discriminative correlation filters. In *ICME*, 2019.

[56] Y. Zhang, D. Wang, L. Wang, J. Qi, and H. Lu. Learning regression and verification networks for long-term visual tracking. *CoRR*, abs/1809.04320, 2018.

[57] Z. Zhang, H. Peng, and Q. Wang. Deeper and wider siamese networks for real-time visual tracking. *CoRR*, abs/1901.01660, 2019.

[58] P. Zhu, L. Wen, D. Du, X. Bian, H. Ling, Q. Hu, and *et al.* Visdrone-det2018: The vision meets drone object detection in image challenge results. In *ECCVW*, pages 437–468, 2018.

[59] P. Zhu, L. Wen, D. Du, X. Bian, H. Ling, Q. Hu, and *et al.* Visdrone-vdt2018: The vision meets drone video detection and tracking challenge results. In *ECCVW*, pages 496–518, 2018.

[60] Z. Zhu, Q. Wang, B. Li, W. Wu, J. Yan, and W. Hu. Distractor-aware siamese networks for visual object tracking. In *ECCV*, pages 103–119, 2018.