VisDrone-SOT2018: The Vision Meets Drone Single-Object Tracking Challenge Results

Longyin Wen[1], Pengfei Zhu[2]✉*, Dawei Du[3], Xiao Bian[4], Haibin Ling[5], Qinghua Hu[2], Chenfeng Liu[2], Hao Cheng[2], Xiaoyu Liu[2], Wenya Ma[2], Qinqin Nie[2], Haotian Wu[2], Lianjie Wang[2], Asanka G Perera[23], Baochang Zhang[8], Byeongho Heo[17], Chunlei Liu[8], Dongdong Li[22], Emmanouil Michail[15], Hanlin Chen[10], Hao Liu[22], Haojie Li[12], Ioannis Kompatsiaris[15], Jian Cheng[28,29], Jiaqing Fan[29], Jie Zhang[21], Jin Young Choi[17], Jing Li[27], Jinyu Yang[8], Jongwon Choi[17,19], Juanping Zhao[6], Jungong Han[9], Kaihua Zhang[30], Kaiwen Duan[14], Ke Song[20], Konstantinos Avgerinakis[15], Kyuewang Lee[17], Lu Ding[6], Martin Lauer[16], Panagiotis Giannakeris[15], Peizhen Zhang[25], Qiang Wang[28], Qianqian Xu[31], Qingming Huang[13,14], Qingshan Liu[30], Robert Laganière[7], Ruixin Zhang[24], Sangdoo Yun[18], Shengyin Zhu[11], Sihang Wu[12], Stefanos Vrochidis[15], Wei Tian[16], Wei Zhang[20], Weidong Chen[14], Weiming Hu[28], Wenhao Wang[20], Wenhua Zhang[21], Wenrui Ding[8], Xiaohao He[26], Xiaotong Li[21], Xin Zhang[21], Xinbin Luo[6], Xixi Hu[20], Yang Meng[21], Yangliu Kuai[22], Yanyun Zhao[11], Yaxuan Li[20], Yifan Yang[14], Yifan Zhang[28,29], Yong Wang[7], Yuankai Qi[13], Zhipeng Deng[22], and Zhiqun He[11]

[1]JD Finance, Mountain View, CA, USA.
[2]Tianjin University, Tianjin, China.
[3]University at Albany, SUNY, Albany, NY, USA.
[4]GE Global Research, Niskayuna, NY, USA.
[5]Temple University, Philadelphia, PA, USA.
[6]Shanghai Jiao Tong University, China.
[7]University of Ottawa, Canada.
[8]Beihang University, China.
[9]Lancaster University, UK.
[10]Shenyang Aerospace University, China.
[11]Beijing University of Posts and Telecommunications, China.
[12]South China University of Technology, China.
[13]Harbin Institute of Technology, China.
[14]University of Chinese Academy of Sciences, China.
[15]Centre for Research & Technology Hellas, Greece.
[16]Karlsruhe Institute of Technology, Germany.
[17]Seoul National University, South Korea.
[18]NAVER Corp, South Korea.
[19]Samsung R&D Campus, South Korea.
[20]Shandong University, China.
[21]Xidian University, China.
[22]National University of Defense Technology, China.
[23]University of South Australia, Australia.
[24]Tencent, China.
[25]Sun yat-sen university, China.
[26]Tsinghua University, China.

[27]Civil Aviation University Of China, China.
[28]Institute of Automation, Chinese Academy of Sciences, China.
[29]Nanjing Artificial Intelligence Chip Research, Institute of Automation, Chinese Academy of Sciences, China.
[30]Nanjing University of Information Science and Technology, China.
[31]Institute of Computing Technology, Chinese Academy of Sciences.

Abstract. Single-object tracking, also known as visual tracking, on the drone platform attracts much attention recently with various applications in computer vision, such as filming and surveillance. However, the lack of commonly accepted annotated datasets and standard evaluation platform prevent the developments of algorithms. To address this issue, the Vision Meets Drone Single-Object Tracking (VisDrone-SOT2018) Challenge workshop was organized in conjunction with the 15th European Conference on Computer Vision (ECCV 2018) to track and advance the technologies in such field. Specifically, we collect a dataset, including 132 video sequences divided into three non-overlapping sets, *i.e.*, training (86 sequences with 69,941 frames), validation (11 sequences with 7,046 frames), and testing (35 sequences with 29,367 frames) sets. We provide fully annotated bounding boxes of the targets as well as several useful attributes, *e.g.*, occlusion, background clutter, and camera motion. The tracking targets in these sequences include pedestrians, cars, buses, and animals. The dataset is extremely challenging due to various factors, such as occlusion, large scale, pose variation, and fast motion. We present the evaluation protocol of the VisDrone-SOT2018 challenge and the results of a comparison of 22 trackers on the benchmark dataset, which are publicly available on the challenge website: `http://www.aiskyeye.com/`. We hope this challenge largely boosts the research and development in single object tracking on drone platforms.

Keywords: Performance evaluation, drone, single-object tracking.

1 Introduction

Drones, or general UAVs, equipped with cameras have been fast deployed to a wide range of applications, including agricultural, aerial photography, fast delivery, and surveillance. Consequently, automatic understanding of visual data collected from drones becomes highly demanding, which makes computer vision and drones more and more closely. Despite the great progresses in general computer vision algorithms, such as tracking and detection, these algorithms are not usually optimal for dealing with sequences or images generated by drones, due to various challenges such as view point change and scales.

Developing and evaluating new vision algorithms for drone generated visual data is a key problem in drone-based applications. However, as pointed out

⋆ Email address: zhupengfei@tju.edu.cn

in recent studies (*e.g.*, [43, 26]), the lack of public large-scale benchmarks or datasets is the bottleneck to achieve this goal. Some recent preliminary efforts [43, 49, 26] have been devoted to construct datasets with drone platforms focusing on single-object tracking. These datasets are still limited in size and scenarios covered, due to the difficulties in data collection and annotation. Thus, a more general and comprehensive benchmark is desired for further boost research on computer vision problems with drones.

To advance the developments in single-object tracking, we organize the Vision Meets Drone Single-Object Tracking (VisDrone-SOT2018) challenge, which is one track of the "Vision Meets Drone: A Challenge"[1] on September 8, 2018, in conjunction with the 15th European Conference on Computer Vision (ECCV 2018) in Munich, Germany. In particular, we collected a single-object tracking dataset with various drone models, *e.g.*, DJI Mavic, and Phantom series 3, 3A, in different scenarios with various weather and lighting conditions. All video sequences are labelled per-frame with different visual attributes to aid a less biased analysis of the tracking results. The objects to be tracked are of various types including pedestrians, cars, buses, and sheep. We invite the authors to submit the tracking results in the VisDrone-SOT2018 dataset. The authors of submitted algorithms in the challenge have an opportunity to share their ideas in the workshop and further publish the source code at our website: `http://www.aiskyeye.com/`, which are helpful to push the development of the single-object tracking field.

2 Related Work

Single-object tracking or visual tracking, is one of the fundamental problems in computer vision, which aims to estimate the trajectory of a target in a video sequence, given its initial state. In this section, we briefly review the related datasets and recent tracking algorithms.

Existing datasets. In recent years, numerous datasets have been developed for single object tracking. Wu *et al.* [65] create a standard benchmark to evaluate the single-object tracking algorithms, which includes 50 video sequences. After that, they further extend the dataset with 100 video sequences. Concurrently, Liang *et al.* [36] collect 128 video sequences for evaluating the color enhanced trackers. To track the progress in single-object tracking field, Kristan *et al.* [56, 31, 29, 30] organize the VOT competition from 2013 to 2018, where the new datasets and evaluation strategies are proposed for tracking evaluation. The series of competitions promote the developments of visual tracking. Smeulders *et al.* [52] present the ALOV300 dataset, containing 314 video sequences with 14 visual attributes, such as long duration, zooming camera, moving camera and transparency. Li *et al.* [32] construct a large-scale dataset with 365 video sequences, covering 12 different kinds of objects captured from moving cameras. Du *et al.* [15] design a dataset with 50 fully annotated video sequences, focusing on

[1] `http://www.aiskyeye.com/`.

deformable object tracking in unconstrained environments. To evaluate tracking algorithms in high frame rate videos (*e.g.*, 240 frame per second), Galoogahi *et al.* [21] propose a dataset containing 100 video clips (380,000 frames in total), recorded in real world scenarios. Besides using video sequences captured by RGB cameras, Felsberg *et al.* [20, 57, 30] organize a series of competitions from 2015 to 2017, focusing on single-object tracking on thermal video sequences recorded by 8 different types of sensors. In [53], a RGB-D tracking dataset is presented, which includes 100 RGB-D video clips with manually annotated ground truth bounding boxes. UAV123 [43] is a large UAV dataset including 123 fully annotated high-resolution video sequences captured from the low-altitude aerial view points. Similarly, UAVDT [16] describes a new UAV benchmark focusing on several different complex scenarios. Müller *et al.* [45] present a large-scale benchmark for object tracking in the wild, which includes more than 30,000 videos with more than 14 million dense bounding box annotations. Recently, Fan *et al.* [18] propose a large tracking benchmark with 1,400 videos, with each frame manually annotated. Most of the above datasets cover a large set of object categories, but do not focus on drone based scenarios as our dataset.

Review of recent single-object tracking methods. Single-object tracking is a hot topic with various applications (*e.g.*, video surveillance, behavior analysis and human-computer interaction). It attracts much research such as graph model [4, 15, 35, 64], subspace learning [50, 28, 62, 63] and sparse coding [42, 39, 69, 47]. Recently, the correlation filter algorithm becomes popular in visual tracking field due to its high efficiency. Henriques *et al.* [25] derive a kernelized correlation filter and propose a fast multi-channel extension of linear correlation filters using a linear kernel. Danelljan *et al.* [10] propose to learn discriminative correlation filters based on the scale pyramid representation to improve the tracking performance. To model the distribution of feature attention, Choi *et al.* [7] develop an attentional feature-based correlation filter evolved with multiple trained elementary trackers. The Staple method [2] achieves a large gain in performance over previous methods by combining color statistics and correlations. Danelljan *et al.* [11] demonstrate that learning the correlation filter coefficients with spatial regularization is effective for tracking task. Li *et al.* [34] integrate the temporal regularization into the SRDCF framework [11] with single sample, and propose the spatial-temporal regularized correlation filters to provide a more robust appearance model in the case of large appearance variations. Du *et al.* [17] design a correlation filter based method that integrates the target part selection, part matching, and state estimation into a unified energy minimization framework.

On the other hand, the deep learning based methods achieve a dominant position in the single-object tracking field with the impressive performance. Some methods directly use the deep Convolutional Neural Networks (CNNs) to extract the features to replace the hand-crafted features in the correlation filter framework, such as CF2 [41], C-COT [13], ECO [9], CFNet [59], and PTAV [19]. In [60], different types of features are combined to construct multiple experts through discriminative correlation filter algorithm, and each of them tracks the target independently. With the proposed robustness evaluation strategy, the

most confident expert is selected to produce the tracking results in each frame. Besides, another way is to construct an end-to-end deep model to complete target appearance learning and tracking [3, 55, 46, 54, 67, 33, 14]. In SiamFC [3] and SINT [55], the researchers employ siamese deep neural network to learn the matching function between the initial patch of the target in the first frame and the candidate in the subsequent frames. Li *et al.* [33] propose the siamese region proposal network, which consists of a siamese sub-network for feature extraction and a region proposal sub-network for classification and regression. MDNet [46] uses a pre-trained CNN model on a large set of video sequences with manually annotated ground-truths to obtain a generic target representation, and then evaluates the candidate windows randomly sampled around the previous target state to find the optimal location for tracking. After that, Song *et al.* [54] present the VITAL algorithm to generate more discriminative training samples via adversarial learning. Yun *et al.* [67] design a tracker controlled by sequentially pursuing actions learned by deep reinforcement learning. Dong *et al.* [14] propose a hyperparameter optimization method that is able to find the optimal hyperparameters for a given sequence using an action-prediction network leveraged on continuous deep Q-learning.

3 The VisDrone-SOT2018 Challenge

As described above, to track and promote the developments in single-object tracking field, we organized the Vision Meets Drone Single-Object Tracking (or VisDrone-SOT2018, for short) challenge, which is one track of the workshop challenge "Vision Meets Drone: A Challenge" on September 8, 2018, in conjunction with the 15th European Conference on Computer Vision (ECCV 2018) in Munich, Germany. The VisDrone-SOT2018 challenge focuses on single-object tracking on the drone platform. Specifically, given an initial bounding box enclosing the target in the first frame, the submitted algorithm is required to estimate the region of target in the subsequent video frames. We released a single-object tracking dataset, *i.e.*, the VisDrone-SOT2018 dataset, which consists of 132 video sequences formed by 106,354 frames, captured by various drone-mounted cameras, covering a wide range of aspects including location (taken from 14 different cities in China), environment (urban and country), objects (pedestrian, vehicles, bicycles, etc.), and density (sparse to crowded scenes). We invited researchers to participate the challenge and to evaluate and discuss their research on the VisDrone-SOT2018 dataset at the workshop. We believe the workshop challenge will be helpful to the research in the video object tracking community.

3.1 Dataset

The released VisDrone-SOT2018 dataset in this workshop includes 132 video clips with 106,354 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 (35 sequences with 29,367 frames).

Table 1: Comparison of Current State-of-the-Art Benchmarks and Datasets. Note that the resolution indicates the maximum resolution of the video frames included in the dataset. Notably, we have $1k = 1,000$.

datasets	scenarios	#sequences	#frames	year
ALOV300 [52]	life	314	151.6k	2014
OTB100 [66]	life	100	59.0k	2015
TC128 [36]	life	128	55.3k	2015
VOT2016 [29]	life	60	21.5k	2016
UAV123 [43]	drone	123	110k	2016
NfS [21]	life	100	383k	2017
POT 210 [37]	planar objects	210	105.2k	2018
VisDrone-SOT2018	drone	132	106.4k	2018

The video clips in these three subsets are taken at different locations, but share similar environments and attributes. The dataset is collected in various real-world scenarios by various drone platforms (*i.e.*, different drone models) under various weather and lighting conditions, which is helpful for the researchers to improve the algorithm performance in real-world scenarios. We manually annotated the bounding boxes of targets (*e.g.*, pedestrians, dogs, and vehicles) as well as several useful attributes (*e.g.*, occlusion, background clutter, and camera motion) for algorithm analysis. We present the number of frames *vs.* the aspect ratio (*i.e.*, object height divided by width) change rate with respect to the first frame in Fig. 2 (a), and show the number of frames *vs.* the area change rate with respect to the first frame in Fig. 2 (b). We plot the distributions of the number of frames of video clips in the `training`, `validation`, and `testing` sets in Fig. 2(c). In addition, some annotated examples in the VisDrone-SOT2018 dataset are presented in Fig. 1.

3.2 Evaluation Protocol

Following the evaluation methodology in [66], we use the success and precision scores to evaluate the performance of the trackers. The success score is defined as the area under the success plot. That is, with each bounding box overlap threshold t_o in the interval $[0, 1]$, we compute the percentage of successfully tracked frames to generate the successfully tracked frames *vs.* bounding box overlap threshold plot. The overlap between the the tracker prediction B_t and the ground truth bounding box B_g is defined as $O = \frac{|B_t \bigcap B_g|}{|B_t \bigcup B_g|}$, where $\bigcap$ and $\bigcup$ represent the intersection and union between the two regions, respectively, and $|\cdot|$ calculates the number of pixels in the region. Meanwhile, the precision score is defined as the percentage of frames whose estimated location is within the given threshold distance of the ground truth based on the Euclidean distance in the image plane. Here, we set the distance threshold to 20 pixels in evaluation. Notably, the success score is used as the primary metric for ranking methods.

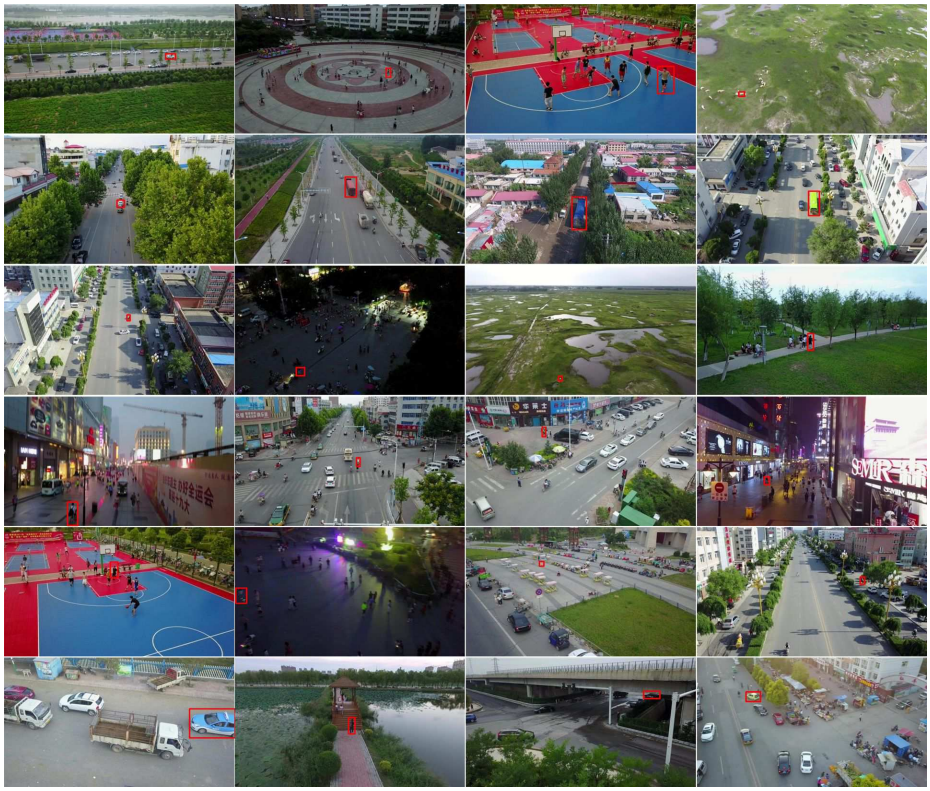

Fig. 1: Some annotated example video frames of single object tracking. The first frame with the bounding box of the target object is shown for each sequence.

3.3 Trackers Submitted

We have received 17 entries from 26 different institutes in the VisDrone-SOT2018 challenge. The VisDrone committee additionally evaluates 5 baseline trackers with the default parameters on the VisDrone-SOT2018 dataset. If the default parameters are not available, some reasonable values are used for evaluation. Thus, there are in total 22 algorithms are included in the single-object tracking task of VisDrone2018 challenge. In the following we briefly overview the submitted algorithms and provide their descriptions in the Appendix A.

Among in the submitted algorithms, 4 trackers are improved based on the correlation filter algorithm, including CFWCRKF (A.3), CKCF (A.6), DCST (A.16) and STAPLE_SRCA (A.17). Four trackers, *i.e.*, C3DT (A.4), VITALD (A.5), DeCom (A.8) and BTT (A.10), are developed based on the MDNet [46] algorithm, which is the winner of the VOT2015 challenge [31]. Seven trackers combine the CNN models and correlation filter algorithm, namely OST (A.1), CFCNN (A.7), TRACA+ (A.9), LZZ-ECO (A.11), SECFNet (A.12), SDRCO (A.14) and DCFNet (A.15), where OST (A.1), CFCNN (A.7) and LZZ-ECO

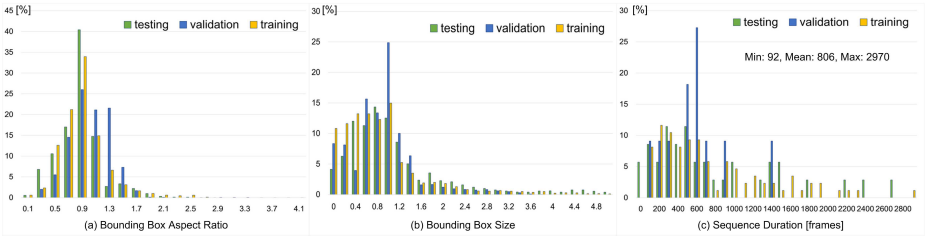

Fig. 2: (a) The number of frames *vs.* the aspect ratio (height divided by width) change rate with respect to the first frame, (b) the number of frames *vs.* the area change rate with respect to the first frame, and (c) the distributions of the number of frames of video clips, in the `training`, `validation`, and `testing` sets for single object tracking.

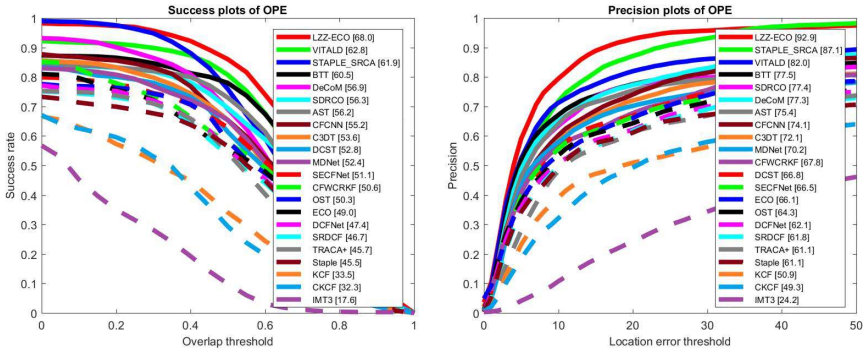

Fig. 3: The success and precision plots of the submitted trackers. The success and precision scores for each tracker are presented in the legend.

(A.11) apply object detectors to conduct target re-detection. One tracker (*i.e.*, AST (A.2)) is based on saliency map, and another tracker (*i.e.*, IMT3 (A.13)) is based on the normalized cross correlation filter.

3.4 Overall Performance

The overall success and precision plots of all submissions are shown in Fig. 3. Meanwhile, we also report the success and precision scores, tracking speed, implementation details, pre-trained dataset, and the references of each method in Table 2. As shown in Table 2 and Appendix A, we find that the majority of the top 5 trackers are using the deep CNN model. LZZ-ECO (A.11) employs the deep detector YOLOv3 [48] as the re-detection module and use the ECO [9] algorithm as the tracking module, which achieves the best results among all the 22 submitted trackers. VITALD (A.5) (rank 2), BTT (A.10) (rank 4) and DeCom (A.8) (rank 5) are all improved from the MDNet [46] algorithm, and VITALD (A.5) fine-tunes the state-of-the-art object detector RefineDet [68] on the

VisDrone-SOT2018 `training` set to re-detect the target to mitigate the drifting problem in tracking. Only the STAPLE_SRCA algorithm (A.17) (rank 3) in top 5 is the variant of the correlation filter integrated with context information. SDRCO (A.14) (rank 6) is an improved version of the correlation filter based tracker CFWCR [24], which uses the ResNet50 [23] network to extract discriminative features. AST (A.2) (rank 7) calculates the saliency map via aggregation signature for target re-detection, which is effective to track small target. CFCNN (A.7) combines multiple BACF trackers [22]) with the CNN model (*i.e.*, VGG16) by accumulating the weighted response of both trackers. This method ranks 8 among all the 22 submissions. Notably, most of the submitted trackers are improved from recently (after year 2015) leading computer vision conferences and journals.

4 Results and Analysis

According to the success scores, the best tracker is LZZ-ECO (A.11), followed by the VITALD method (A.5). STAPLE_SRCA (A.17) performs slightly worse with the gap of 0.9%. In terms of precision scores, LZZ-ECO (A.11) also performs the best. The second and third best trackers based on the precision score are STAPLE_SRCA (A.17) and VITALD (A.5). It is worth pointing out that the top two trackers employ the combination of state-of-the-art object detectors (*e.g.*, YOLOv3 [48] and RefineDet [68]) for target re-detection and an accurate object tracking algorithm (*e.g.*, ECO [9] and VITAL [54]) for object tracking.

In addition, the baseline trackers (*i.e.*, KCF (A.18), Staple (A.19), ECO (A.20), MDNet (A.21) and SRDCF (A.22)) submitted by the VisDrone committee, rank at the lower middle level of all the 22 submissions based on the success and precision scores. This phenomenon demonstrates that the submitted methods achieve significant improvements from the baseline algorithms.

4.1 Performance Analysis by Attributes

Similar to [43], we annotate each sequence with 12 attributes and construct subsets with different dominant attributes that facilitate the analysis of the performance of trackers under different challenging factors. We show the performance of each tracker of 12 attributes in Fig. 4 and 5. We present the descriptions of 12 attributes used in evaluation, and report the median success and precision scores under different attributes of all 22 submissions in Table 3. We find that the most challenging attributes in terms of success score are *Similar Object* (36.1%), *Background Clutter* (41.2%) and *Out-of-View* (41.5%).

As shown in Fig. 4 and 5, LZZ-ECO (A.11) achieves the best performance in all 12 attribute subsets, and other trackers rank the second place in turn. Specifically, VITALD (A.5) achieves the second best success score in terms of the *Aspect Ratio Change*, *Camera Motion*, *Fast Motion*, *Illumination Variation*, *Out-of-View* and *Scale Variation* attributes. We speculate that the object detection module in VITALD is effective to re-detect the target to mitigate the drift

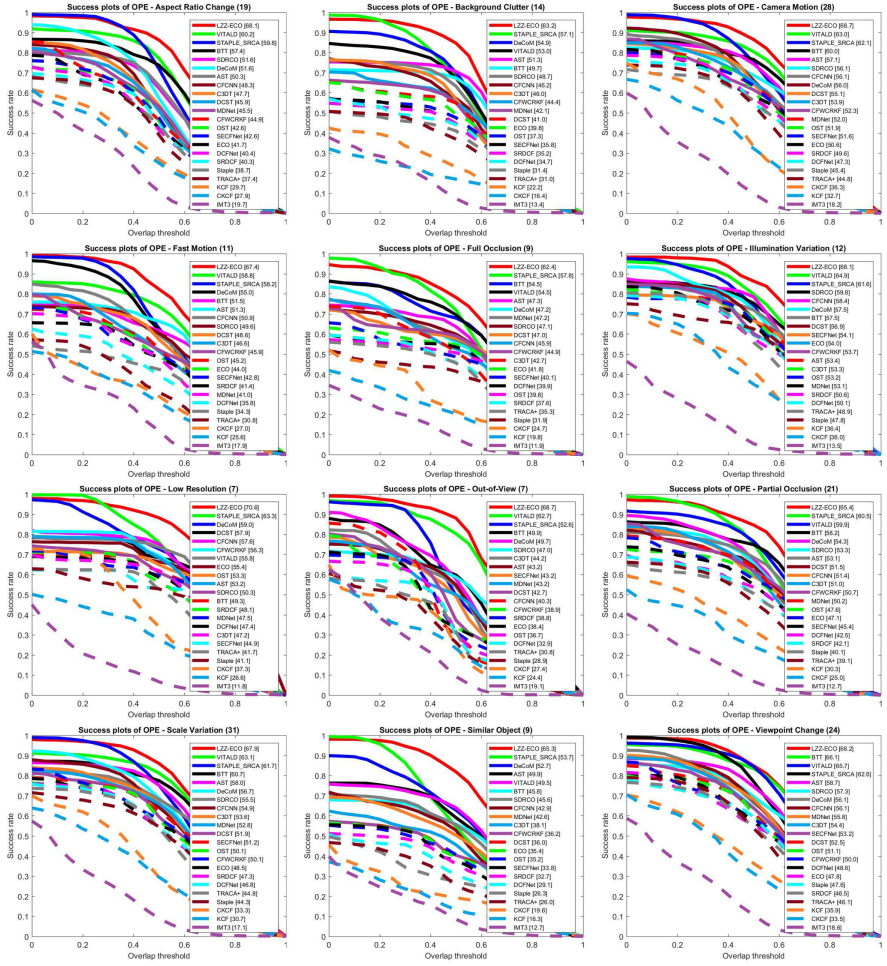

Fig. 4: The success plots for the submitted trackers in different attributes, *e.g.*, aspect ratio change, background clutter, camera motion, etc.). The number presented in the title indicates the number of sequences with that attribute.

problem to produce more accurate results. STAPLE_SRCA (A.17) performs the second best in *Background Clutter*, *Full Occlusion*, *Low Resolution*, *Partial Occlusion* and *Similar Object* attributes, which demonstrates the effectiveness of the proposed sparse response context-aware correlation filters. BTT (A.10) only performs worse than LZZ-ECO (A.11) in *Viewpoint Change* attribute, which benefits from the backtracking-term, short-term and long-term model updating mechanism based on the discriminative training samples.

We also report the comparison between the MDNet and ECO trackers in the subsets of different attributes in Fig. 6. The MDNet and ECO trackers are

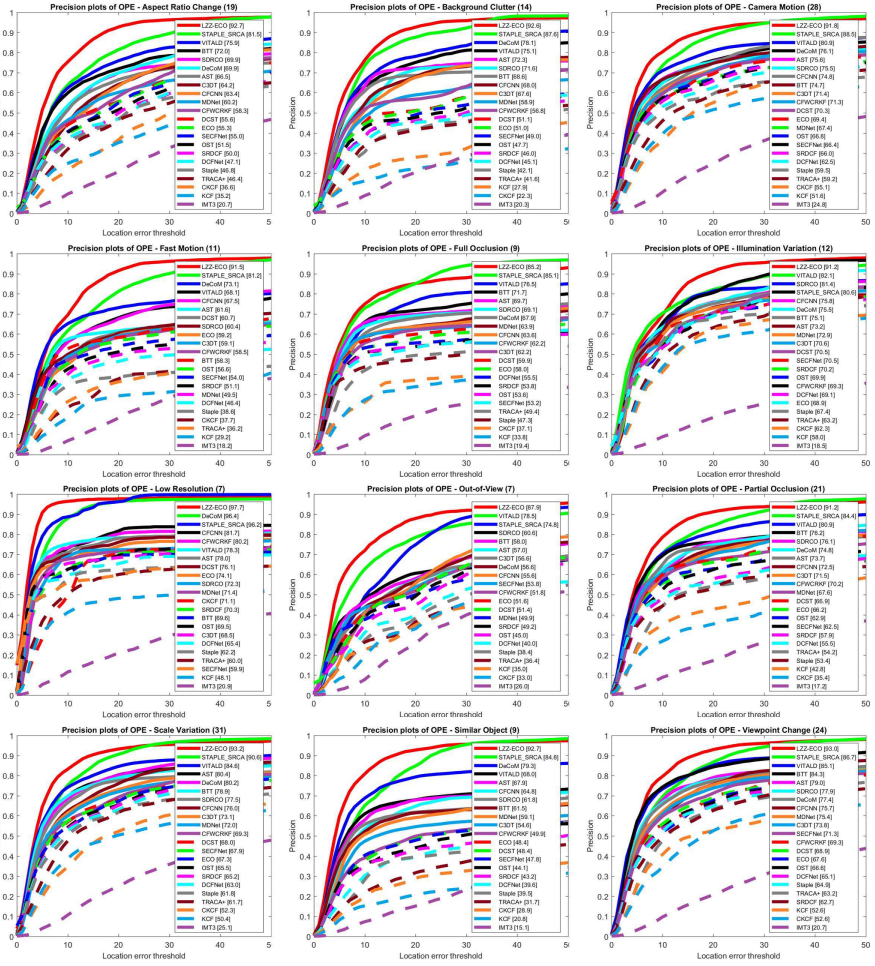

Fig. 5: The precision plots for the submitted trackers in different attributes, *e.g.*, aspect ratio change, background clutter, and camera motion. The number presented in the title indicates the number of sequences with that attribute.

two popular methods in single-object tracking field. We believe the analysis is important to understand the progress of the tracking algorithms on the drone-based platform. As shown in Fig. 6, ECO achieves favorable performance against MDNet in the subsets of the fast motion (FM), illumination variation (IV), and low resolution (LR) attributes, while MDNet performs better than ECO in the other attribute subsets. In general, the deep CNN model based MDNet is able to produce more accurate results than ECO. However, the ECO tracker still has some advantages worth to learn. For the FM subset, it is difficult for MDNet to train a reliable model using such limited training data. To solve this issue, BTT

Table 2: Comparison of all submissions in the VisDrone-SOT2018 challenge. The success score, precision score, tracking speed (in FPS), implementation details (M indicates Matlab, P indicates Python, and G indicates GPU), pre-trained dataset (I indicates ImageNet, L indicates ILSVRC, P indicates PASCAL VOC, V indicates the VisDrone-SOT2018 `training` set, O indicates other additional datasets, and × indicates that the methods do not use the pre-trained datasets) and the references are reported. The ∗ mark indicates the methods submitted by the VisDrone committee.

Submission	Success	Precision	Speed	Impl.	Pre-trained	Reference
OST (A.1)	50.3	54.3	54.2	M,G	I,V	CVPR'17 [9]
AST (A.2)	56.2	75.4	5.9	M,G	V	ICCV'15 [11]
CFWCRKF (A.3)	50.6	67.8	11.7	M,G	I,V	ICCVW'17 [24]
C3DT (A.4)	53.6	72.1		P,G	I,V,O	CVPR'16 [46]
VITALD (A.5)	62.8	82.0	0.6	M,P,G	I,V,O	CVPR'18 [54]
CKCF (A.6)	32.3	49.3	59	P,G	×	TPAMI'15 [25]
CFCNN (A.7)	55.2	74.1	12	M,G	×	ICCV'17 [22]
DeCoM (A.8)	56.9	77.3	3.3	P,G	I,V	CVPR'16 [46]
TRACA+ (A.9)	45.7	61.1	46.2	M,G	I,P	CVPR'18 [6]
BTT (A.10)	60.5	77.5	2.1	M,G	I,V,O	CVPR'16 [46]
LZZ-ECO (A.11)	**68.0**	**92.9**		M,G	×	CVPR'17 [9]
SECFNet (A.12)	51.1	66.5	13.6	M,G	L,V	CVPR'17 [59]
IMT3 (A.13)	17.6	24.2		M	×	NCC
SDRCO (A.14)	56.3	77.4	0.3	M,G	V	ICCVW'17 [24]
DCFNet (A.15)	47.4	62.1	35.1	M,G	L	arXiv'17 [61]
DCST (A.16)	52.8	66.8	25.5	M	×	CVPR'16 [2]
STAPLE_SRCA (A.17)	61.9	87.1		M	×	CVPR'17 [44]
KCF* (A.18)	33.5	50.9	254.4	M	×	TPAMI'15 [25]
Staple* (A.19)	45.5	61.1	39.9	M	×	CVPR'16 [2]
ECO* (A.20)	49.0	66.1	1.3	M	×	CVPR'17 [9]
MDNet* (A.21)	52.4	70.2	2.6	M,G	I	CVPR'16 [46]
SRDCF* (A.22)	46.7	61.8	6.5	M	×	ICCV'15 [11]

(A.10) uses an extra backtracking-term updating strategy when the tracking score is not reliable. For the IV subset, ECO constructs a compact appearance representation of target to prevent overfitting, producing better performance than MDNet. For the LR subset, the appearance of small object is no longer informative after several convolutional layers, resulting in inferior performance of deep CNN based methods (*e.g.*, MDNet and VITALD (A.5)). Improved from MDNet, DeCoM (A.8) introduces an auxiliary tracking algorithm based on color template matching when deep tracker fails. It seems that color cue is effective to distinguish small objects.

4.2 Discussion

Compared to previous single-object tracking datasets and benchmarks, such as OTB100 [66], VOT2016 [29], and UAV123 [43], the VisDrone-SOT2018 dataset

Table 3: Attributes used to characterize each sequence from the drone-based tracking perspective. The median success and precision scores under different attributes of all 22 submissions are reported to describe the tracking difficulties. The three most challenging attributes are presented in bold **red**, **blue** and **green** fonts, respectively.

Attribute	Success	Precision	Description
Aspect Ratio Change (ARC)	45.2	57.0	The fraction of ground truth aspect ratio in the first frame and at least one subsequent frame is outside the range $[0.5, 2]$.
Background Clutter (BC)	**41.2**	**54.0**	The background near the target has similar appearance as the target.
Camera Motion (CM)	52.2	69.9	Abrupt motion of the camera.
Fast Motion (FM)	45.6	58.4	Motion of the ground truth bounding box is larger than 20 pixels between two consecutive frames.
Full Occlusion (FOC)	43.8	61.1	The target is fully occluded.
Illumination Variation (IV)	53.6	70.5	The illumination of the target changes significantly.
Low Resolution (LR)	49.8	71.3	At least one ground truth bounding box has less than 400 pixels.
Out-of-View (OV)	**51.7**	**41.5**	Some portion of the target leaves the view.
Partial Occlusion (POC)	50.5	67.3	The target is partially occluded.
Scale Variation (SV)	51.6	68.7	The ratio of initial and at least one subsequent bounding box is outside the range $[0.5, 2]$.
Similar Object (SOB)	**36.1**	**49.2**	There are objects of similar shape or same type near the target.
Viewpoint Change (VC)	52.9	70.3	Viewpoint affects target appearance significantly.

involves very wide viewpoint, small objects and fast camera motion challenges, which puts forward the higher requirements of the single-object tracking algorithms. To make the tracker more effective in such scenarios, there are several directions worth to explore, described as follows.

– **Object detector based target re-identification.** Since the target appearance is easily changed in drone view, it is quite difficult for traditional trackers to describe the appearance variations accurately for a long time. State-of-the-art object detectors, such as YOLOv3 [48], R-FCN [8] and RefineDet [68], are able to help the trackers recover from the drifting problem and generate more accurate results, especially for the targets with large deformation or in the fast moving camera. For example, LZZ-ECO (A.11) outperforms the ECO (A.20) tracker with a large margin, *i.e.*, generates 19% higher success score and 26.8% higher precision score.

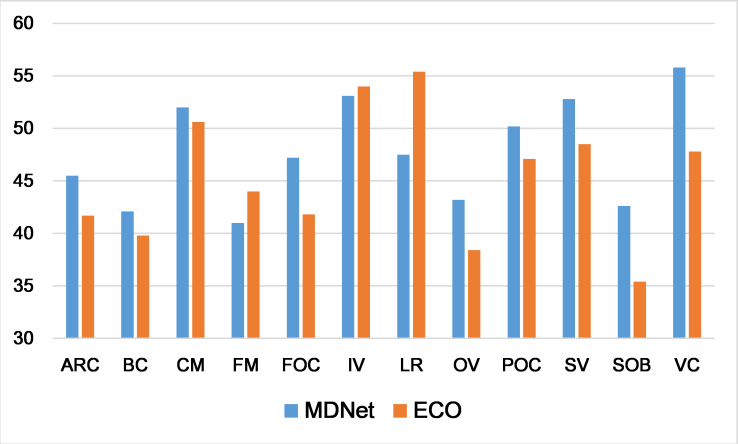

Fig. 6: Comparison of the MDNet and ECO algorithms with each attribute. The x-axis is the abbreviation of the 12 attributes, and the y-axis is the success scores of MDNet and ECO.

- **Searching region.** Since the video sequences in the VisDrone-SOT2018 dataset often involves wide viewpoint, it is critical to expand the search region to ensure that the target is able to be detected by the tracker, even if the fast motion or occlusion happen. For example, BTT (A.10) improves 8.1% and 7.3% higher success and precision scores, compared to MDNet (A.21).
- **Spatio-temporal context.** The majority of the CNN-based trackers only consider the appearance features in the video frames, and are hard to benefit from the consistent information included in consecutive frames. The spatio-temporal context information is useful to improve the robustness of the trackers, such as the optical flow [1], RNN [61] and 3DCNN [58] algorithms. In addition, the spatio-temporal regularized correlation filter (*e.g.*, DCST (A.16)) is another effective algorithm to deal with the appearance variations by exploiting the spatio-temporal information.
- **Multi-modal features.** It is important for the trackers to employ multiple types of features (*e.g.*, deep features, texture features and color features) to improve the robustness in different scenarios in tracking. The comparison results between DeCoM (A.8) and MDNet (A.21) show that the integration of different features is very useful to improve the tracking accuracy. Moreover, adding the appropriate weights on the responses of correlation filters is effective in tracking task (see SDRCO (A.14)).
- **Long-term and short-term updating.** During the tracking process, the foreground and background samples are usually exploited to update the appearance model to prevent the drifting problem when fast motion and occlusion happen. Long-term and short-term updates are always used to capture gradual and instantaneous variations of object appearance (see LZZ-ECO

(A.11)). It is important to design an appropriate updating mechanism for both long-term and short-term updating for better performance.

5 Conclusions

In this paper, we give a brief review of the VisDrone-SOT2018 challenge. The challenge releases a dataset formed by 132 video sequences, *i.e.*, 86 sequences with 69,941 frames for training, 11 sequences with 7,046 frames for validation, and 35 sequences with 29,367 frames for testing. We provide fully annotated bounding boxes of targets as well as several useful attributes, *e.g.*, occlusion, background clutter, and camera motion. A total of 22 trackers have been evaluated on the collected dataset. A large percentage of them are inspired from the state-of-the-art object algorithms. The top three trackers are LZZ-ECO (A.11), VITALD (A.5), and STAPLE_SRCA (A.17), achieving 68.0, 62.8, and 61.9 success scores, respectively.

We are glad to organize the VisDrone-SOT2018 challenge in conjunction with ECCV 2018 in Munich, Germany, successfully. A large amount of researchers participate the workshop to share their research progress. This workshop will not only serve as a meeting place for researchers in this area but also present major issues and potential opportunities. We believe the released dataset allows for the development and comparison of the algorithms in the single-object tracking field, and workshop challenge provide a way to track the process. Our future work will be focused on improving the dataset and the evaluation kit based on the feedbacks from the community.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grant 61502332 and Grant 61732011, in part by Natural Science Foundation of Tianjin under Grant 17JCZDJC30800, in part by US National Science Foundation under Grant IIS-1407156 and Grant IIS-1350521, in part by Beijing Seetatech Technology Co., Ltd and GE Global Research.

A Submitted Trackers

In this appendix, we provide a short summary of all algorithms participated in the VisDrone2018-SOT competition. These are ordered according to the submissions of their final results.

A.1 Ottawa-sjtu-tracker (OST)

Yong Wang, Lu Ding, Robert Laganière, Xinbin Luo
ywang6@uottawa.ca, dinglu@sjtu.edu.cn, laganier@eecs.uottawa.ca
losinbin@sjtu.edu.cn

OST is the combination of R-FCN detector [8] and ECO tracker [9]. Our algorithm is as follows: the tracker tracks the target. If the response is below a threshold, it indicates tracking failure. The detector provides detection results and the tracker searches target in the candidates and finally locate the target. The feature for tracker is HOG. The tracking results are based on the original R-FCN which is trained on ImageNet [51]. The detector is trained on VisDrone2018 training set and implemented offline at present.

A.2 Aggregation signature tracker (AST)

Chunlei Liu, Wenrui Ding, Jinyu Yang, Baochang Zhang, Jungong Han, Hanlin Chen
liuchunlei@buaa.edu.cn, ding@buaa.edu.cn, 17801004216@163.com
bczhang@buaa.edu.cn, jungonghan77@gmail.com, 15734029010@163.com

AST includes the base tracker and re-detection stages, particularly for small objects. The part of aggregation signature calculation illustrates the saliency map calculation in the re-detection procedure. Once a drifting is detected, we choose the searching region around the center of the previous target location to calculate the saliency map via aggregation signature. In the learning process, the target prior and the context information are used to learn the saliency map that helps find a new searching initial position, where the base tracker will be performed again for re-detection.

A.3 Correlation Filters with Weighted Convolution Responses and Kalman Filter (CFWCRKF)

Shengyin Zhu, Yanyun Zhao
lichenggang@bupt.edu.cn, zyy@bupt.edu.cn

CFWCRKF is built upon a correlation filters based tracker known as the Correlation Filters with Weighted Convolution Responses (CFWCR) [24], an improved version of the popular tracker Efficient Convolution Operators Tracker (ECO) [9]. ECO is an improved version of the tracker C-COT [13] and has achieved impressive results on the visual tracking benchmark. We have made some modifications to the algorithm of CFWCR, such as search area scale and weights factor. The most significant change is that we add Kalman Filter in the algorithm to deal with occlusion and fast motion.

A.4 3D Convolutional Networks for Visual Tracking (C3DT)

Haojie Li, Sihang Wu
{*201721011386, eesihang*}*@mail.scut.edu.cn*

C3DT improves the existing tracker MDNet [46] by introducing spatio-temporal information using the C3D network [58]. MDNet treats the tracking as classification and regression, which utilizes the appearance feature from the current frame to determine which candidate frame is object or background, and then gets a accurate bounding box by a linear regression. This network ignores the importance of spatio-temporal information for visual tracking. To address this problem, our approach adopts two-branch network to extract features. One branch is used to get features from the current frame by the VGG-S [5]; another is the C3D network, which extracts spatio-temporal information from the previous frames. C3DT fuses the features between two branch network to do the task of classification and regression.

A.5 VIsual Tracking via Adversarial Learning and Object Detection (VITALD)

Yuankai Qi, Yifan Yang, Weidong Chen, Kaiwen Duan, Qianqian Xu, Qingming Huang
qykshr@gmail.com, yangyifan@yeah.net, cwd2123@gmail.com
duankaiwen17@mails.ucas.ac.cn, xuqianqian@ict.ac.cn, qmhuang@ucas.ac.cn

VITALD is based on the VITAL tracker [54]. We improve VITAL from three aspects. First, we randomly augment fifty percent of the training data via flipping, rotation, and blurring. Second, we propose to adaptively adjust the size of the target searching region when the target scale change-ratio and translation between two contiguous frames exceed the thresholds α and β, respectively. Third, we train a pedestrian detection model and a vehicle (car, truck) detection model based on RefineDet [68] to provide additional target candidates for the target/background classification. According the given ground truth and detection results of these two models in the first frame, our method adaptively determines whether the detection should be used and to use which detection model.

A.6 CERTH's KCF algorithm on Visdrone (CKCF)

Emmanouil Michail, Konstantinos Avgerinakis, Panagiotis Giannakeris, Stefanos Vrochidis, Ioannis Kompatsiaris
{*michem, koafgeri, giannakeris, stefanos, ikom*}*@iti.gr*

CKCF is based on KCF [25]. For specific sequences that needed excessive memory resource, the algorithm was applied sequentially, by splitting the whole sequence in shorter sequences and using as initial bounding boxes, the predicted bounding boxes of the previous sequence.

A.7 Jointly weighted correlation filter and convolutional neural network (CFCNN)

Wei Tian and Martin Lauer
{*wei.tian, martin.lauer*}*@kit.edu*

CFCNN combines both the correlation filter and the convolutional neural network into a single framework by accumulating the weighted response of each tracker model. For implementation, we employ the BACF tracker as our correlation filter model and keep the parameters from its paper [22]. For the CNN model, we deploy a simple residual network structure consisting of 2 base layers and 3 residual layers. The input for CF is the concatenation of HOG and Color Name [12] features while the input of our CNN model is the response map from the layer conv4-3 of a pre-trained VGG16 network. The channel number of response map from VGG16 is shrinked to 32 by PCA approach for computational efficiency. To cope with abrupt motion, we employ a very large searching area for each tracker model, *i.e.*, 10 times of the target size.

A.8 Deep tracker with Color and Momentum (DeCoM)

Byeongho Heo, Sangdoo Yun, Jin Young Choi
bhheo@snu.ac.kr, sangdoo.yun@navercorp.com, jychoi@snu.ac.kr

DeCoM applies color and motion based tracking algorithm based on MDNet [46]. The scenes in the VisDrone dataset is very wide, and in most cases the object does not return to the same place. Therefore, we introduce an auxiliary tracking algorithm that can roughly follow the object even if the deep tracker fails. Classical color-based template matching is more efficient than deep features and edge-based features in the situations such as motion blur and heavy occlusion. In our tracking algorithm, if the deep tracker fails, an auxiliary tracker based on template matching is activated and tracks the object until the deep tracker is successful again. The tracking target of auxiliary tracker is the area around the object including the background for robust tracking. Besides, we introduce momentum in the auxiliary tracker to cope with heavy occlusion. Since the target of auxiliary tracker includes the background, the tracking position is closer to the background position than the actual object position. Thus, the difference between the position of a deep tracker and the auxiliary tracker approximates the relative speed of the background and the object. When the deep tracker is successful, we accumulate this difference to measure the momentum of the object, and when the deep tracker fails, the tracking result is made to move as much as the momentum, so as to predict where the object exits from the occlusion.

A.9 Extended context-aware deep feature compression for high-speed visual tracking (TRACA+)

Kyuewang Lee, Jongwon Choi, Jin Young Choi
{kyuewang5056, jwchoi.pil}@gmail.com, jychoi@snu.ac.kr

TRACA+ is a fast and effective deep feature-based tracker which is suitable to UAV camera environments. To address the issues such as confusing appearance of small objects, frequent occlusion in an urban environment, and abrupt camera motion due to swift change of UAV position, we have extended TRACA [6] to be applied to UAV environments. The reason to choose TRACA is that it achieves both high speed and high performance at the same time. Since the computing power of the embedded systems on drones is low, TRACA can be a viable tracking solution. Although TRACA shows superior performance in many of the benchmark datasets, UAV camera environments such as drones remain challenging due to the following hindrances: confusing appearance of small objects, frequent occlusion in an urban environment, and heavy or abrupt camera motion. To handle these hindrances, we extend TRACA by adding two-fold techniques. First, we concatenate RGB color feature in addition to the compressed feature to relieve the effects of confusing appearance of small objects and motion blur from the abrupt camera motion. Second, we propose a homography-based Kalman filtering method to predict the next frame target position which is combined with the CF tracking position in a convex combination manner to get the next frame final position. This method can not only handle occlusion problems to some degree but also predict object motion regardless of camera motion.

A.10 Visual Tracking using Backtracking (BTT)

Ke Song, Xixi Hu, Wenhao Wang, Yaxuan Li, and Wei Zhang
201613125@mail.sdu.edu.cn, huxixity@gmail.com, 201400040023@mail.sdu.edu.cn
yaxuanli2018@gmail.com, davidzhangsdu@mail.sdu.edu.cn

BTT is improved from the MDNet [46] algorithm to handle fast motion (FM), partial occlusion (POC) and full occlusion (FOC). The modifications are mainly in two aspects: First, we generate 500 positive samples in the first frame of sequence then extract and store the features of them. These features are used to update network to prevent the model drift caused by background when fast motion and occlusion arise. In detail, besides the long-term and short-term updates, we add an extra backtracking-term update, which is performed when the positive score of the estimated target is less than 0.3. The samples used for backtracking-term update contains three parts: The first one are the positive samples generated from the first frame as stated above. The second one are the samples generated from the last 20 frames that the result confidence score is greater than 0.5. The last one are the negative samples. Considering that the old negative examples are often redundant or irrelevant to the current frame we only select the last 10 frames to generate negative samples. The negative samples are

collected in the manner of hard negative mining. Second, correspondingly, we expand the search scale in one frame and increase the number of target candidates aimed at effective re-detection to fast motion and occlusion situation.

A.11 An improved ECO algorithm for preventing camera shake, long-term occlusion and adaptation to target deformation (LZZ-ECO)

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

LZZ-ECO is based on ECO [9] and has made the following improvements based on ECO:

(1) We add the object detection algorithm YOLOv3 [48] to optimize the location of the target, especially when the target has a large deformation or camera angle changes. When the target is violently deformed or the camera's perspective changes, the traditional ECO tracking box may only contain a part of the target. At this time, using the detection results of the detection algorithm to optimize the tracking results will achieve good results. Specifically, when the above situation is detected, a pixel block of 400×400 (in order to approximate the input picture size of YOLOv3) extracted around the center of the tracking box will be input to YOLOv3. Then the IOU of tracking box and each detection box are calculated in the detection result to select the detection box with the highest IOU as the optimized box.

(2) To deal with the long time occlusion problem, we use the optical flow method [1] to estimate the approximate motion trajectory of the target in the occluded stage when the target is detected to be occluded. Thus the tracking algorithm can track the target successfully when it appears again. Moreover, when the target is detected to be occluded, we stop update the correlation filters in ECO because the image used for filter training may already be an occlusion rather than a target at this time.

(3) To deal with the camera violent shaking problem, we use the sift feature based matching algorithm [40] to calculate the offset of the target between the current frame and the previous frame to accurately locate the position of the target in the current frame. It can successfully track several sequences of camera shakes in the testing sequences, which improves significantly in those with the sheep target.

A.12 Feature learning in CFNet and channel attention in SENet by focal loss (SECFNet)

Dongdong Li, Yangliu Kuai, Hao Liu, Zhipeng Deng, Juanping Zhao
{lidongdong12, kuaiyangliu09}@nudt.edu.cn)

SECFNet is based on the feature learning study in CFNet [59], channel attention

in SENet [27] and focal loss in [38]. The proposed tracker introduces channel attention and focal loss into the network design to enhance feature representation learning. Specifically, a Squeeze-and-Excitation (SE) block 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. To alleviate the foreground-background data imbalance, we propose a focal logistic loss by adding a modulating factor to the logistic loss, with two tunable focusing parameters. The focal logistic loss down-weights the loss assigned to easy examples in the background area. Both the SE block and focal logistic loss are computationally lightweight and impose only a slight increase in model complexity. Our tracker is pre-trained on the ILSVRC2015 dataset and fine-tuned on the VisDrone2018 train set.

A.13 Iteratively Matching Three-tier Tracker (IMT3)

Asanka G Perera
asanka.perera@mymail.unisa.edu.au

IMT3 is a method to use with Normalized cross-correlation (NCC) filter for rotation and scale invariant object tracking. The proposed solution consists of three modules: (i) multiple appearance generation in the search image at different rotation angles and scales, (ii) bounding box drifting correction by a re-initialization step, and (iii) failure handling by tracker combination. A point tracker that uses the Kanade-Lucas-Tomasi feature-tracking algorithm and a histogram-based tracker that uses the continuously adaptive mean shift (CAMShift) algorithm have been used as supporting trackers.

A.14 Convolution Operators for Tracking using Resnet features using Rectangle Rectifier with Similarity Network to Solve the Occlusion Problem (SDRCO)

Zhiqun He, Ruixin Zhang, Peizhen Zhang, Xiaohao He
he010103@bupt.edu.cn, ruixinzhang@tencent.com, zhangpzh5@mail2.sysu.edu.cn
hexh17@mails.tsinghua.edu.cn

SDRCO is an improved version of the baseline tracker CFWCR [24]. We use ResNet features and new formulation to solve the correlation filter formula. Besides, we use Kalman filter to help smooth the results. After the tracking, we use a detector trained in the SOT training data to rectify the rectangle of RCO. We have a similarity network (ResNet50) to find out the occlusion frame and the Kalman filter to predict the location of the target and re-detect the target using the rectifier.

A.15 Discriminant correlation filters network for visual tracking (DCFNet)

Jing Li, Qiang Wang, and Weiming Hu
jli24@outlook.com, {qiang.wang, wmhu}@ia.ac.cn

DCFNet [61] is an end-to-end lightweight network architecture to learn the convolutional features and perform the correlation tracking process simultaneously. Specifically, we treat DCF as special correlation filter layer added in a Siamese network, and carefully derive the back-propagation through it by defining the network output as the probability heatmap of object location. Since the derivation is still carried out in Fourier frequency domain, the efficiency property of DCF is preserved. This enables our tracker to run at more than 60 FPS during test time, while achieving a significant accuracy gain compared with KCF using HoGs.

A.16 Dual Color clustering and Spatio-temporal regularized regressions based complementary Tracker (DCST)

Jiaqing Fan, Yifan Zhang, Jian Cheng, Kaihua Zhang, Qingshan Liu
fjq199407@163.com, {yfzhang, jcheng}@nlpr.ia.ac.cn, zhkhua@gmail.com, qsliu@nuist.edu.cn

DCST is improved from Staple [2], which is equipped with complementary learners of Discriminative Correlation Filters (DCFs) and color histograms to deal with color changes and deformations. Staple has some weakness: (i) It only employs a standard color histogram with the same quantization step for all sequences, which does not consider the specific structural information of target in each sequence, thereby affecting its discriminative capability to separate target from background. (ii) The standard DCFs are efficient but suffer from unwanted boundary effects, leading to failures in some challenging scenarios. Based on these issues, we make two significant improvements in color histogram regressor and DCF regressor, respectively. First, we design a novel color clustering based histogram model that first adaptively divides the colors of the target in the 1st frame into several cluster centers, and then the cluster centers are taken as references to construct adaptive color histograms for targets in the coming frames, which enable to adapt significant target deformations. Second, we propose to learn spatio-temporal regularized CFs, which not only enables to avoid boundary effects but also provides a more robust appearance model than DCFs in Staple in the case of large appearance variations. Finally, we fuse these two complementary merits.

A.17 Sparse response context-aware correlation filter tracking (STAPLE_SRCA)

Wenhua Zhang, Yang Meng
{zhangwenhua_nuc, xdyangmeng}@163.com

STAPLE_SRCA [44] is a context-aware tracking proposed based on the framework of correlation filter. A problem is that when the target moves out of the scene or is completely covered by other objects, it is possible that the target will be lost forever. When the target comes out again, the tracker cannot track the target. Focusing on this problem, we propose a sparse response context-aware correlation filter tracking method based on STAPLE [2]. In the training process, we force the expected response to be as sparse as possible, then most responses are close to 0. When the target disappears, all the responses will be close to 0. Then in the tracking process, the case that the target moves out of the scene or be covered by other objects can be easily recognized and this frame is taken as a pending frame. As a consequence, those frames will not influence the frames where the target comes out.

A.18 High-Speed Tracking with Kernelized Correlation Filters (KCF)

Submitted by VisDrone Committee

KCF is the Kernelized Correlation Filter [25] with HOG features. Based on a linear kernel, the linear multi-channel filters are performed with very low computational complexity (*i.e.*, running at hundreds of frames-per-second). It is equivalent to a kernel ridge regression trained with thousands of sample patches around the object at different translations. Please refer to [25] for more details.

A.19 Complementary Learners for Real-Time Tracking (Staple)

Submitted by VisDrone Committee

Staple improves the traditional correlation filters based tracker by combining complementary cues in a ridge regression framework. Correlation filter-based trackers usually sensitive to deformation while color statistics based on models can handle variation in shape well. Staple combines both representations to learn a model that is inherently robust to color changes and deformations. Specifically, it is solved with two independent ridge-regression problems efficiently. Please refer to [2] for more details.

A.20 Efficient Convolution Operators for Tracking (ECO)

Submitted by VisDrone Committee

ECO significantly improves the tracking performance of the Discriminative Correlation Filter (DCF) based methods in three-folds. (1) A factorized convolution operator is developed to reduce the number of parameters in the model drastically. (2) A compact generative model of the training sample distribution are

proposed to reduce memory and time complexity significantly while provide better diversity of samples. (iii) A conservative model update strategy is introduced for robustness and reduced complexity. Please refer to [9] for more details.

A.21 Learning Multi-Domain Convolutional Neural Networks for Visual Tracking (MDNet)

Submitted by VisDrone Committee

MDNet is a single object tracking algorithm based on the representations from a discriminatively trained CNN model. Specifically, the network consists of shared layers and multiple branches of domain-specific layers. The "domains" indicate individual training sequences, and each branch is responsible for binary classification to identify target in each domain. Each domain is train iteratively to obtain generic target representations in the shared layers for binary classification. The tracking is performed by sampling target candidates around the previous target state, evaluating them on the CNN, and selecting the sample with the maximum score. Please refer to [46] for more details.

A.22 Learning Spatially Regularized Correlation Filters for Visual Tracking (SRDCF)

Submitted by VisDrone Committee

SRDCF is the abbreviation of Spatially Regularized Discriminative Correlation Filters. Specifically, we introduce a novel spatial regularization component in the learning to penalize correlation filter coefficients depending on their spatial location. The proposed formulation allows the correlation filters to be learned on a significantly larger set of negative training samples, without corrupting the positive samples. Please refer to [11] for more details.

References

1. Beauchemin, S.S., Barron, J.L.: The computation of optical flow. ACM Comput. Surv. **27**(3), 433–467 (1995)
2. Bertinetto, L., Valmadre, J., Golodetz, S., Miksik, O., Torr, P.H.S.: Staple: Complementary learners for real-time tracking. In: CVPR. pp. 1401–1409 (2016)
3. Bertinetto, L., Valmadre, J., Henriques, J.F., Vedaldi, A., Torr, P.H.S.: Fully-convolutional siamese networks for object tracking. In: ECCV. pp. 850–865 (2016)
4. Cai, Z., Wen, L., Lei, Z., Vasconcelos, N., Li, S.Z.: Robust deformable and occluded object tracking with dynamic graph. TIP **23**(12), 5497–5509 (2014)
5. Chatfield, K., Simonyan, K., Vedaldi, A., Zisserman, A.: Return of the devil in the details: Delving deep into convolutional nets. In: BMVC (2014)
6. Choi, J., Chang, H.J., Fischer, T., Yun, S., Lee, K., Jeong, J., Demiris, Y., Choi, J.Y.: Context-aware deep feature compression for high-speed visual tracking. In: CVPR (2018)

7. Choi, J., Chang, H.J., Jeong, J., Demiris, Y., Choi, J.Y.: Visual tracking using attention-modulated disintegration and integration. In: CVPR. pp. 4321–4330 (2016)
8. Dai, J., Li, Y., He, K., Sun, J.: R-FCN: object detection via region-based fully convolutional networks. In: NIPS. pp. 379–387 (2016)
9. Danelljan, M., Bhat, G., Khan, F.S., Felsberg, M.: ECO: efficient convolution operators for tracking. In: CVPR. pp. 6931–6939 (2017)
10. Danelljan, M., Häger, G., Khan, F.S., Felsberg, M.: Accurate scale estimation for robust visual tracking. In: BMVC (2014)
11. Danelljan, M., Häger, G., Khan, F.S., Felsberg, M.: Learning spatially regularized correlation filters for visual tracking. In: ICCV. pp. 4310–4318 (2015)
12. Danelljan, M., Khan, F.S., Felsberg, M., van de Weijer, J.: Adaptive color attributes for real-time visual tracking. In: CVPR. pp. 1090–1097 (2014)
13. Danelljan, M., Robinson, A., Khan, F.S., Felsberg, M.: Beyond correlation filters: Learning continuous convolution operators for visual tracking. In: ECCV. pp. 472–488 (2016)
14. Dong, X., Shen, J., Wang, W., Liu, Y., Shao, L., Porikli, F.: Hyperparameter optimization for tracking with continuous deep q-learning. In: CVPR. pp. 518–527 (2018)
15. Du, D., Qi, H., Li, W., Wen, L., Huang, Q., Lyu, S.: Online deformable object tracking based on structure-aware hyper-graph. TIP **25**(8), 3572–3584 (2016)
16. Du, D., Qi, Y., Yu, H., Yang, Y., Duan, K., Li, G., Zhang, W., Huang, Q., Tian, Q.: The unmanned aerial vehicle benchmark: Object detection and tracking. In: ECCV (2018)
17. Du, D., Wen, L., Qi, H., Huang, Q., Tian, Q., Lyu, S.: Iterative graph seeking for object tracking. TIP **27**(4), 1809–1821 (2018)
18. Fan, H., Lin, L., Yang, F., Chu, P., Deng, G., Yu, S., Bai, H., Xu, Y., Liao, C., Ling, H.: LaSOT: A high-quality benchmark for large-scale single object tracking. arXiv (2018)
19. Fan, H., Ling, H.: Parallel tracking and verifying: A framework for real-time and high accuracy visual tracking. In: ICCV. pp. 5487–5495 (2017)
20. Felsberg, M., Berg, A., Häger, G., Ahlberg, J., Kristan, M., Matas, J., Leonardis, A., Cehovin, L., Fernández, G., Vojír, T., Nebehay, G., Pflugfelder, R.P.: The thermal infrared visual object tracking VOT-TIR2015 challenge results. In: ICCVWorkshops. pp. 639–651 (2015)
21. Galoogahi, H.K., Fagg, A., Huang, C., Ramanan, D., Lucey, S.: Need for speed: A benchmark for higher frame rate object tracking. In: ICCV. pp. 1134–1143 (2017)
22. Galoogahi, H.K., Fagg, A., Lucey, S.: Learning background-aware correlation filters for visual tracking. In: ICCV. pp. 1144–1152 (2017)
23. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016)
24. He, Z., Fan, Y., Zhuang, J., Dong, Y., Bai, H.: Correlation filters with weighted convolution responses. In: ICCVWorkshops. pp. 1992–2000 (2017)
25. Henriques, J.F., Caseiro, R., Martins, P., Batista, J.: High-speed tracking with kernelized correlation filters. TPAMI **37**(3), 583–596 (2015)
26. Hsieh, M., Lin, Y., Hsu, W.H.: Drone-based object counting by spatially regularized regional proposal network. In: ICCV (2017)
27. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. CoRR **abs/1709.01507** (2017), `http://arxiv.org/abs/1709.01507`

28. Hu, W., Li, X., Zhang, X., Shi, X., Maybank, S.J., Zhang, Z.: Incremental tensor subspace learning and its applications to foreground segmentation and tracking. IJCV **91**(3), 303–327 (2011)
29. Kristan, M., Leonardis, A., Matas, J., Felsberg, M., Pflugfelder, R.P., Cehovin, L., Vojír, T., Häger, G., Lukezic, A., Fernández, G., *et al.*: The visual object tracking VOT2016 challenge results. In: ECCVWorkshops. pp. 777–823 (2016)
30. Kristan, M., Leonardis, A., Matas, J., Felsberg, M., Pflugfelder, R.P., Zajc, L.C., Vojir, T., Häger, G., Lukezic, A., Eldesokey, A., Fernández, G., *et al.*: The visual object tracking VOT2017 challenge results. In: ICCVWorkshops. pp. 1949–1972 (2017)
31. Kristan, M., Matas, J., Leonardis, A., Felsberg, M., Cehovin, L., Fernández, G., Vojír, T., Häger, G., Nebehay, G., Pflugfelder, R.P.: The visual object tracking VOT2015 challenge results. In: ICCVWorkshops. pp. 564–586 (2015)
32. Li, A., Li, M., Wu, Y., Yang, M.H., Yan, S.: NUS-PRO: A new visual tracking challenge. In: IEEE Transactions on Pattern Analysis and Machine Intelligence. pp. 1–15 (2015)
33. Li, B., Yan, J., Wu, W., Zhu, Z., Hu, X.: High performance visual tracking with siamese region proposal network. In: CVPR. pp. 8971–8980 (2018)
34. Li, F., Tian, C., Zuo, W., Zhang, L., Yang, M.: Learning spatial-temporal regularized correlation filters for visual tracking. In: CVPR (2018)
35. Li, S., Du, D., Wen, L., Chang, M., Lyu, S.: Hybrid structure hypergraph for online deformable object tracking. In: ICPR. pp. 1127–1131 (2017)
36. Liang, P., Blasch, E., Ling, H.: Encoding color information for visual tracking: Algorithms and benchmark. TIP **24**(12), 5630–5644 (2015)
37. Liang, P., Wu, Y., Lu, H., Wang, L., Liao, C., Ling, H.: Planar object tracking in the wild: A benchmark. In: ICRA (2018)
38. Lin, T., Goyal, P., Girshick, R.B., He, K., Dollár, P.: Focal loss for dense object detection. In: ICCV. pp. 2999–3007 (2017)
39. Liu, B., Huang, J., Kulikowski, C.A., Yang, L.: Robust visual tracking using local sparse appearance model and k-selection. TPAMI **35**(12), 2968–2981 (2013)
40. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. IJCV **60**(2), 91–110 (2004)
41. Ma, C., Huang, J., Yang, X., Yang, M.: Hierarchical convolutional features for visual tracking. In: ICCV. pp. 3074–3082 (2015)
42. Mei, X., Ling, H.: Robust visual tracking using $\ell 1$ minimization. In: ICCV. pp. 1436–1443 (2009)
43. Mueller, M., Smith, N., Ghanem, B.: A benchmark and simulator for UAV tracking. In: ECCV. pp. 445–461 (2016)
44. Mueller, M., Smith, N., Ghanem, B.: Context-aware correlation filter tracking. In: CVPR. pp. 1387–1395 (2017)
45. Müller, M., Bibi, A., Giancola, S., Al-Subaihi, S., Ghanem, B.: Trackingnet: A large-scale dataset and benchmark for object tracking in the wild. In: ECCV (2018)
46. Nam, H., Han, B.: Learning multi-domain convolutional neural networks for visual tracking. In: CVPR. pp. 4293–4302 (2016)
47. Qi, Y., Qin, L., Zhang, J., Zhang, S., Huang, Q., Yang, M.: Structure-aware local sparse coding for visual tracking. TIP **27**(8), 3857–3869 (2018)
48. Redmon, J., Farhadi, A.: Yolov3: An incremental improvement. CoRR **abs/1804.02767** (2018), `http://arxiv.org/abs/1804.02767`
49. Robicquet, A., Sadeghian, A., Alahi, A., Savarese, S.: Learning social etiquette: Human trajectory understanding in crowded scenes. In: ECCV. pp. 549–565 (2016)

50. Ross, D.A., Lim, J., Lin, R., Yang, M.: Incremental learning for robust visual tracking. IJCV **77**(1-3), 125–141 (2008)
51. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M.S., Berg, A.C., Li, F.: Imagenet large scale visual recognition challenge. IJCV **115**(3), 211–252 (2015)
52. Smeulders, A.W.M., Chu, D.M., Cucchiara, R., Calderara, S., Dehghan, A., Shah, M.: Visual tracking: An experimental survey. TPAMI **36**(7), 1442–1468 (2014)
53. Song, S., Xiao, J.: Tracking revisited using RGBD camera: Unified benchmark and baselines. In: ICCV. pp. 233–240 (2013)
54. Song, Y., Ma, C., Wu, X., Gong, L., Bao, L., Zuo, W., Shen, C., Lau, R.W.H., Yang, M.: VITAL: visual tracking via adversarial learning. In: CVPR (2018)
55. Tao, R., Gavves, E., Smeulders, A.W.M.: Siamese instance search for tracking. In: CVPR. pp. 1420–1429 (2016)
56. *et al.*, M.K.: The visual object tracking VOT2014 challenge results. In: ECCV-Workshops. pp. 191–217 (2014)
57. *et al.*, M.F.: The thermal infrared visual object tracking VOT-TIR2016 challenge results. In: ECCVWorkshops. pp. 824–849 (2016)
58. Tran, D., Bourdev, L.D., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3d convolutional networks. In: CVPR. pp. 4489–4497 (2015)
59. Valmadre, J., Bertinetto, L., Henriques, J.F., Vedaldi, A., Torr, P.H.S.: End-to-end representation learning for correlation filter based tracking. In: CVPR. pp. 5000–5008 (2017)
60. Wang, N., Zhou, W., Tian, Q., Hong, R., Wang, M., Li, H.: Multi-cue correlation filters for robust visual tracking. In: CVPR. pp. 4844–4853 (2018)
61. Wang, Q., Gao, J., Xing, J., Zhang, M., Hu, W.: Dcfnet: Discriminant correlation filters network for visual tracking. CoRR **abs/1704.04057** (2017), `http://arxiv.org/abs/1704.04057`
62. Wen, L., Cai, Z., Lei, Z., Yi, D., Li, S.Z.: Online spatio-temporal structural context learning for visual tracking. In: ECCV. pp. 716–729 (2012)
63. Wen, L., Cai, Z., Lei, Z., Yi, D., Li, S.Z.: Robust online learned spatio-temporal context model for visual tracking. TIP **23**(2), 785–796 (2014)
64. Wu, T., Lu, Y., Zhu, S.: Online object tracking, learning and parsing with and-or graphs. TPAMI **39**(12), 2465–2480 (2017)
65. Wu, Y., Lim, J., Yang, M.: Online object tracking: A benchmark. In: CVPR. pp. 2411–2418 (2013)
66. Wu, Y., Lim, J., Yang, M.: Object tracking benchmark. TPAMI **37**(9), 1834–1848 (2015)
67. Yun, S., Choi, J., Yoo, Y., Yun, K., Choi, J.Y.: Action-decision networks for visual tracking with deep reinforcement learning. In: CVPR (2017)
68. Zhang, S., Wen, L., Bian, X., Lei, Z., Li, S.Z.: Single-shot refinement neural network for object detection. In: CVPR (2018)
69. Zhong, W., Lu, H., Yang, M.: Robust object tracking via sparse collaborative appearance model. TIP **23**(5), 2356–2368 (2014)