VisDrone-MOT2020: The Vision Meets Drone Multiple Object Tracking Challenge Results

Heng Fan[1], Dawei Du[2], Longyin Wen[3], Pengfei Zhu[4(✉)], Qinghua Hu[4], Haibin Ling[1], Mubarak Shah[5], Junwen Pan[4], Arne Schumann[10], Bin Dong[7], Daniel Stadler[8], Duo Xu[12], Filiz Bunyak[17], Guna Seetharaman[18], Guizhong Liu[6], V. Haritha[15], P. S. Hrishikesh[15], Jie Han[6], Kannappan Palaniappan[17], Kaojin Zhu[14], Lars Wilko Sommer[9], Libo Zhang[19], Linu Shine[15], Min Yao[19], Noor M. Al-Shakarji[16,17], Shengwen Li[13], Ting Sun[6], Wang Sai[7], Wentao Yu[6], Xi Wu[12], Xiaopeng Hong[6], Xing Wei[6], Xingjie Zhao[6], Yanyun Zhao[13], Yihong Gong[6], Yuehan Yao[7], Yuhang He[6], Zhaoze Zhao[11], Zhen Xie[12], Zheng Yang[14], Zhenyu Xu[7], Zhipeng Luo[7], and Zhizhao Duan[12]

[1] Stony Brook University, New York, NY, USA
[2] Kitware, Inc., Clifton Park, NY, USA
[3] JD Finance America Corporation, Mountain View, CA, USA
[4] Tianjin University, Tianjin, China
zhupengfei@tju.edu.cn
[5] University of Central Florida, Orlando, FL, USA
[6] Xi'an Jiaotong University, Xi'an, China
[7] DeepBlue Technology (Shanghai), Shanghai, China
[8] Karlsruhe Institute of Technology, Karlsruhe, Germany
[9] Fraunhofer IOSB, Karlsruhe, Germany
[10] Fraunhofer Center for Machine Learning, Karlsruhe, Germany
[11] Southwestern University of Finance and Economics, Chengdu, China
[12] Zhejiang University, Hangzhou, China
[13] Beijing University of Posts and Telecommunications, Beijing, China
[14] Xidian University, Xi'an, China
[15] College of Engineering Trivandrum, Thiruvananthapuram, India
[16] University of Technology, Baghdad, Iraq
[17] University of Missouri-Columbia, Columbia, MO, USA
[18] U.S. Naval Research Laboratory, Washington, DC, USA
[19] Institute of Software, Chinese Academy of Sciences, Beijing, China

Abstract. The Vision Meets Drone (VisDrone2020) Multiple Object Tracking (MOT) is the third annual UAV MOT tracking evaluation activity organized by the VisDrone team, in conjunction with European Conference on Computer Vision (ECCV 2020). The VisDrone-MOT2020 consists of 79 challenging video sequences, including 56 videos (∼24K frames) for training, 7 videos (∼3K frames) for validation and 17 videos (∼6K frames) for evaluation. All frames in these sequences are manually annotated with high-quality bounding boxes. Results of 12 participating MOT algorithms are presented and analyzed in detail. The challenging

A. Bartoli and A. Fusiello (Eds.): ECCV 2020 Workshops, LNCS 12538, pp. 713–727, 2020.
https://doi.org/10.1007/978-3-030-66823-5_43

results, video sequences as well as the evaluation toolkit are made available at http://aiskyeye.com/. By holding VisDrone-MOT2020 challenge, we hope to facilitate future research and applications of MOT algorithms on drone videos.

Keywords: Drone-based multiple object tracking · Drone · Performance evaluation

1 Introduction

The goal of multiple object tracking (MOT) is to simultaneously determine the identities of multiple moving target objects and estimate their trajectories in a video sequence. MOT is one of the most important components for video understanding in computer vision and has a long list of applications such as video surveillance, human-machine interaction, and robotics.

In order to boost the development of MOT, benchmarks have played a crucial role in developing and evaluating MOT algorithms. To this end, many MOT benchmarks have been proposed in recent years, such as UA-DETRAC [42], MOT challenge [13,32], TAO [12], and KITTI [17]. Nevertheless, these MOT benchmarks focus on general scenes for person or vehicle tracking. Recently, drone based algorithms and algorithms have drawn extensive attention owing to the flexibility of drone platform. Therefore, there are several attempts to constructing drone tracking datasets [14,20,33,38]. However, these drone benchmarks are often limited in size and do not aim at multiple object tracking. Based on these analysis, a large-scale MOT benchmark based on drone platform is still desired.

As mentioned by [42], a typical MOT algorithm consists of two components including an object detection component and a multi-object tracking component. In this challenge, unlike traditional MOT benchmark [32], we argue that it is more reasonable to assess the whole MOT system without using common prior detection results as inputs. For this purpose, we organize the challenge workshop "Vision Meets Drone Multiple Object Tracking (VisDrone-MOT2020)", in conjunction with European Conference on Computer Vision (ECCV 2020), followed by previous successful editions VisDrone-MOT2019 [45] (in conjunction with ICCV 2019) and VisDrone-VDT2018 [59] (in conjunction with ECCV 2018). We present 12 MOT algorithms with in-depth analysis and discussion. It is worth noting that many of these algorithms are modified based on existing state-of-the-art detection methods or multi-object tracking approaches. The results and videos of the VisDrone-MOT2020 challenge are available at http://aiskyeye.com/.

2 Related Work

2.1 Multi-Object Tracking

One of the most common strategy in MOT is tracking-by-detection. The approach of [44] introduces a new data association technique based on hierarchical relation hypergraph. The main idea is to formulate the MOT problem as a dense neighborhoods searching task on the dynamic affinity graph. The method in [22] proposes to leverage long-short term memory (LSTM) to incorporate temporal information of different targets into appearance modeling, which effectively improves the performance. In order to hand the noisy in detection results, the method of [57] utilizes a single object tracker and apply the tracking result for data association. The approach of [39] proposes to employ both spatial and temporal information for appearance modeling for improvement. The method of [21] takes motion information into consideration and proposes to combine low- and high-level cues for multi-object tracking. The approach in [43] proposes to learn a non-unified hypergraph for multi-object tracking. To fully explore powerful deep feature representation, the algorithm of [11] proposes an end-to-end architecture for feature learning, affinity estimation and multi-dimensional assignment. The method of [10] introduces an instance-aware tracker to combine single object tracking methods for MOT by encoding awareness both within and between target models, significantly boosting the performance. The approach of [5] learns a neural solver based on message passing networks for multi-object tracking in an end-to-end fashion. The approach of [3] exploits the bounding box regression in detection and apply it to improve the performance of multi-object tracking.

2.2 Similarity Learning in Person Re-Identification

The goal of person re-identification (Re-ID) is to recognize the person of interest from a set of gallery images. It is often adopted to deal with the data associate problem in MOT. The approach of [50] proposes an unsupervised strategy to learn multi-level descriptors from pixel-level, patch and image levels for person re-identification. The method in [27] introduce a filter pairing neural network to handle various misalignment problem in person re-identification. The work of [53] leverages salient cues from human body for recognition. The method in [55] proposes a graph learning method to deal with misalignment in person re-identification using local structures. The approach of [52] proposes to utilize deep learning method to learn an aligned representation for person re-identification. Further, this method is extend by [7] through incorporating bilinear coding to improve the robustness of feature representation. The method of [8] proposes to exploit high-order attention for person re-identification.

In addition to image based person re-identification, video based person re-identification has also drawn increasing interest. The work of [24] explores both global and local temporal information for video person re-identification. The approach of [25] utilizes 3D convolution network for video person re-identification. The work of [26] combines both motion and appearance information to improve video person re-identification.

2.3 Tracking Benchmark

Multi-object tracking is one of the most important problems in computer vision. In order to advance the research of MOT, many benchmarks have been proposed in recent years to evaluation different MOT approaches. Different from single object tracking benchmarks [16,47], constructing multi-object tracking benchmarks is more challenging as the number of targets in each frame are much larger than one. The KITTI benchmark [17] focuses on multi-object tracking in traffic scenes. MOT challenge 2016 [32] is one of the most popular MOT benchmarks for pedestrian tracking. Later, this challenge is extended to MOT challenge 2020 [13] by introducing more video sequences. The UA-DETRAC [42] proposes a MOT benchmark in various traffic sceneries and new protocols for evaluation. Considering the requirement of large-scale dataset in deep learning era, the recently proposed TAO [12] contributes a large set of videos for multi-object tracking. Especially, this dataset provides both training and evaluation videos for MOT. Despite the above MOT benchmarks, there is a lack of drone based MOT benchmark, which motivates the proposal of this challenge.

3 The VisDrone-MOT2020 Challenge

3.1 The VidDrone-MOT2020 Dataset

Similar to VisDrone-MOT2019 [45], VisDrone-MOT2020 uses the same video sequences for a fair comparison. In specific, VisDrone-MOT2020 consists of 79 video clips with around 70,000 frames in total. We divide VisDrone-MOT2020 into three subsets, including training set (56 video clips with around 24,000 frames), validation set (7 video clips with around 3,000 frames) and testing set (16 video clips with around 6,000 frames). Each frame in these videos are manually labeled in high quality. Similar to VisDrone-MOT2019 [45], we focus on five selected target classes in this challenge, including *pedestrian*, *car*, *van*, *bus* and *truck*.

As discussed early, VisDrone-MOT2020 does not provide the common detection results as inputs to the trackers, we encourage participants to use their own detectors to offer detection results. In this way, we are able to evaluate the complete multi-object tracking system more reliably. Following VisDrone-MOT2019 [45], we utilize the same evaluation protocol in [35] to evaluate the performance of submitted trackers. Specifically, each submitted tracker needs to generate a list of (axis-aligned) bounding box with confidence scores and the corresponding identities. Then the tracklets, which are formed by the bounding box detection results with the same identity, are sorted based on the average confidence scores over the detection results. Each tracklet is measured by its intersection over union (IoU) overlap with the groundtruth tracklet. If the IoU is larger than a pre-defined threshold (*i.e.*, 0.25, 0.50 and 0.75), the tracklet is correct. Finally, all participating algorithms are ranked by averaging the mean average precision (mAP) per object class over different thresholds. For more details, please refer to [35].

Table 1. The summary of the submitted MOT algorithms in the VisDrone-MOT2020 Challenge. GPUs for training, implementations (Python or Matlab), framework, pre-trained datasets ('V' indicates VisDrone-MOT2020 and 'C' indicates COCO [29]).

Method	GPU	Implementation	Framework	Pre-trained
COFE	TITAN Xp	Python	Cascade R-CNN [6]+OSNet [54]	V,C
SOMOT	Tesla V100	Python	Cascade R-CNN [6]	V,C
PAS tracker	Tesla V100	Python	Cascade R-CNN [6]+FPN [28]	V,C
Deepsort	n/a	Python	Fast R-CNN [18]+SORT [46]	C
YOLO-TRAC	Tesla P100	Python	YOLO-V5+V-IOU [4]	V,C
VDCT	GTX 1080Ti	Python	CenterTrack [56]+OSNet [54]	V,C
Cascade RCNN+IOU	n/a	Python	Cascade R-CNN [6]+IOU [4]	V,C
HTC+IOU	n/a	Python	HTC [9]+IOU [4]	V,C
HR-GNN	GTX 1080Ti	Python	HRNet [41]	V
TNT	Colab	Python	TNT [39]	V
anchor-free_mot	RTX 2080	Python	FairMOT [51]	V
SCTrack	GTX 960M	Matlab	Faster R-CNN [37]+YOLOv3 [36]	C

3.2 Submitted Trackers

In VisDrone-MOT2020 challenge, we receive 12 different multi-object tracking algorithms with detailed description, as described in Appendix A. Many of them are modified based on existing state-of-the-art detection models such as Cascade R-CNN [6], FPN [28] and HTC [9] and MOT methods such as SORT [46] and FairMOT [51]. Moreover, Re-ID techniques [31,40,54] are introduced to improve the accuracy of detection association.

All of these participating trackers are based on the tracking-by-detection framework. In specific, the submissions COFE (A.1), SOMOT (A.2), PAS tracker (A.3) and Cascade RCNN+IOU (A.7) apply the Cascade R-CNN detector [6] to obtain detection inputs. COFE (A.1) proposes a coarse-to-fine strategy to refine the tracking results in various kinds of vehicles (*e.g.*, van, bus, and car). SOMOT (A.2) relies on the embedding model by Multiple Granularity Network [40] to deal with detection association. PAS tracker (A.3) builds the similarity measure that integrates position, appearance and size information of objects, where the appearance of an object is represented by a feature vector computed with a re-identification model from [31]. The submissions Deepsort (A.4) and VDCT (A.6) are built based on the state-of-the-art MOT methods, *i.e.*, SORT [46] and CenterTrack [56]. The submission YOLO-TRAC (A.5) leverages the latest YOLO-v5 detector for multi-object tracking. The submission HTC+IOU (A.8) uses instance segmentation technique for MOT. The submission HR-GNN (A.9) employs stronger backbone to extract features and apply graph neural network for MOT. The submission TNT (A.10) combines temporal and appearance information together to develop a unified MOT framework. The submission anchor-free_mot (A.11) introduces anchor-free detector for multi-object tracking [51]. The submission SCTrack (A.12) develops a cascade architecture for multi-object tracking by exploiting various information. The summary of these trackers is shown in Table 1.

Table 2. Multi-object tracking results on the VisDrone-MOT2020 test-challenge set. The best three performers are highlighted by the **red**, **green** and **blue** fonts, respectively.

Method	AP	AP@0.25	AP@0.50	AP@0.75	AP_{car}	AP_{bus}	AP_{trk}	AP_{ped}	AP_{van}
COFE (A.1)	**61.88**	**64.99**	**62.00**	**58.65**	**79.09**	**65.26**	**50.91**	**56.87**	**57.26**
SOMOT (A.2)	**57.65**	**70.06**	**60.13**	**42.75**	**68.52**	**62.10**	**47.98**	**54.94**	**54.69**
PAS tracker (A.3)	**50.80**	**62.24**	**50.74**	**39.43**	**62.59**	**50.59**	**42.18**	**44.34**	**54.30**
Deepsort (A.4)	42.11	58.82	42.64	24.86	55.06	43.18	41.30	29.10	41.88
YOLO-TRAC (A.5)	42.10	52.94	41.86	31.49	52.81	48.98	39.17	28.92	40.59
VDCT (A.6)	35.76	45.86	35.46	25.96	56.94	24.62	28.16	34.00	35.06
Cascade RCNN+IOU (A.7)	27.23	36.14	28.25	17.31	49.56	16.27	30.18	10.78	29.36
HTC+IOU (A.8)	26.46	34.39	27.43	17.57	51.18	19.05	21.55	10.77	29.76
HR-GNN (A.9)	19.54	26.52	19.67	12.42	37.72	15.48	9.98	18.87	15.65
TNT (A.10)	6.55	10.93	7.00	1.70	1.88	19.51	2.07	1.96	7.32
anchor-free_mot (A.11)	4.88	9.73	3.38	1.53	10.69	2.04	1.51	5.89	4.26
SCTrack (A.12)	3.01	5.01	2.77	1.24	8.99	1.21	2.16	1.68	0.98

4 Results and Analysis

4.1 Overall Performance

The results of the submitted 12 multi-object tracking algorithms are presented in Table 2. COEF (A.1), SOMOT (A.2) and PAS tracker (A.3) achieve the top 3 AP score among all submissions, respectively. In specific, COFE obtains the AP score of 61.88, SOMOT AP score of 57.65 and PAS Tracker AP score of 50.80. In addition, under two out of three thresholds, COFE achieves the best performance on AP@0.50 and AP@0.75 with 62.00 and 58.65 scores. Moreover, it shows the best results for all five target classes. Further, we have observed that all these three solutions adopt the state-of-the-art detector Cascade R-CNN [6] to obtain inputs. We argue that the reason is that the targets in VisDrone-MOT2020 have various scales. The cascade strategy in Cascade R-CNN is essential for detecting these targets on drone-captured scenes. Moreover, the top 3 performers introduce the re-identification model in the MOT framework, resulting in considerable improvement. Following the top 3 MOT methods, Deepsort (A.4) and YOLO-TRAC (A.5) are typical tracking-by-detection frameworks without re-identification, achieving the AP score of 42.

It is worth noting that, compared with the solutions in VisDrone-MOT2019 [45], the submitted multi-object trackers in VisDrone-MOT2020 performs much better. In specific, the top three trackers in VisDrone-MOT2019 [45] achieve the AP scores of 43.94, 39.19 and 30.87, respectively. In comparison, the top three trackers COEF (A.1), SOMOT (A.2) and PAS tracker (A.3) respectively obtain the AP scores of 61.88, 57.65 and 50.80, significantly improving the performance. A potential reason accounting for this is that the submitted trackers in this challenge employ more powerful detection models and multi-object tracking components, leading to better performance.

4.2 Performance Analysis by Categories

In order to further analyze different tracking algorithms, we report the AP score for each target classes in VisDrone-MOT2020, as shown in Table 2. From Table 2, we observe that COFE (A.1) consistently achieves the best performance under all target categories. In specific, COFE achieves AP scores of 79.09, 65.26, 50.91, 56.87 and 57.26 for categories *car*, *bus*, *truck*, *pedestrian* and *van*. SOMOT (A.2) achieves the second best performance of AP scores 68.52, 62.10, 47.98, 54.94 and 54.69 for these five classes. PAS tracker (A.3) obtains the third best results of 62.59, 50.59, 42.18, 44.34 and 54.30 for five categories.

5 Discussion

It is challenging to develop a robust MOT approach on drone videos. Based on the results above, there is still large room to improvement for future research. Here we summarize some effective strategies for boosting MOT performance on drone-captured scenarios.

- **Robust detection.** Different from existing MOT benchmarks, we aim to evaluate the complete MOT system. According the top performers in this challenge, a robust detection model (*e.g.*, Cascade R-CNN [6]) with a simple detection association method (*e.g.*, Kalman Filter, Hungarian method [23] and SORT [46]) can achieve state-of-the-art performance compared with joint detection and tracking methods (*e.g.*, CenterTrack [56] and FairMOT [51]). Therefore, it is necessary to choose a robust detector to pursue better performance of multi-object tracking.
- **Motion information.** Although the current state-of-the-art MOT methods focus on modeling appearance information of the targets, motion information is also crucial for multi-object tracking, especially when targets have similar appearance or are occluded by others. In this challenge, the submission VDCT (A.6) incorporate motion information of targets and show promising results.

6 Conclusion

This paper concludes the VisDrone-MOT2020 challenge, which is the third annual UAV MOT tracking evaluation activity, in conjunction with ECCV 2020. In this challenge, we present a more challenging dataset consisting of 79 video clips with around 33K frames. 12 MOT algorithms based on existing state-of-the-art detectors or multi-object trackers are submitted and analyzed. The top three solutions are COFE (A.1), SOMOT (A.2) and PAS tracker (A.3) with the AP scores of 61.88, 57.65 and 50.80, respectively. The experiment shows the effectiveness of Cascade R-CNN detector and re-identification techniques in the tracking-by-detection MOT framework. We hope that this challenge can advance future research and applications of drone based MOT algorithms [58].

Acknowledgements. This work was supported in part by the National Natural Science Foundation of China under Grant 61876127 and Grant 61732011, in part by Natural Science Foundation of Tianjin under Grant 17JCZDJC30800.

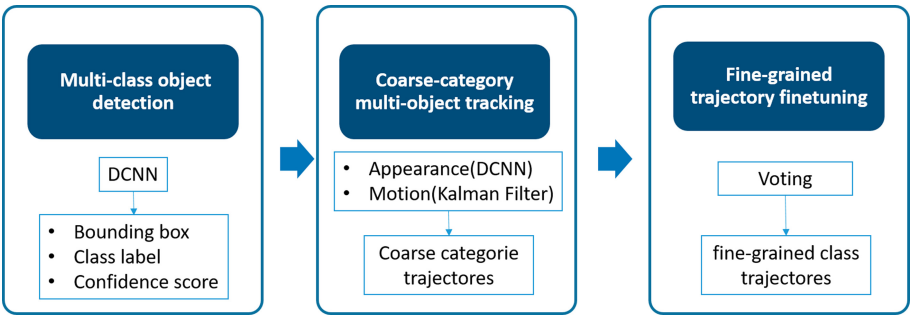

Fig. 1. The framework of COFE.

A Descriptions of Submitted Trackers

In the appendix, we summarize 12 trackers submitted in the VisDrone-MOT2020 Challenge, which are ordered according to the submissions of their final results.

A.1 Coarse-to-Fine Multi-Class Multi-Object Tracking (COFE)

Yuhang He, Wentao Yu, Jie Han, Xiaopeng Hong, Xing Wei and Yihong Gong
{*hyh1379478,yu1034397129,hanjie1997*}*@stu.xjtu.edu.cn,*
{*hongxiaopeng,weixing,ygong*}*@mail.xjtu.edu.cn*

COFE is proposed to track multiple targets in different categories under different scenarios. As shown in Fig. 1, the proposed method contains three major modules: 1) Multi-class object detection, 2) Coarse-category multi-object tracking, and 3) Fine-grained trajectory finetuning. Firstly, we use a Deep Convolutional Neural Network (DCNN) based object detector [6] to detect interested targets in the image plane, where each detection is denoted by a bounding box with a class label and a confidence score. Secondly, we track multiple targets in coarse categories, where fine-grained classes (such as van, bus, car) are summarized into coarse categories (*e.g.*, vehicle). For each coarse category, we perform multi-object tracking by exploiting the appearance and motion information of targets, where the appearance feature is extracted using a DCNN feature extractor [54] and the motion pattern of each target is modeled by a Kalman Filter. Finally, for each obtained trajectory, we finetune its fine-grained class label by a simple voting and refine the tracking results by post processing (*i.e.*, bounding box smoothing).

A.2 Simple Online Multi-Object Tracker (SOMOT)

Zhipeng Luo, Yuehan Yao, Zhenyu Xu, Bin Dong and Wang Sai
{luozp,yaoyh,xuzy,dongb,wangs}@deepblueai.com

Following separate detection and embedding model, we build a strong detector based on Cascade R-CNN [6] and a embedding model based on Multiple Granularity Network (MGN). For association step, we build simple online multi-object tracker, which is inspired by DeepSORT [46] and FairMOT [51]. For detector, Cascade R-CNN [6] pretrained on COCO [29] is applied. For embedding model, bag of tricks are used to improve the performance of MGN [40]. For association step, we initialize a number of tracklets based on the estimated boxes in the first frame. In the subsequent frames, we associate the boxes to the existing tracklets (all activated tracklets) according to their distances measured by embedding features. We update the appearance features of the trackers in each time step to handle appearance variations. Then, unmatched activated tracklets and estimated boxes are associated by their distance of Intersection over Union (IoU). Also, inactivated tracklets and estimated boxes are associated by their distance of IoU.

A.3 Position-, Appearance- and Size-aware Tracker (PAS tracker)

Daniel Stadler, Lars Wilko Sommer and Arne Schumann
daniel.stadler@kit.edu,{lars.sommer,arne.schumann}@iosb.fraunhofer.de

The PAS algorithm follows the tracking-by-detection paradigm. As detectors, we train two Cascade R-CNN [6] with FPN [28] on the VisDrone2020 MOT train and val set applying as backbone ResNeXt-101 [49] and HRNetV2p-W32 [41], respectively. Training is performed on randomly sampled image crops (608×608 pixels) and the SSD [30] data augmentation pipeline is used. To improve the quality of the detections, we utilize test-time strategies like horizontal flipping and multi-scale testing. Additionally, we generate category-specific expert models using weights from different epochs and from the two detectors with different backbones. For associating detections, we build a similarity measure that integrates position, appearance and size information of objects. A constant velocity model is assumed for the motion prediction of objects and a camera motion compensation model based on the Enhanced Correlation Coefficient Maximization [15] is also applied. The appearance of an object is represented by a feature vector computed with a re-identification model from [31] based on a ResNet-50 [19]. The association of tracks and new detections is solved by the Hungarian method [23]. Additionally, to remove false positive detections in crowded scenarios, a simple filtering approach considering the overlap of existing tracks and new detections is proposed. Finally, we remove short tracks with less than 10 frames and small tracks with a mean size of less than 100 pixels as most of them are false positives.

A.4 Simple Online and Realtime Tracking with a Deep Association (Deepsort)

Zhaoze Zhao
hanjie@smail.swufe.edu.cn

Simple Online and Realtime Tracking (SORT) [46] is a pragmatic approach to multiple object tracking with a focus on simple, effective algorithms. In this paper, we integrate appearance information to improve the performance of SORT. Due to this extension we are able to track objects through longer periods of occlusions, effectively reducing the number of identity switches. In spirit of the original framework we place much of the computational complexity into an offline pre-training stage where we learn a deep association metric on a large-scale person re-identification dataset. During online application, we establish measurement-to-track associations using nearest neighbour queries in visual appearance space.

A.5 YOLOv5 based V-IOU tracker (YOLO-TRAC)

Zhizhao Duan, Xi Wu, Duo Xu and Zhen Xie
{Duanai,21725018}@zju.edu.cn,wuxi9410@gmail.com,zjutxz@hotmail.com

Trac is a track by detection framework. We use YOLO-V5[1] as our detection network, and V-IOU Tracker [4] is used for tracking.

A.6 An improved multi-object tracking method for the VisDrone videos based on CenterTrack (VDCT)

Shengwen Li and Yanyun Zhao
{2019140337,zyy}@bupt.edu.cn

VDCT is improved from CenterTrack, which is a point-based framework that combines detection and tracking [56]. Its inputs include the current frame, the previous frame, and the tracked objects in the previous frame; and it outputs the displacements of tracked objects. Our improvements include: (1) The tracked objects which do not match within 20 frames are allowed to associate with objects detected in current frame by properly extending the survival time of the tracked objects. (2) The motion direction of adjacent frame objects usually does not change abruptly due to the continuity of object motion, so we calculate the dot product of the displacements of adjacent frame objects and decide whether to associate the objects. (3) We use the NIOU method [34] to perform non-maximum suppression on vehicle objects. (4) We adopt the hierarchical matching strategy in DeepSORT [46] to solve the long occlusion problem. (5) OSNet [54] is used to extract each trajectory's appearance feature, measure

[1] https://github.com/ultralytics/yolov5.

their distance from others and we simply merge two trajectories if their distance is close enough. The experimental results show the effectiveness of our improved method.

A.7 Cascade RCNN based IOU tracker (Cascade RCNN+IOU)

Ting Sun and Xingjie Zhao
sunting9999@stu.xjtu.edu.cn,1243273854@qq.com

We use Cascade R-CNN [6] as the detector with three improvements: (1) We use Group normalization [48] instead of Batch normalization; (2) We use online hard example mining to select positive and negative samples; (3) We use multiple scales to test our data; (4) We use two stronger backbones to train models and integrate them. Then, we perform detection association using the IOU tracker [4].

A.8 Hybrid task cascade based IOU tracker (HTC+IOU)

Ting Sun, Xingjie Zhao and Guizhong Liu
sunting9999@stu.xjtu.edu.cn,1243273854@qq.com

We use hybrid task cascade for instance segmentation [9] as the detector with three improvements: (1) We use Group normalization [48] instead of Batch normalization; (2) We use online hard example mining to select positive and negative samples; (3) We use multiple scales to test our data; (4) We use two stronger backbones to train models and integrate them. Then, we perform detection association using the IOU tracker [4].

A.9 Multi-object Tracking based on HRNet (HR-GNN)

Zheng Yang and Kaojin Zhu
151776257@qq.com,1320531351@qq.com

HR-GNN is built based on the detector using HRNet [41] as backbone. Then the tracking results are generated by using GNN to analyze the detection results.

A.10 Multi-object tracking with TrackletNet (TNT)

Haritha V, Melvin Kuriakose, Hrishikesh PS and Linu Shine
vakkatharitha@gmail.com

TNT is based on the work of [39] by merging temporal and appearance information together as a unified framework. We learn appearance similarity among tracklets by a graph model, where we use CNN features and intersection-over-union (IOU) with epipolar constraints to compensate camera movement between adjacent frames. Finally, the tracklets can be clustered into groups, resulting in trajectories with individual object IDs.

A.11 A simple baseline for one-shot multi-object tracking (anchor-free_mot)

Min Yao and Libo Zhang
libo@iscas.ac.cn

The anchor-free_mot method is based on FairMOT [51]. Specifically, we use the encoder-decoder network to extract feature maps. Then, two simple parallel heads are used to predict the bounding box and re-ID features of the targets, respectively. Notably, the targets are represented by points from the anchor-free object detection method.

A.12 Semantic Color Correlation Tracker (SCTrack)

Noor M. Al-Shakarji, Filiz Bunyak, Guna Seetharaman and Kannappan Palaniappan
{nmahyd,bunyak,palaniappank}@mail.missouri.edu,
gunasekaran.seetharaman@rl.af.mil

SCTrack is a time-efficient detection-based multi-object tracking method. Specifically, we use a three-step cascaded data association scheme to combine a fast spatial distance only short-term data association, a robust tracklet linking step using discriminative object appearance models, and an explicit occlusion handling unit relying not only on tracked objects' motion patterns but also on environmental constraints such as presence of potential occluders in the scene. The details can be referred to [1,2].

References

1. Al-Shakarji, N.M., Bunyak, F., Seetharaman, G., Palaniappan, K.: Multi-object tracking cascade with multi-step data association and occlusion handling. In: AVSS (2018)
2. Al-Shakarji, N.M., Seetharaman, G., Bunyak, F., Palaniappan, K.: Robust multi-object tracking with semantic color correlation. In: AVSS (2017)
3. Bergmann, P., Meinhardt, T., Leal-Taixe, L.: Tracking without bells and whistles. In: ICCV (2019)
4. Bochinski, E., Eiselein, V., Sikora, T.: High-speed tracking-by-detection without using image information. In: AVSS (2017)
5. Brasó, G., Leal-Taixé, L.: Learning a neural solver for multiple object tracking. In: CVPR (2020)
6. Cai, Z., Vasconcelos, N.: Cascade R-CNN: delving into high quality object detection. In: CVPR (2018)
7. Chang, Z., et al.: Weighted bilinear coding over salient body parts for person re-identification. Neurocomputing **407**, 454–464 (2020)
8. Chen, B., Deng, W., Hu, J.: Mixed high-order attention network for person re-identification. In: ICCV (2019)
9. Chen, K., et al.: Hybrid task cascade for instance segmentation. In: CVPR (2019)

10. Chu, P., Fan, H., Tan, C.C., Ling, H.: Online multi-object tracking with instance-aware tracker and dynamic model refreshment. In: WACV (2019)
11. Chu, P., Ling, H.: FAMNet: joint learning of feature, affinity and multi-dimensional assignment for online multiple object tracking. In: ICCV (2019)
12. Dave, A., Khurana, T., Tokmakov, P., Schmid, C., Ramanan, D.: TAO: a large-scale benchmark for tracking any object. arXiv (2020)
13. Dendorfer, P., et al.: MOT20: a benchmark for multi object tracking in crowded scenes. arXiv (2020)
14. Du, D., et al.: The unmanned aerial vehicle benchmark: object detection and tracking. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11214, pp. 375–391. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01249-6_23
15. Evangelidis, G.D., Psarakis, E.Z.: Parametric image alignment using enhanced correlation coefficient maximization. PAMI **30**(10), 1858–1865 (2008)
16. Fan, H., et al.: LaSOT: a high-quality benchmark for large-scale single object tracking. In: CVPR (2019)
17. Geiger, A., Lenz, P., Stiller, C., Urtasun, R.: Vision meets robotics: the KITTI dataset. Int. J. Robot. Res. **32**(11), 1231–1237 (2013)
18. Girshick, R.: Fast R-CNN. In: ICCV (2015)
19. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
20. Hsieh, M.R., Lin, Y.L., Hsu, W.H.: Drone-based object counting by spatially regularized regional proposal network. In: ICCV (2017)
21. Keuper, M., Tang, S., Andres, B., Brox, T., Schiele, B.: Motion segmentation & multiple object tracking by correlation co-clustering. PAMI **42**(1), 140–153 (2018)
22. Kim, C., Li, F., Rehg, J.M.: Multi-object tracking with neural gating using bilinear LSTM. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11212, pp. 208–224. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01237-3_13
23. Kuhn, H.W.: The Hungarian method for the assignment problem. Naval Res. Logist. Q. **2**(1–2), 83–97 (1955)
24. Li, J., Wang, J., Tian, Q., Gao, W., Zhang, S.: Global-local temporal representations for video person re-identification. In: ICCV (2019)
25. Li, J., Zhang, S., Huang, T.: Multi-scale 3D convolution network for video based person re-identification. In: AAAI (2019)
26. Li, S., Yu, H., Hu, H.: Appearance and motion enhancement for video-based person re-identification. In: AAAI (2020)
27. Li, W., Zhao, R., Xiao, T., Wang, X.: DeepReID: deep filter pairing neural network for person re-identification. In: CVPR (2014)
28. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: CVPR (2017)
29. Lin, T.Y., et al.: Microsoft COCO: common objects in context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8693, pp. 740–755. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10602-1_48
30. Liu, W., et al.: SSD: single shot multibox detector. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9905, pp. 21–37. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46448-0_2
31. Luo, H., Gu, Y., Liao, X., Lai, S., Jiang, W.: Bag of tricks and a strong baseline for deep person re-identification. In: CVPRW (2019)
32. Milan, A., Leal-Taixé, L., Reid, I., Roth, S., Schindler, K.: MOT16: a benchmark for multi-object tracking. arXiv (2016)

33. Mueller, M., Smith, N., Ghanem, B.: A benchmark and simulator for UAV tracking. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9905, pp. 445–461. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46448-0_27
34. Pan, S., Tong, Z., Zhao, Y., Zhao, Z., Su, F., Zhuang, B.: Multi-object tracking hierarchically in visual data taken from drones. In: ICCVW (2019)
35. Park, E., Liu, W., Russakovsky, O., Deng, J., Li, F.F., Berg, A.: Large Scale Visual Recognition Challenge 2017. http://image-net.org/challenges/LSVRC/2017
36. Redmon, J., Farhadi, A.: YOLOv3: an incremental improvement. arXiv (2018)
37. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. In: NIPS (2015)
38. Robicquet, A., Sadeghian, A., Alahi, A., Savarese, S.: Learning social etiquette: human trajectory understanding in crowded scenes. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9912, pp. 549–565. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46484-8_33
39. Wang, G., Wang, Y., Zhang, H., Gu, R., Hwang, J.: Exploit the connectivity: multi-object tracking with trackletnet. In: ACM MM, pp. 482–490 (2019)
40. Wang, G., Yuan, Y., Chen, X., Li, J., Zhou, X.: Learning discriminative features with multiple granularities for person re-identification. In: ACM MM (2018)
41. Wang, J., et al.: Deep high-resolution representation learning for visual recognition. PAMI (2020)
42. Wen, L., et al.: UA-DETRAC: a new benchmark and protocol for multi-object detection and tracking. Comput. Vis. Image Underst. **193**, 102907 (2020)
43. Wen, L., Du, D., Li, S., Bian, X., Lyu, S.: Learning non-uniform hypergraph for multi-object tracking. In: AAAI, pp. 8981–8988 (2019)
44. Wen, L., Li, W., Yan, J., Lei, Z., Yi, D., Li, S.Z.: Multiple target tracking based on undirected hierarchical relation hypergraph. In: CVPR (2014)
45. Wen, L., Zhang, Y., Bo, L., Shi, H., Zhu, R., et al.: VisDrone-MOT2019: the vision meets drone multiple object tracking challenge results. In: ICCVW, pp. 189–198 (2019)
46. Wojke, N., Bewley, A., Paulus, D.: Simple online and realtime tracking with a deep association metric. In: ICIP (2017)
47. Wu, Y., Lim, J., Yang, M.H.: Online object tracking: a benchmark. In: CVPR (2013)
48. Wu, Y., He, K.: Group normalization. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11217, pp. 3–19. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01261-8_1
49. Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K.: Aggregated residual transformations for deep neural networks. In: CVPR (2017)
50. Yang, Y., Wen, L., Lyu, S., Li, S.Z.: Unsupervised learning of multi-level descriptors for person re-identification. In: AAAI (2017)
51. Zhan, Y., Wang, C., Wang, X., Zeng, W., Liu, W.: A simple baseline for multi-object tracking. arXiv (2020)
52. Zhao, L., Li, X., Zhuang, Y., Wang, J.: Deeply-learned part-aligned representations for person re-identification. In: ICCV (2017)
53. Zhao, R., Ouyang, W., Wang, X.: Unsupervised salience learning for person re-identification. In: CVPR (2013)
54. Zhou, K., Yang, Y., Cavallaro, A., Xiang, T.: Omni-scale feature learning for person re-identification. In: ICCV (2019)
55. Zhou, Q., et al.: Graph correspondence transfer for person re-identification. In: AAAI (2018)

56. Zhou, X., Koltun, V., Krähenbühl, P.: Tracking objects as points. arXiv (2020)
57. Zhu, J., Yang, H., Liu, N., Kim, M., Zhang, W., Yang, M.-H.: Online multi-object tracking with dual matching attention networks. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11209, pp. 379–396. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01228-1_23
58. Zhu, P., Wen, L., Du, D., Bian, X., Hu, Q., Ling, H.: Vision meets drones: past, present and future. CoRR abs/2001.06303 (2020)
59. Zhu, P., et al.: VisDrone-VDT2018: the vision meets drone video detection and tracking challenge results. In: Leal-Taixé, L., Roth, S. (eds.) ECCV 2018. LNCS, vol. 11133, pp. 496–518. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-11021-5_29