

A Review of Detection-related Multiple Object Tracking in Recent Times

Suya Li[†], Ying Cao*, Xin Xie[†]

[†]Henan University, School of Computer and Information Engineering, Jinming Campus, Kaifeng, China

*Corresponding Author

Email: henu_work_cy@163.com

Henan University,

Henan Key Laboratory of Big Data Analysis and Processing,

School of Computer and Information Engineering, Jinming Campus, Kaifeng, China

Abstract—Multi-object tracking (MOT) is garnering more attention due to its widespread application in the area of autonomous driving, human-computer interaction, and intelligent video surveillance. Especially in recent years, MOT has rapidly developed thanks to related technologies such as object detection, which has helped in handling interfering factors such as crowded scene occlusion, small objects, and similar appearances. Among these, Detection-based MOT is the mainstream for accurately forming objects' trajectories. Therefore, according to the analysis of the last three years' research, this paper particularly focuses on discussing the continuous optimization strategies of MOT around the development of object detection at each stage. In addition, this article also introduces the commonly used benchmark datasets and related applications of MOT.

Index Terms—Multi-object Tracking (MOT), Object detection, Deep learning, Research progress

I. INTRODUCTION

With the development of Deep-learning(DL)-based Multi-object tracking (MOT), it has been widely utilized in many related real-life problems like Intelligent transportation[1], Video surveillance[2], autonomous driving[3], human-computer interaction[4]. These applications also highlight the important academic significance of MOT. To accomplish the task of MOT, a common way is locating objects' position frame per frame first and associating them inter frames then. There, the tracking objects could be anything such as pedestrians, sports players, birds, dogs, vehicles, or all of the above. Due to the urgent need for continuous improvement of the applications performance and the challenges such as occlusion, small objects, and similar appearances, the schemes in MOT are always being optimized. Among them, detection-related multi-object tracking methods are the mainstream solution for forming the trajectories of objects[5, 6, 7, 8].

Therefore, different from related review papers on MOT in previous years[9, 10, 11], this paper mainly focuses on object detection-related multi-object research, and through the analysis of the methods to divides them into detection-based MOT,

joint detection-based MOT. The core ideas and algorithmic features of each technique are elaborated on separately.

Consequently, to sum up, the work of this paper is organized as follows:

- 1) The development summary of standard object detection algorithms used in MOT;
- 2) The list of tracking-by-detection and joint detection and tracking MOT approaches;
- 3) The overview of MOT datasets;
- 4) The exploration of various applications.

II. OBJECT DETECTION FOR MOT

As the fundamental technology in Computer vision, object detection has been used in much other vision tasks[12] and the related vision application[2, 3, 4]. Also, as the basis of MOT, the performance of object detection is the crucial key. A comprehensive understanding of MOT relies heavily on the development of object detection. Therefore, it is essential to first examine the common object detection. As shown in Fig. 1, the development of object detection could be broadly delineated into two overarching periods: the era of traditional object detection (1998-2014) and the epoch of DL-based (2014-present). The latter could be further sub-categorized into anchor-based (one-stage, two-stage separately) and anchor-free methods.

The conventional object detection algorithms[13, 14, 15] primarily rely on feature extraction methods that are manually crafted[16]. The proposal of RCNN[17] marked the dawn of deep learning in object detection. After filtering related papers, the research could be divided into anchor-based and anchor-free, to be more precise, the anchor-based would be described as two-stage and one-stage methods.

Firstly, two-stage object detection methods: the two-stage methods refer to an object detection approach that involves generating candidates first and then performing classification as well as regression. In this mode, Kaiming et al. aimed for handling the overlapping parts between candidate boxes, and adopted a spatial pyramid pooling layer to process regions of interest in different scales, thereby avoiding redundant

Key Scientific and Technological Project of Henan Province, China232102240023.

* Ying Cao is corresponding author.

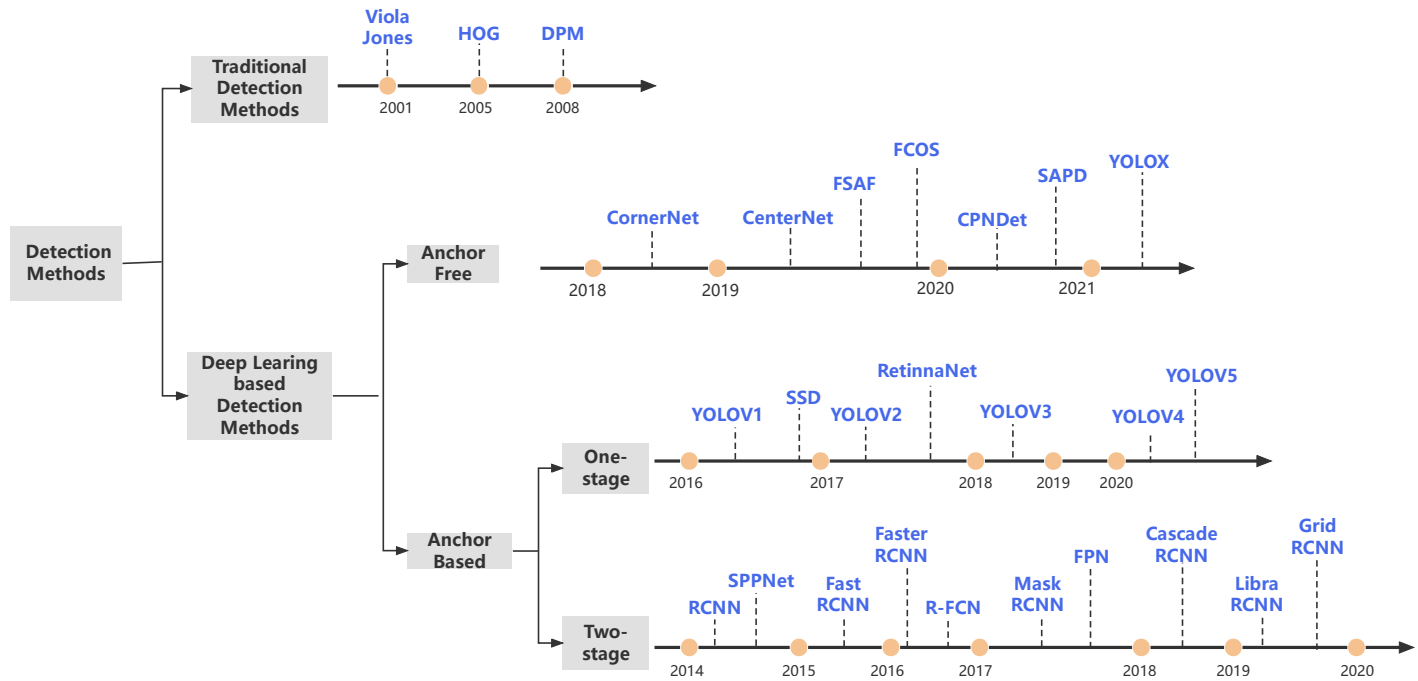


Fig. 1. The development of object detection.

computation in overlapping areas and improving computational efficiency. Meanwhile, they also replaced the fixed size maintain in RCNN with any size of the input image to better preserve the information of input image[18]. To solve the complex training process of SPPNET, Girshick et al. proposed Fast R-CNN, which significantly decreased the computational time by leveraging the Region of Interest (ROI) pooling layer[19]. Subsequently, Ren et al. proposed the Faster R-CNN algorithm, which replaced selective search with the Region Proposal Network (RPN) to further reduce training and testing time while enabling end-to-end training[20]. Later, Lin et al. designed Feature Pyramid Network (FPN) to include both high-level semantic features and low-level visual feature in image features representation for achieving multi-scale object detection[21]. Also based on Faster R-CNN, Cascade R-CNN incorporated the cascaded training to filter more accurate object boxes, thereby achieving higher detection performance[22]. Though the superior performance of these two-stage methods, the complexity, speed, and inaccuracy of candidates may be the limitations of two-stage model.

One-stage object detection methods: compared with two-stage methods, one-stage can achieve high detection speeds while maintaining a high level of accuracy by eliminating the region proposal step. for example: Redmon et al. designed YOLOv1 to predict the class probability and position information for each grid which is generated by input image partition[23]. The same year, Liu et al. utilized the newly introduced techniques of Multi-reference and Multi-resolution detection, along with the employment of multi-scale feature maps, to significantly enhance the precision of multi-scale object detection[24]; Another, YOLOv2[23]

modified the network structure of YOLOv1 with a utilization of a more efficient feature extraction network DarkNet-19[25]. YOLOv3 enhanced the DarkNet-19 as DarkNet-53 and took the advantage of FPN (Feature Pyramid Network) structure to achieve multi-scale prediction[26]. YOLOv4 integrated various techniques to improve detection accuracy[27]. Except for the YOLO series, RetinaNet employed Focal Loss to replace the standard cross-entropy loss function, and through weight adjustment automatically making the model more focused on training hard samples[28].

While anchor-based methods can achieve good performance, the bounding boxes (bbox)-based design would be not appropriated. For instance, the different settings of number, size and aspect ratio of bboxes in various datasets, the imbalanced positive and negative samples caused by large number of bboxes, and the more non-object information covered by unfit bboxes shape, these would all lead to detection performance decreasing. In response to these, researches have been conducted on Anchor-free object detection methods, such as CornerNet[29] transformed object detection into corner points detection problem, and used the accurate corner points prediction of objects to eliminate the need for anchor design and regression processes in traditional object detection methods, thus achieving superior object localization accuracy and multi-scale, occlusion issues handled effectively, as well the reduction of parameter number and computational complexity. CornerNet[29] has made a significant contribution in the anchor-free filed. CenterNet also predicted the center point location of the target, but additionally regressed other target attributes through the center point position to achieve higher accuracy and greater versatility[30]. Followed the high

detection efficiency characteristic of YOLO series, YOLOx adapted the YOLO detector to an anchor-free design and achieved significant performance improvements by integrating other advanced technique[31].

III. MULTI-OBJECT TRACKING

A. Tracking-by-Detection

Tracking-by-Detection (TBD) is a two-stage paradigm for MOT that generated candidates by a detector firstly, and then associated them inter-frame to obtain the objects' motion trajectories[32]. In TBD paradigm, SORT is the classic method, which formed the objects' trajectories by measuring the relationship among frames based on the Intersection over Union (IOU) between tracked detection and predictions. Specifically, relying on the Kalman filter [39] and the Hungarian matching[40], SORT could predict the current position based on the position of the target in the previous frame, and match the detection bboxes with the prediction bboxes. Though SORT is fast, it still limited in handling object occlusion, and this would lead to high number of ID switches[5]. For this, based on the utilization recursive Kalman filter and a frame-by-frame Hungarian matching, in data association, DeepSORT added a cascade matching scheme to enhance the tracking performance in the crowded scenes and reduce the frequency of ID switches in real-time tracking task[6].

Different with DeepSORT[6] which took advantage of Hungarian matching, Guillem et al. proposed a neural network-based solver to automatically learn the features and spatio-temporal variation patterns of targets, thereby better adapting to different tracking scenarios[33]. Another, to addressed the issues of unstable MOT performance caused by a reliance on appearance features, Han et al. proposed the Motion-aware Tracker (MAT) method to pay more attention on motion features[34], and achieve enhanced performance by integrating Integrated Motion Localization (ML), Dynamic Reconnection Context (DRC) and 3D Integral Image (3DII) modules. Among them, ML is designed to solve camera motion and non-rigid object motion, DRC is utilized to deal with long-range motion-based reconnection and interface caused by occlusion or blur, 3DII is employed in association stage to cut useless track-detection association connections with temporal-spatial constraints.

With the introduction and advancement of anchor-free detectors, MOT can also benefit from the superiors of this detection paradigm, including better adaptability to diverse sizes, shapes, and aspect ratios of targets, stronger generalization ability without prior knowledge or manual feature engineering, and other advantages. For example, based on CenterNet[30] or YOLOx[31]: compared to other methods that only associate high-scoring detection bboxes, ByteTrack has improved the accuracy of data association by incorporating low-scoring detection bboxes into the process[7]. Also integrated with YOLOx[31], Yang et al. designed a Cascaded-Buffered Intersection over Union (C-BIoU) tracker to solve the performance reduction of traditional methods with the interference of unreliable appearance features and irregular

motions. Specifically, in C-BIoU's matching stage, instead of using IoU as the above methods, this tracker utilized BiIoU and set a proportion to expand the size of the matching bbox, thereby increasing the chances of properly matching. Additionally, a cascaded matching approach is adapted to prevent BiIoU from expanding arbitrarily and causing mismatching. Based on these to address the issues of tracking failures caused by non-overlapping detections and tracks of identical objects in adjacent frames, as well as mismatches between detection and tracks due to inaccurate motion estimation[35]. Compared to the optimization of C-BIoU in an IoU matching way, based on IoU, Aharon et al. introduced a new tracker, BoT-SORT, which attempted to adapt a simple fusion method of IoU and Re-ID's cosine distance, to establish a stronger correlation between detections and tracks in data association. along with the better bboxes positions obtained by adding camera-motion compensation features and a more accurate Kalman filter[39] state vector, achieving better performance of MOT[36]. Also based on "SORT-like" methods as BoT-SORT[36], to tackle the phenomenon of DeepSORT's[6] under-performance relative to recent existing advanced methods, StrongSORT enhances its tracking capabilities through the equipped modules such as the Appearance-Free Linking model (AFLink) and the Gaussian Smoothed Interpolation algorithm (GSI)[8]. Similarly focused on SORT[5], OC-SORT attempted to optimize the motion model and emphasized the critical role of detection information (i.e., observation) in recovering lost trajectories and reducing the accumulation of Kalman filter errors during occlusion periods, thus improving the overall robustness of MOT in occlusion or non-linear motion scenarios[37]. Furthermore, Deep OC-SORT achieved performance improvement by introducing dynamic visual appearance and camera motion compensation upon OC-SORT[37], as well as the addition of adaptive weighting factors during the trajectory association stage to balance the weights between motion and appearance models in constructing a reasonable cost distance[38].

Tracking-by-detection paradigm could take advantage of the superior of object detection to obtain more detailed information about targets and employ these to assist effectively tracking in occlusion, small objects, or blurred scenes. However, in this two-stage paradigm, the detect-first-then-track execution logic would result in the over-reliance of tracking performance on the detector's ability, and increases the computational complexity.

B. Joint Detection and Tracking

With the rapid advancement of multi-task learning in deep learning, there is increasing interest in one-shot MOT research where the core concept is to perform object detection and re-identification embedding simultaneously in a unified architecture with the aim of reducing inference time. Also according to the used backbone detector, the recent Joint-Detection-and-Tracing (JDT) methods can be presented in two-stage, one-stage, and anchor-free perspectives.

1) *Joint Two-stage Detection and Tracking*: D & T is one of the pioneering approaches that adopted the JDT paradigm to

TABLE I
SUMMARY OF TBD RELATED PAPERS

Algorithm	Year	Detection	Dataset	MOTA(%)
SORT[5]	2016	Faster-RCNN	MOT15, MOT16	33.4, 59.8
DeepSORT[6]	2017	Faster-RCNN	MOT16	61.4
[33]	2019	Faster-RCNN	MOT16, MOT17	58.6, 58.8
MAT[34]	2022	Faster-RCNN	MOT16, MOT17	70.5 69.5
ByteTrack[7]	2021	YOLOX	MOT17, MOT20	80.3, 77.8
StrongSORT[8]	2021	YOLOX	MOT17, MOT20	78.3, 72.2
C-BIou[35]	2022	YOLOX	MOT17, DanceTrack	81.1, 91.6
BoT-SORT[36]	2022	YOLOX	MOT17, MOT20	80.6, 77.8
OC-SORT[37]	2023	YOLOX	MOT17, MOT20	78.0, 75.5
DEEP OC-SORT[38]	2023	YOLOX	MOT17, MOT20	79.4, 75.6

optimize MOT performance through mutual feedback between detection and tracking[41]. Following the JDT paradigm, Tracktor++ additionally integrated a Siamese CNN-based re-ID network and a motion model based on Constant Velocity Model (CMC), Constant Velocity Assumption (CVA) to enhance MOT performance, where CMC was used to address occlusion and camera jitter, and CVA was designed for fast-moving target in low frame rate scenarios in object tracking[42]. Though the importance of motion information, based on a Siamese-like architecture, Tobias et al. created a quasi-dense tracking (QDTrack) to achieve tracking in scenes with significant appearance changes and dense targets only by appearance information effectively utilizing[43]. In contrast to the aforementioned papers' focus on motion or appearance features extraction, Yu et al. mainly considered the feature representation and proposed ReleationTrack. In their scenario, for a better feature presentation, ReleationTrack addressed the issue of feature dependence in the JDT paradigm by introducing Global Context Disentangling (GCD) to decouple the extracted features into task-specific features. additionally, ReleationTrack identified the problem of neglecting global semantic relevance in existing ReID features and designed the Guided Transformer Encoder (GTE) module to learn more globally aware ReID features[44]. In any case, reasonable feature acquisition or effective feature representation would lead to the more efficient development of the current MOT framework.

2) *Joint One-stage Detection with Tracking*: With the introduction and optimization of the one-stage detectors, they were also employed in JDT paradigm for further improvement of MOT performance. For instance: Wang et al. established the Joint-detection-with-tracking (JDE) paradigm to integrate object detection and Re-ID feature extraction in a shared model, their method reduced the inference time and provided a benchmark of JDT[45]. Following with JDE[45], CStrack[46] constructed a Cross-correlation network (CCN)

to focus on the problem of over-competition between detection and ReID, and the performance of both branches was enhanced by cross-relating the two branches with the utilization of an attention mechanism, the Scale-Aware-Attention-Network (SAAN) is also developed to effectively prevent semantic misalignment and improve the association capability of ID embeddings, finally achieving superior MOT performance. As further, CStrackV2[47] re-stored the misclassified targets caused by the detector through the temporal clue for secondary inspection, based on which to effectively recover missed targets and obtain smoother track trajectories. Comparatively, AttTrack adopted the dual-model structure that utilized an attention mechanism to transfer knowledge from the complex network (teacher model) to the lightweight network (student model) for MOT acceleration. This approach further enhancing the tracking accuracy through an interleaved model, which leverages updated predictions from the teacher model as prior knowledge to assist the student model[48].

Also focusing on studying the complexity and invalidity of TBD paradigm, but unlike the above optimizations, RetinaTrack[49] mainly considered the feature representation. Specifically, Lu et al. added post-FPN prediction layers in utilizing the RetinaNet[28], which has demonstrated advantages in small object detection, to obtain instance-level features for each target by splitting different-size features corresponding to different anchor points in advance, by continuously matching the anchor points with instance-level features to solve the problem of incorrect target matching (ID Switch) caused by occlusion.

In contrast to the above methods, which only implemented detection and embedding within a unified framework and actually still relied on two-stage tracking of detection+embedding and data association, Chained-Tracker[50] transformed the data association into a regression task of paired detection bboxes within two frames, and by leveraging the feature enhancement of the joint attention module, Chained-Tracker

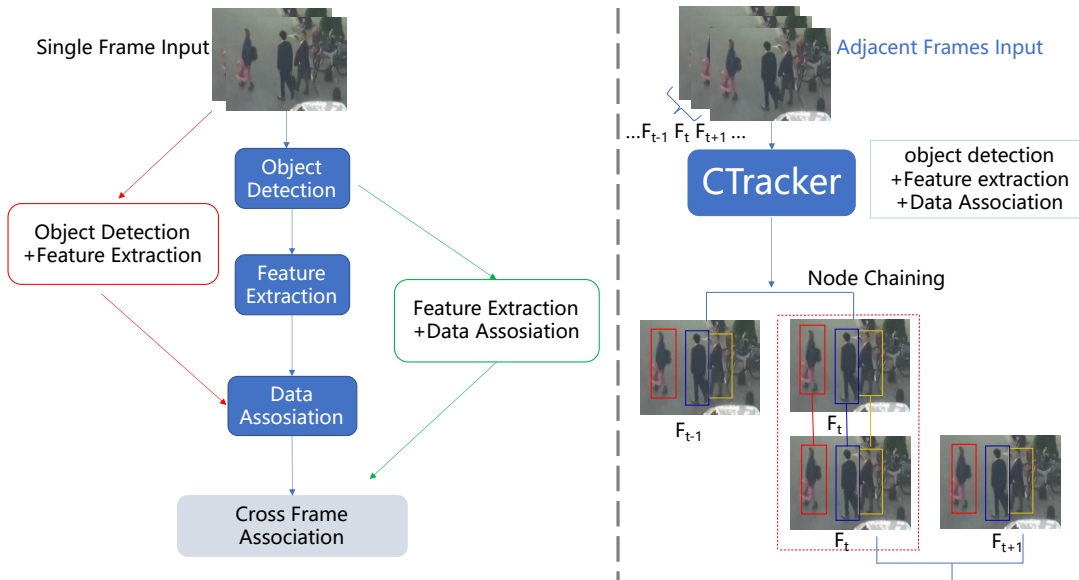


Fig. 2. Comparison between Chained-Tracker and Traditional MOT algorithm.

achieves a full end-to-end implementation of JDT, thereby further improving the performance of MOT[50], the comparison between Chained-Tracker and traditional MOT algorithm is illustrated in Fig. 2.

3) *Joint Anchor-free Detection with Tracking*: Though the one-shot trackers exhibit superior performance, the anchor-based architecture integrated detectors such as YOLO have been found not ideal for learning ReID features. Despite the good results, but also led to a relatively larger number of ID switches. Therefore, tracking with anchor-free detectors has been conducted to address the computational complexity and inefficiency issues for the JDT paradigm. Examples include:

Following JDE, FairMOT[51] proposed an anchor-free approach to tackle the issue of anchor confusion that plagues anchor-based detectors, which impeded the learning of appropriate Re-ID information and identification of the same target. Specifically, the authors integrated the anchor-free detector CenterNet[30] and Re-ID module into two homogeneous branches, this innovation helped to alleviate the problem of over-competition between detection and ReID, resulting in a better balance between the two and the acquisition of high-quality ReID embedding. Contrasting to the aforementioned methods that optimized the branches in JDT, CorrTracker[52] utilized self-supervised learning for training the local correlation module to model the relationships between the target and its surrounding environment, in this way, to strengthen the model's ability in discriminating similar objects, ultimately improving its recognition performance in complex scenarios. Similarly, to improve the discriminability of target features and solve the issue of missed low-confidence detection caused by occlusion or blur, SGT[53] adopted GNN in FairMOT[51], and represented the detection, connection, and relationship features as nodes, edges, edge features. With the aid of edge and edge features, high-order relationship features are generated by

aggregating adjacent detection features and relationships with the current, based on which to enhance the discriminability of detection and recover missed detection. Furthermore, Ren et al. considered the unreliability of coarse-grained global object representation generated according to bbox or center features which were in the presence of occlusion and other interfering factors. FineTrack was designed to adopt a comprehensive approach that describes the object's appearance from both global and local perspectives. Concretely, FineTrack proposed the Flow Alignment FPN (FAFPN) and Multi-head Part Mask Generator (MPMG) to explore multi-dimensional fine-grained feature representation, Among which, FAFPN corrected misalignment and aggregated multi-scale features, MPMG focused on extracting fine-grained representations of the aligned feature map[54].

The JDE-followed solutions mentioned above all demonstrated superior performance. However, JDE paradigm only implements a unified architecture of detection and Re-ID embedding, which is still a two-stage approach that involves detection + embedding and data association. For this, Tokmakov et al. thought of a thoroughly one-stage network, CenterTrack, that combined detection and data association in a more meaningful way. In this model, the effective migration of the detection and data association were undertaken by the utilization of CenterNet[30]. In detail, CenterTrack used the CenterNet[30] to locate the center of the target as a simplified target representation, based on this, model took the current and previous frame image pairs as detector inputs, combined with the previous frame's trajectory heatmap, CenterTrack[55] could predict the center points, object box size, and offset of the targets in the current frame, especially for occluded targets. After that, the greedy matching would be performed relying only on the predicted offset and the distance between center points detected in the previous frame to solve all

TABLE II
SUMMARY OF JDT RELATED PAPERS

Algorithm	Year	Detection	Dataset	MOTA(%)
D&T[41]	2017	R-FCN	ImageNet VID	74.2%(mAP)
Tracktor++[42]	2019	Faster-RCNN	MOT17	61.9
QDTrack[43]	2022	Faster-RCNN	MOT17	64.6
RelationTrack[44]	2022	Faster-RCNN	MOT16, MOT17, MOT20	75.6, 73.8, 67.2
JDE[45]	2019	YOLOv3	MOT16	64.4
CSTrack[46]	2020	YOLOv5	MOT16, MOT17, MOT20	75.6, 74.9, 66.6
CSTrackV2[47]	2021	YOLOv3	MOT16, MOT17, MOT20	76.4, 76.3, 70.7
AttTrack[48]	2022	YOLOv5	MOT15, MOT17	52.9, 43.6
RetinaTrack[49]	2020	RetinaNet	MOT17, Waymo	39.19, 44.92
ChainedTrack[50]	2020	RetinaNet	MOT16, MOT17	67.6, 66.6
FairMOT[51]	2021	CenterNet	MOT16, MOT17, MOT20	74.9, 73.7, 61.8
CorrTracker[52]	2021	CenterNet	MOT16, MOT17, MOT20	76.6, 76.5, 65.2
SGT[53]	2022	CenterNet	MOT16, MOT17, MOT20	76.8, 76.5, 72.8
FineTrack[54]	2023	YOLOX	MOT17, MOT20	80.0, 77.9
CenterTrack[55]	2020	CenterNet	MOT17, KITTI	67.8, 89.44

targets effectively location in each frame and the association of the correct target over time, thus generating trajectories with minimal local input.

Compared to TBD, JDT paradigm could reduce the cases of missed and false detections, improving the accuracy and stability of MOT, and importantly, more suitable for real-time implementation. However, drawbacks still exist, such as the over-competition between detection and tracking[56, 57], the need for sufficient training with dataset preparation, and the performance bottleneck faced with occlusion or motion blur. All of these need further exploration.

IV. MOT BENCHMARKS AND APPLICATION

A. Benchmarks

In recent years, numerous MOT public datasets have been released by universities, companies, and research teams, their published annotated videos along with unified hypotheses, annotations, and evaluation tools ensure the feasibility of validating MOT schemes and promote the development of MOT. Specifically, the development of MOT datasets is shown in Fig. 3.

B. Application

With the development of MOT, it can be integrated in many fields, including but not limited to:

(1) Military Filed

As the deployment of unmanned aerial systems (UAS)[58] continues to rise, the demand for effective MOT in aerial surveillance is increasing. Leveraging MOT can enable precise

positioning and tacking of enemies and military weapons, thereby significantly enhancing combat efficiency.

(2) Video Surveillance

As one of the primary supporting technologies in video surveillance[59], MOT can track and analyze the trajectory of pedestrians, vehicles, and other objects in specific areas, providing various intelligent services such as monitoring abnormal behaviors of pedestrians or vehicles[60], and warning of safety hazards. This technology has played a significant role in enhancing the effectiveness of video surveillance.

(3) Autonomous Driving

MOT is a critical component in the development of autonomous driving technology, it enables autonomous vehicles to effectively predict the subsequent movement of pedestrians, vehicles, or others based on their current trajectories and assist plan their driving paths accordingly. This allows for accurate avoidance of obstacles in the surroundings, preventing collisions and achieving safe, advanced autonomous driving[61].

(4) Medical Diagnosis

MOT can also be used in medical image analysis to perform the tasks such as cell tracking[62], blood vessel tracking[63], neuron tracking[64], and more. This technology can assist doctors in tracking diseases such as tumors and vascular lesions, ultimately enhancing the accuracy of medical diagnoses.

(5) Motion Analysis

MOT is still employed in motion analysis, particularly in sports[65], by tracking and capturing information on athletes' positions and movements, this technology can assist in analyzing and tracking athletes' actions and performance, aiding in performance evaluation and the development of effective

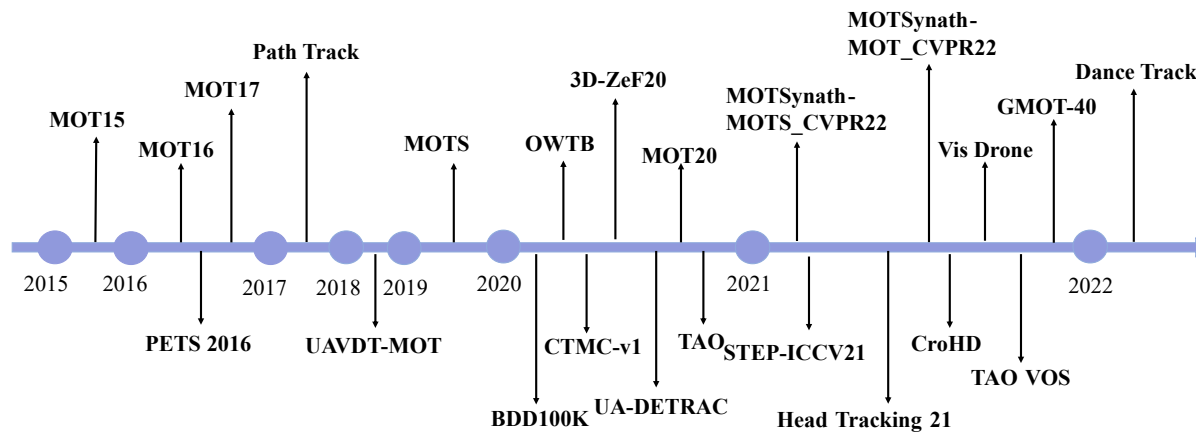


Fig. 3. Development of MOT Dataset.

training plans. Ultimately, leading to improved training outcomes.

V. CONCLUSIONS

According to the analysis of the latest research, this paper particularly focuses on discussing the continuous optimization strategies of MOT around the development of object detection at each stage. Explored the relevant algorithms and characteristics of TBD and JDT methods separately. In addition, we also introduced datasets and application areas related to multi-target tracking. Finally, we hope to provide useful references for researchers in this field.

REFERENCES

- [1] Zheng Tang et al. "Single-Camera and Inter-Camera Vehicle Tracking and 3D Speed Estimation Based on Fusion of Visual and Semantic Features". In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 2018, pp. 108–1087. DOI: 10.1109/CVPRW.2018.00022.
- [2] Jingxuan Hao et al. "A Review of Target Tracking Algorithm Based on UAV". In: *2018 IEEE International Conference on Cyborg and Bionic Systems (CBS)*. 2018, pp. 328–333. DOI: 10.1109/CBS.2018.8612263.
- [3] Hou-Ning Hu et al. "Joint Monocular 3D Vehicle Detection and Tracking". In: *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. 2019, pp. 5389–5398. DOI: 10.1109/ICCV.2019.00549.
- [4] Cyril Robin and Simon Lacroix. "Multi-robot target detection and tracking: taxonomy and survey". In: *Autonomous Robots* (2015), pp. 729–760. DOI: 10.1007/s10514-015-9491-7.
- [5] Alex Bewley et al. "Simple online and realtime tracking". In: *2016 IEEE International Conference on Image Processing (ICIP)*. 2016, pp. 3464–3468. DOI: 10.1109/ICIP.2016.7533003.
- [6] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. "Simple online and realtime tracking with a deep association metric". In: *2017 IEEE International Conference on Image Processing (ICIP)*. 2017, pp. 3645–3649. DOI: 10.1109/ICIP.2017.8296962.
- [7] Yifu Zhang et al. *ByteTrack: Multi-Object Tracking by Associating Every Detection Box*. 2022. arXiv: 2110.06864 [cs.CV].
- [8] Yunhao Du et al. "StrongSORT: Make DeepSORT Great Again". In: *IEEE Transactions on Multimedia* (2023), pp. 1–14. DOI: 10.1109/TMM.2023.3240881.
- [9] Sankar K. Pal et al. "Deep learning in multi-object detection and tracking: state of the art". In: *Applied Intelligence* 51.9 (2021), pp. 6400–6429. ISSN: 1573-7497. DOI: 10.1007/s10489-021-02293-7.
- [10] Gioele Ciaparrone et al. "Deep learning in video multi-object tracking: A survey". In: *Neurocomputing* 381 (2020), pp. 61–88. ISSN: 0925-2312. DOI: 10.1016/j.neucom.2019.11.023. URL: %5Curl%7Bhttps://www.sciencedirect.com/science/article/pii/S0925231219315966%7D.
- [11] Yan Dai et al. "A survey of detection-based video multi-object tracking". In: *Displays* 75 (2022), p. 102317. ISSN: 0141-9382. DOI: 10.1016/j.displa.2022.102317. URL: %5Curl%7Bhttps://www.sciencedirect.com/science/article/pii/S0141938222001354%7D.
- [12] Xiaoding Yuan et al. "Robust Instance Segmentation through Reasoning about Multi-Object Occlusion". In: *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2021, pp. 11136–11145. DOI: 10.1109/CVPR46437.2021.01099.
- [13] P. Viola and M. Jones. "Rapid object detection using a boosted cascade of simple features". In: *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR. Vol. 1*, pp. I–I. ISBN: 1063-6919. DOI: 10.1109/CVPR.2001.990517.

- [14] N. Dalal and B. Triggs. "Histograms of oriented gradients for human detection". In: *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*. Vol. 1. 2005, 886–893 vol. 1. DOI: 10.1109/CVPR.2005.177.
- [15] Pedro Felzenszwalb, David McAllester, and Deva Ramanan. "A discriminatively trained, multiscale, deformable part model". In: *2008 IEEE Conference on Computer Vision and Pattern Recognition*. 2008, pp. 1–8. DOI: 10.1109/CVPR.2008.4587597.
- [16] David G. Lowe. "Distinctive Image Features from ScaleInvariant Keypoints". In: *International Journal of Computer Vision* (2004), pp. 91–110. DOI: 10.1023/b:visi.0000029664.99615.94.
- [17] Ross Girshick et al. "Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation". In: *2014 IEEE Conference on Computer Vision and Pattern Recognition*. 2014, pp. 580–587. DOI: 10.1109/CVPR.2014.81.
- [18] Kaiming He et al. "Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 37.9 (2015), pp. 1904–1916. DOI: 10.1109/TPAMI.2015.2389824.
- [19] Ross Girshick. "Fast R-CNN". In: *2015 IEEE International Conference on Computer Vision (ICCV)*. 2015, pp. 1440–1448. DOI: 10.1109/ICCV.2015.169.
- [20] Shaoqing Ren et al. "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39.6 (2017), pp. 1137–1149. DOI: 10.1109/TPAMI.2016.2577031.
- [21] Chenchen Zhu, Yihui He, and Marios Savvides. "Feature Selective Anchor-Free Module for Single-Shot Object Detection". In: *2019 IEEE /CVF Conference on Computer Vision and Pattern Recognition (CVP R)*. 2019, pp. 840–849. DOI: 10.1109/CVPR.2019.00093.
- [22] Zhaowei Cai and Nuno Vasconcelos. "Cascade R-CNN: Delving Into High Quality Object Detection". In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2018, pp. 6154–6162. DOI: 10.1109/CVPR.2018.00644.
- [23] Joseph Redmon et al. "You Only Look Once: Unified, Real-Time Object Detection". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 779–788. DOI: 10.1109/CVPR.2016.91.
- [24] Wei Liu et al. "SSD: Single Shot MultiBox Detector". In: *Computer Vision (ECCV) 2016*. Springer International Publishing, 2016, pp. 21–37. DOI: 10.1007/978-3-319-46448-0_2.
- [25] Joseph Redmon and Ali Farhadi. "YOLO9000: Better, Faster, Stronger". In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017, pp. 6517–6525. DOI: 10.1109/CVPR.2017.690.
- [26] Joseph Redmon and Ali Farhadi. *YOLOv3: An Incremental Improvement*. 2018. arXiv: 1804.02767 [cs.CV].
- [27] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. *YOLOv4: Optimal Speed and Accuracy of Object Detection*. 2020. arXiv: 2004.10934 [cs.CV].
- [28] Tsung-Yi Lin et al. "Focal Loss for Dense Object Detection". In: *2017 IEEE International Conference on Computer Vision (ICCV)*. 2017, pp. 2999–3007. DOI: 10.1109/ICCV.2017.324.
- [29] Hei Law and Jia Deng. "CornerNet: Detecting Objects as Paired Keypoints". In: *International Journal of Computer Vision* 128.3 (2020), pp. 642–656. ISSN: 1573-1405. DOI: 10.1007/s11263-019-01204-1.
- [30] Kaiwen Duan et al. "CenterNet: Keypoint Triplets for Object Detection". In: *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. 2019, pp. 6568–6577. DOI: 10.1109/ICCV.2019.00667.
- [31] Zheng Ge et al. *YOLOX: Exceeding YOLO Series in 2021*. 2021. arXiv: 2107.08430 [cs.CV].
- [32] Zhihong Sun et al. "A Survey of Multiple Pedestrian Tracking Based on Tracking-by-Detection Framework". In: *IEEE Transactions on Circuits and Systems for Video Technology* 31.5 (2021), pp. 1819–1833. DOI: 10.1109/TCSVT.2020.3009717.
- [33] Guillem Brasó and Laura Leal-Taixé. "Learning a Neural Solver for Multiple Object Tracking". In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020, pp. 6246–6256. DOI: 10.1109/CVPR42600.2020.00628.
- [34] Shoudong Han et al. "MAT: Motion-aware multi-object tracking". In: *Neurocomputing* 476 (2022), pp. 75–86. ISSN: 0925-2312. DOI: 10.1016/j.neucom.2021.12.104. URL: %5Curl%7Bhttps://www.sciencedirect.com/science/article/pii/S0925231221019627%7D.
- [35] Fan Yang et al. "Hard to Track Objects with Irregular Motions and Similar Appearances? Make It Easier by Buffering the Matching Space". In: *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. 2023, pp. 4788–4797. DOI: 10.1109/WACV56688.2023.00478.
- [36] Nir Aharon, Roy Orfaig, and Ben-Zion Bobrovsky. *BoT-SORT: Robust Associations Multi-Pedestrian Tracking*. 2022. arXiv: 2206.14651 [cs.CV].
- [37] Jinkun Cao et al. *Observation-Centric SORT: Rethinking SORT for Robust Multi-Object Tracking*. 2023. arXiv: 2203.14360 [cs.CV].
- [38] Gerard Maggolino et al. *Deep OC-SORT: Multi-Pedestrian Tracking by Adaptive Re-Identification*. 2023. arXiv: 2302.11813 [cs.CV].
- [39] Du Yong Kim and Moongu Jeon. "Data fusion of radar and image measurements for multi-object tracking via Kalman filtering". In: *Information Sciences* 278 (2014), pp. 641–652. ISSN: 0020-0255. DOI: 10.1016/j.ins.2014.03.080. URL: %5Curl%7Bhttps://www.sciencedirect.com/science/article/pii/S0020025514003715%7D.

- [40] H. W. Kuhn. "The Hungarian method for the assignment problem". In: *Naval Research Logistics Quarterly* 2.1-2 (1955), pp. 83–97. DOI: 10.1002/nav.3800020109. eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/nav.3800020109>. URL: <https://onlinelibrary.wiley.com/doi/abs/10.1002/nav.3800020109%7D>.
- [41] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. *Detect to Track and Track to Detect*. 2018. arXiv: 1710.03958 [cs.CV].
- [42] Philipp Bergmann, Tim Meinhardt, and Laura Leal-Taixé. "Tracking Without Bells and Whistles". In: *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. 2019, pp. 941–951. DOI: 10.1109/ICCV.2019.00103.
- [43] Tobias Fischer et al. *QDTrack: Quasi-Dense Similarity Learning for Appearance-Only Multiple Object Tracking*. 2022. arXiv: 2210.06984 [cs.CV].
- [44] En Yu et al. "RelationTrack: Relation-aware Multiple Object Tracking with Decoupled Representation". In: *IEEE Transactions on Multimedia* (2022), pp. 1–1. DOI: 10.1109/TMM.2022.3150169.
- [45] Zhongdao Wang et al. *Towards Real-Time Multi-Object Tracking*. 2020. arXiv: 1909.12605 [cs.CV].
- [46] Chao Liang et al. "Rethinking the Competition Between Detection and ReID in Multiobject Tracking". In: *IEEE Transactions on Image Processing* 31 (2022), pp. 3182–3196. DOI: 10.1109/TIP.2022.3165376.
- [47] Chao Liang et al. *One More Check: Making "Fake Background" Be Tracked Again*. 2021. arXiv: 2104.09441 [cs.CV].
- [48] Keivan Nalaie and Rong Zheng. "AttTrack: Online Deep Attention Transfer for Multi-object Tracking". In: *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. 2023, pp. 1654–1663. DOI: 10.1109/WACV56688.2023.00170.
- [49] Zhichao Lu et al. "RetinaTrack: Online Single Stage Joint Detection and Tracking". In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020, pp. 14656–14666. DOI: 10.1109/CVPR42600.2020.01468.
- [50] Jinlong Peng et al. *Chained-Tracker: Chaining Paired Attentive Regression Results for End-to-End Joint Multiple-Object Detection and Tracking*. 2020. arXiv: 2007.14557 [cs.CV].
- [51] Zhang et al. "FairMOT: On the Fairness of Detection and Re-identification in Multiple Object Tracking". In: *International Journal of Computer Vision* (2021), pp. 3069–3087. DOI: 10.1007/s11263-021-01513-4.
- [52] Ying Wang et al. "Learning Correlation for Online Multiple Object Tracking". In: *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2022, pp. 4713–4717. DOI: 10.1109/ICASSP43922.2022.9746986.
- [53] Jeongseok Hyun et al. "Detection Recovery in Online Multi-Object Tracking with Sparse Graph Tracker". In: *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. 2023, pp. 4839–4848. DOI: 10.1109/WACV56688.2023.00483.
- [54] Hao Ren et al. *Focus On Details: Online Multi-object Tracking with Diverse Fine-grained Representation*. 2023. arXiv: 2302.14589 [cs.CV].
- [55] Xingyi Zhou, Vladlen Koltun, and Philipp Krähenbühl. *Tracking Objects as Points*. 2020. arXiv: 2004.01177 [cs.CV].
- [56] Jieming Yang et al. "Online multi-object tracking using multi-function integration and tracking simulation training". In: *Applied Intelligence* (2021), pp. 1268–1288. DOI: 10.1007/s10489-021-02457-5.
- [57] Haidong Wang et al. "JDAN: Joint Detection and Association Network for Real-Time Online Multi-Object Tracking". In: *ACM Trans. Multimedia Comput. Commun. Appl.* 19.1s (Feb. 2023). ISSN: 1551-6857. DOI: 10.1145/3533253.
- [58] Giancarmine Fasano et al. "Sense and avoid for unmanned aircraft systems". In: *IEEE Aerospace and Electronic Systems Magazine* 31.11 (2016), pp. 82–110. DOI: 10.1109/MAES.2016.160116.
- [59] Yuhao Luo et al. "Pedestrian tracking in surveillance video based on modified CNN". In: *Multimedia Tools and Applications* (2018), pp. 24041–24058. DOI: 10.1007/s11042-018-5728-8.
- [60] Kuei-Hsiang Chao and Pi-Yun Chen. "An Intelligent Traffic Flow Control System Based on Radio Frequency Identification and Wireless Sensor Networks". In: *International Journal of Distributed Sensor Networks* 10.5 (2014), p. 694545. ISSN: 1550-1329. DOI: 10.1155/2014/694545.
- [61] Ming Gao et al. "Multiple object tracking using a dual-attention network for autonomous driving". In: *IET Intelligent Transport Systems* (2020), pp. 842–848. DOI: 10.1049/iet-its.2019.0536.
- [62] Yoonas Imani et al. "A new method for multiple sperm cells tracking". In: *Journal of medical signals and sensors* 4.1 (2014), pp. 35–42.
- [63] Dengqiang Jia and Xiahai Zhuang. "Learning-based algorithms for vessel tracking: A review". In: *Computerized Medical Imaging and Graphics* 89 (2021), p. 101840. ISSN: 0895-6111. DOI: 10.1016/j.compmedimag.2020.101840. URL: <https://www.sciencedirect.com/science/article/pii/S089561112030135X%7D>.
- [64] Linus Manubens-Gil et al. "BigNeuron: a resource to benchmark and predict performance of algorithms for automated tracing of neurons in light microscopy datasets". In: *Nature Methods* (2023). ISSN: 1548-7105. DOI: 10.1038/s41592-023-01848-5.
- [65] Sungwon Moon et al. "A comparative study on multi-object tracking methods for sports events". In: *2017 19th International Conference on Advanced Communication Technology (ICACT)*. 2017, pp. 883–885. DOI: 10.23919/ICACT.2017.7890221.