一体化多目标跟踪算法研究综述

周 雪[1,2*]，梁 超[2]，何均洋[2]，唐瀚林[2]

(1. 电子科技大学 (深圳) 高等研究院 广东 深圳 518110；
2. 电子科技大学自动化工程学院 成都 611731)

【摘要】视觉多目标跟踪算法 (MOT) 一直是计算机视觉与视频图像智能分析领域的一个研究热点。近年来，随着深度学习的发展及实际应用需要，越来越多性能优异的一体化多目标跟踪算法被提出，受到研究者的青睐。对近年来广受关注的一体化多目标跟踪算法进行了系统性的综述。从不同的一体化构建思路出发，梳理包括构建出发点、框架设计、方法优缺点、研究趋势等方面的内容，并在权威的 MOT Challenge 系列数据集上进行性能比较，定量地分析不同的一体化方法的优势和局限性。最后，结合研究现状，提出了一体化多目标跟踪需要重点关注的若干问题及未来展望。

关 键 词 数据关联； 多目标跟踪； 目标检测； 一体化深度神经网络； 单目标跟踪
中图分类号 TP391 **文献标志码** A doi:10.12178/1001-0548.2021349

A Survey on One-Shot Multi-Object Tracking Algorithm

ZHOU Xue[1,2*], LIANG Chao[2], HE Junyang[2], and TANG Hanlin[2]

(1. Shenzhen Institute for Advanced Study, University of Electronic Science and Technology of China Shenzhen Guangdong 518110;
2. School of Automation Engineering, University of Electronic Science and Technology of China Chengdu 611731)

Abstract Visual multiple object tracking (MOT) has become a hot issue in computer vision and intelligent analysis of video images. In recent years, with the development of deep learning and practical application needs, more and more one-shot MOT algorithms with outstanding performance have been proposed, attracting much attention from researchers. This paper systematically reviews the popular one-shot MOT algorithms. From different construction ideas, the paper summarizes the motivation, framework design, strengths and weaknesses of methods, research trends, etc. Afterwards, we compare the performances of the one-shot MOT algorithms on the public testing set MOT Challenge, and quantitatively analyze the advantages and limitations of different one-shot methods. Finally, some future thoughts, foresight, and key issues that need to be focused on are introduced based on the research status.

Key words data association; multiple object tracking; object detection; one-shot deep neural network; single object tracking

随着计算机技术的发展和人工智能技术的日益成熟，通过计算机视觉来替代人类视觉系统对视频数据进行分析理解的趋势愈加明显。多目标跟踪 (multiple object tracking, MOT)是视频分析理解的热门问题之一，其结合了模式识别、机器学习、计算机视觉、图像处理以及计算机应用等多个学科，构成了一种多目标定位和运动轨迹预测的技术。在智能监控、行为分析、人机交互、体育分析、智能驾驶系统等领域中，多目标跟踪技术有着广泛的应用前景及巨大的潜在经济价值。

在过去数十年里，多目标跟踪技术取得了极大的发展，也涌现出很多优秀的方法。早期的一些工作[1-3]尝试将多目标跟踪建模成多个单目标跟踪任务独立进行，这是一种很直观的解决方法。然而，在多目标跟踪场景中会面临着更加复杂的问题，如目标的频繁遮挡、目标突然出现或消失、目标具有相似的外观等，仅依靠单目标跟踪器很难在该场景下实现鲁棒的跟踪。随着深度学习的发展和高性能

收稿日期：2021－11－22；修回日期：2022－03－24
基金项目：国家自然科学基金 (61972071，U20A20184)；四川省科技计划 (2020YJ0036)；厅市共建智能终端四川省重点实验室开放课题 (SCITLAB-1005)
作者简介：周雪 (1981－)，女，博士，教授，主要从事模式识别及计算机视觉方面的研究.
*通信作者：周雪，E-mail：zhouxue@uestc.edu.cn

检测器的出现，文献 [4-9] 发现基于检测的多目标跟踪 (tracking-by-detection) 在各个场景都可以取得更好的鲁棒性。这类方法将多目标跟踪任务分为两个单独的子任务，即检测和数据关联。第一步是通过高性能的检测器[10-13] 获得每一个目标的目标框预测。第二步是基于重识别 (re-identificaion, ReID)[14-15]、运动预测[16-18] 等方法，构建与目标相关的信息来实现帧间匹配，以形成轨迹。这类方法至今依然在多目标跟踪算法中占据着“统治”地位。虽然基于检测的多目标跟踪方法性能优异，但是堆叠多个模块构成的系统也带来巨大的计算量，并不利于实际应用。为了平衡速度与精度，文献 [19-24] 将注意力转移到如何构建一体化的多目标跟踪模型上，这也是目前多目标跟踪研究的新趋势。

随着多目标跟踪研究的推进，近年来也有不少工作对多目标跟踪研究进行综述。已有综述可分为 3 类：第一类主要从多目标跟踪的模块组成出发，探讨多目标跟踪各组成部分的研究进展[25-27]；第二类梳理了已有的多目标跟踪算法，并进行分类概述[28-29]；第三类主要围绕多目标跟踪中的数据关联方法[30-31] 进行讨论。不同于先前的工作，本文聚焦于多目标跟踪一体化研究进展，对近年来广受关注的一体化多目标跟踪算法进行了系统性地综述。从不同的一体化构建思路出发，梳理包括构建出发点、框架设计、方法优缺点、研究趋势等方面的内容，并结合公开数据集[32-34] 对比分析已有的一体化多目标跟踪方法的优势和局限性，为相关领域做进一步研究提供参考。

1　多目标跟踪系统组成

当前多目标跟踪系统主要分为 4 个模块，即检测、外观建模、运动建模和数据关联。模块间的相互关系如图 1 所示，其中蓝色实线表示定位信息传输，红色实线表示匹配信息。

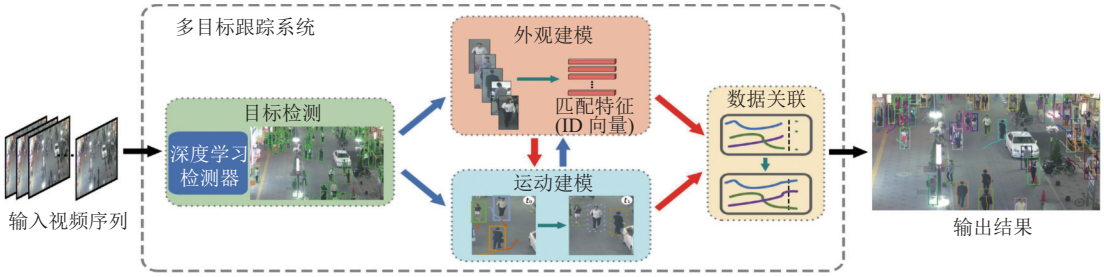

图 1　多目标跟踪系统中 4 大模块相互关系示意图

1) 检测：在多目标跟踪系统中对场景中出现的目标进行定位与尺度预测，其结果往往对多目标跟踪性有决定性影响。当前较先进的方法 (state-of-the-art, SOTA) 的解决方案均采用了基于深度学习的高性能检测框架，如 Faster R-CNN[10]、CenterNet[35]、YOLOX[36] 等。

2) 外观建模：指利用目标的视觉外观，构建具有判别性的匹配特征。当前主流的多目标跟踪方法[6-9] 均采用先进的重识别技术 (ReID)[14-15]，通过深度卷积网络将每一个目标抽象为一个具有高阶判别语义的匹配特征，实现外观建模。该技术有效地弥补时空匹配[4-5] 在密集场景造成的错误匹配，是当前提高算法数据关联能力最有效的方法之一。

3) 运动建模：指通过目标已有的运动状态建立运动模型，以预测目标下一帧可能出现的位置。在多目标跟踪中，运动建模具有极大的应用和研究价值。当前方法主要通过引入 Kalman 滤波[16] 和单目标跟踪器 (single object tracking, SOT)[37-38] 来建模目标的运动信息，不仅有利于通过时序挖掘召回检测器漏检的部分目标，提高了目标定位的鲁棒性，也能利用时序信息增强算法的匹配能力，降低目标漂移的发生概率。

4) 数据关联：指的是利用目标的运动，外观等信息建立跨帧目标间的相似性，找到最优的匹配关系。目前主流的数据关联方式通过匈牙利算法[39] 和正则化策略来计算总代价最小的两两匹配关系作为最优解。也有一些工作引入图卷积神经网络[40] 来替代匈牙利算法，通过深度学习的方式构建更鲁棒的匹配关系生成器。

2　多目标跟踪一体化研究进展

随着研究进一步深入，近年来很多工作关注于如何联合上述两个或多个模块来构建一体化多目标跟踪算法。这些方法无论在处理速度还是性能上都表现优异，受到了广泛关注。下面，将从不同的一体化构建思路出发展开论述。

2.1 联合检测和外观建模的一体化方法

该类方法的出发点是构建一个可以同时输出目标定位和外观匹配信息的模型。具体来说，对于第t帧的图像输入x_t，仅通过一个统一的模型ψ进行处理就可以输出两种任务信息，其具体操作如下：

$$D_t=\psi(x_t),\quad E_t=\psi(x_t) \tag{1}$$

式中，D_t为第t帧目标框的集合；E_t表示D_t中目标框所对应的匹配信息集合。由于减少了反复调用外观模型对每个目标单独提取特征所带来的巨额计算量，该类一体化方法极大地提高了推理速度，甚至在一些非密集场景用单张 GPU 可以实现实时推理。

联合检测和外观建模的一体化方法[41]在两阶段检测器 Faster R-CNN[10]的基础上添加额外的全连接层来提取用于匹配的外观信息。JDE[19]通过重新设计一阶段检测器 YOLOv3[11]的输出模块，实现定位和匹配信息的同步输出。上述两个工作通过输出结构的改进，简单有效地将匹配特征提取融入到不同检测框架中，后续方法均参考该构建思路进行进一步扩展或改进。RetinaTrack[42]在上述思路的基础上设计了多分支头，在不同分辨率的特征图上安排 K 个锚点框，减少目标重叠带来的歧义。FairMOT[43]认为密集的锚点 (anchor) 设置会带来多目标跟踪性能的下降，为此基于无锚框检测器 CenterNet[35]搭建了一体化网络。文献 [20] 分析了一体化模型中检测和 ReID 任务所存在的本质矛盾，指出了这种矛盾导致特征学习存在歧义而造成性能下降。为了解决该问题，CSTrack[20]引入了互相关网络，通过共性和差异性学习促使网络生成与任务相关的特征，有效提高了一体化方法的性能。文献 [44] 提出了 QDTrack，通过在真实标签附近密集采样上百个目标框用于相似性学习，以提高所提取外观特征的匹配能力。

总的来看，该类方法巧妙地统一检测和外观建模，有效地提高了多目标跟踪的效率。然而，当前方法依然强依赖于检测器所获得的检测结果。如果检测结果不可靠，出现漏检、误检的情况，往往会造成轨迹断裂或错误匹配。

2.2 联合检测和运动建模的一体化方法

联合检测和运动建模的一体化方法本质是赋予检测器运动建模的能力。在模型输入上，除了当前帧的图像x_t之外，还需将上一帧的目标定位D_{t-1}一起输入到模型ψ之中，通过模型的运动建模能力实现已有目标的跨帧传播。其操作可总结为：

$$D_t=\psi(x_t,D_{t-1}),\quad D_t=(D_t^{\text{pro}},D_t^{\text{new}}) \tag{2}$$

式中，D_{t-1}和D_t分别表示第$t-1$帧和第t帧的目标框集合。当前帧的结果D_t由两部分组成，一部分是运动建模获得的迁移结果D_t^{pro}，另一部分是由检测器检测到的新出现目标D_t^{new}。

Tracktor[23]首次采用上述思路将检测器转换为跟踪器，利用 Faster R-CNN[10]第二阶段网络的回归能力实现将上一帧目标框传播到当前帧，以一种简单高效的方式将检测器转化为跟踪器。受到 Tracktor 的启发，文献 [24] 基于无锚框检测器 CenterNet[35]搭建了跟踪器 CenterTrack，把目标逐帧传播设定为中心点跟踪问题，通过预测点逐帧的偏移量实现多目标跟踪。虽然上述框架十分简洁，基于检测器的一次前向传播就可完成多目标跟踪，但是其缺点也很明显，即在一些长距离跟踪或者遮挡场景并不鲁棒。针对该问题，基于 CenterTrack 的框架引入卷积门控循环单元[45]，文献 [46] 提出了 PermaTrack。通过编码输入视频中目标的时空演化，PermaTrack可以推断部分或完全遮挡目标的位置，提高了多目标跟踪在遮挡场景的鲁棒性。随着 Transformer 网络[47]的兴起，文献 [48] 提出了 TransTrack，将 Transformer 检测框架 Dert[49]扩展成为一种基于键值 (Key) 查询的运动预测模型，实现了目标的迁移传播。

除了直接利用检测器的回归能力之外，另一种思路是考虑将先进的单目标跟踪融入到检测器之中构建一体化网络。如文献 [50] 提出了 SOTMOT，在 CenterNet[35]的基础上增加一个额外的单目标跟踪分支，通过先进的岭回归目标跟踪方式[51]实现多个目标的运动传播。文献 [52] 提出了 SiamMOT，在 Faster-RCNN[10]的基础上引入了孪生网络跟踪[17]。通过候选区域生成网络，SiamMOT 可直接在编码后的特征上获取每个目标的特征和对应检索区域，并利用互相关操作预测目标在帧间的移动情况。

上述方法高效地将运动建模融入到检测器中，提高了目标一致性预测的鲁棒性，然而在长时或复杂的跟踪场景中，目标运动无法提供可靠的匹配信息，依然存在目标漂移的风险。

2.3 联合检测、外观和运动建模的一体化方法

虽然上述两种一体化思路无论在精度还是处理速度上都取得了 SOTA 的性能，但是其局限性也很明显。可见，运动建模和外观特征是人类观测和跟踪一个物体必须考虑的两方面信息，仅考虑其中一

者难以应对复杂多变的实际场景。因此，为了进一步提高多目标跟踪的性能，后续工作将检测、外观建模和运动建模集成到一个网络中。

文献 [53] 提出 CorrTrack，在联合检测和外观建模的一体化方法 FairMOT[43] 的基础上融入了时空信息，通过局部自注意力的方式建模了目标与周围环境之间的时空拓扑关系，提高了一体化模型的跟踪性能。文献 [54] 提出了 FUFET，采用金字塔光流法[55] 预估目标在场景中的运动情况，弥补了单一外观特征带来的局限，进一步提高了不同帧目标匹配的一致性。文献 [56] 将 CenterTrack[24] 预测目标偏移量的思路融入到联合检测和外观建模的方法中，提出了 TraDeS。TraDeS 利用跟踪线索增强了模型目标检测和分割的性能。文献 [57] 设计了一种轻量化的再查询网络，巧妙地扩展用于匹配的外观特征，以一种极低的开销建模多个目标的时序线索。在当前广受欢迎的联合检测和外观建模的一体化方法上 (如 FairMOT[43] 和 CSTrack[20])，该模块以极小的代价显著提高其跟踪性能。虽然这类方法的已有成果较少，但是其高性能和优异的实时性也正吸引着越来越多的学者投入到其研究中。

2.4　基于视频片段输入的端到端方法

随着基于视频的目标检测技术的发展，一些研究者也关注于是否可以基于视频片段输入来构建端到端的多目标跟踪框架。该类方法目的是通过自动处理一段视频序列输入，直接生成多个目标的运动轨迹和定位信息，不再需要引入额外的数据关联模型或步骤。其具体操作可被总结为：

$$T = \psi(S) \tag{3}$$

式中，ψ 表示端到端的一体化模型；S 表示视频片段输入；T 为所输出的目标定位和匹配结果的集合。

TubeTK[22] 引入 3D 卷积对视频输入进行编码，直接预测目标的时空位置和运动轨迹。由于全局的信息引入，TubeTK 在克服遮挡方面表现出色。CTracker[21] 构建了一种链式的跟踪方法，将目标检测、特征提取、数据关联 3 个模块集成到单个网络中。具体来说，CTracker 将相邻两帧图像建模为一个节点，并将整个视频序列拆分为通过重复帧链接的节点链。通过对节点进行处理，模型可以直接预测相邻帧目标的两两匹配关系和目标定位信息，并通过链接结构完成长时轨迹的预测。虽然该类方法的已有工作较少且性能较低，但是其简单高效的多目标跟踪实现方式，也提供了一种一体化训练和跟踪的新思路。

3　实验分析

本章通过实验定量且定性地分析不同一体化方法的性能表现。在比较不同方法的性能之前，本章首先介绍测试所用的数据集以及评价指标。

3.1　数据集

为了公平比较，采用权威的 MOT Challenge 系列数据集进行测评。MOT Challenge 系列数据集虽然不是最早的一个数据集，但是因为其提供了更丰富的测试场景和更公平的测评环境，自 2015 年后提出的多目标跟踪方法普遍在该数据集上做横向或纵向比较。目前，MOT Challenge 官方针对于行人类别共发布了 4 个数据集供研究者进行研究，分别为 MOT15[32]、MOT16[33]、MOT17[33]、MOT20[34]，其详细情况如表 1 所示。

表 1　MOT Challenge 系列数据集

数据集	视频序列	平均时长/s	轨迹数/序列	目标数/帧	公开检测器	特点
MOT15[32]	22	45.3	55.5	9.0	ACF[32]	场景丰富，密集程度低
MOT16[33]	14	33.1	91.1	20.9	DPM[13]	提高了目标的密度，规范了标注
MOT17[33]	14	33.1	95.1	26.7	DMP[12]、SDP[13]、Faster R-CNN[10]	更关注一些难样本，且提供更多的公开检测器
MOT20[34]	8	66.9	479.1	156.8	Faster R-CNN[10]	密集程度最高，相互遮挡严重

3.2　评价指标

多目标跟踪是一项需要精确定位和长时间关联的任务，评价非常复杂，往往很难用单个指标概括整个系统的性能。目前，在 MOT Challenge 的线上评估系统中有一套公认的指标来评价多目标跟踪系统，主要由 CLEAR MOT 指标[58] 和 ID 指标[59] 构成。本文采用当前研究中最常用的多目标跟踪精度 (multiple object tracking accuracy, MOTA)[58] 和目标识别准确度 (identification F1 score, IDF1)[59] 作为主要评价指标。此外，考虑到不同一体化方法构建的出发点存在差异，为了更直观地评价，本文采用了更多指标，如引入最多跟踪目标数 (mostly tracked,

MT)、最多丢失目标数 (mostly lost, ML)、漏检数量 (false negatives, FN) 和误检数量 (false positives, FP) 来补充评价跟踪器的召回能力；引入 ID 切换数 (identification switch, ID Sw.) 来综合评价生成轨迹的连贯性；引入帧率 (Hz) 来评价跟踪器处理速度。

3.3 方法性能分析

当前所提出基于私有检测 (private detection, PD) 的一体化方法广泛采用 MOT16、MOT17 和近年来提出的 MOT20 进行测试及横向比较。因此，为了保证实验的权威和公平性，采用上述 3 个基准进行实验数据分析。表 2～表 4 根据 MOTA 排序，列举了当前主流的一体化方法和一些经典的多模型堆叠方法在私有检测赛道的性能指标。其中，表 2 的多模型方法用黑体标识。

表 2 MOT16 上基于私有检测的方法性能比较

方法	发布情况	MOTA↑	IDF1↑	MT↑	ML↓	FP↓	FN↓	IDS↓	Hz↑
SORT_POI[4]	**ICIP 2016**	**59.8**	**53.8**	**25.4**	**22.7**	**8 698**	**63 245**	**1 423**	**<8.6**
DeepSORT[6]	**ICIP 2017**	**61.4**	**62.2**	**32.8**	**18.2**	**12 852**	**56 668**	**781**	**<6.7**
RAN[8]	**WACV 2018**	**63.0**	**63.8**	**39.9**	**22.1**	**13 663**	**53 248**	**482**	**<1.5**
TubeTK[22]	CVPR 2020	64.0	59.4	33.5	19.4	10 962	53 626	1 117	1.0
JDE[19]	ECCV 2020	64.4	55.8	35.4	20.0	10 642	52 523	1 544	22.2
POI[7]	**ECCV 2016**	**66.1**	**65.1**	**34.0**	**21.3**	**5 061**	**55 915**	**805**	**<5.2**
CTracker[21]	ECCV 2020	67.6	57.2	32.9	23.1	8 934	48 305	1897	6.8
QDTrack[44]	CVPR 2021	69.8	67.1	41.7	19.8	9 861	44 050	1 097	14.0～30.0
TraDeS[56]	CVPR 2021	70.1	64.7	37.3	20.0	8 091	45 210	1 144	17.5
SOTMOT[50]	CVPR 2021	72.1	72.3	44.0	13.2	14 344	34 784	1 681	16.0
FairMOTv2[43]	IJCV 2021	74.9	72.8	44.7	15.9	–	–	–	25.9
CSTrack[20]	Arxiv 2020	75.6	73.3	42.8	16.5	9 646	33 777	1 121	15.8
OMC[57]	AAAI2022	76.4	74.1	46.1	13.3	10 821	31 044	1 296	12.8
FUFET[54]	Arxiv2020	76.5	68.6	52.8	12.3	12 878	28 982	1 026	–
CorrTrack[53]	CVPR 2021	76.6	74.3	47.8	13.3	10 860	30 756	979	14.8

表 3 MOT17 上基于私有检测的方法性能比较

方法	发布情况	MOTA↑	IDF1↑	MT↑	ML↓	FP↓	FN↓	IDS↓	Hz↑
TubeTK[22]	CVPR 2020	63.0	58.6	31.2	19.9	27 060	177 483	4 137	1.0
CTracker[21]	ECCV 2020	66.6	57.4	32.2	24.2	22 284	160 491	5 529	6.8
CenterTrack[24]	ECCV 2020	67.8	64.7	34.6	24.6	18 498	160 332	3 039	22.0
QDTrack[44]	CVPR 2021	68.7	66.3	40.6	21.8	26 589	146 643	3 378	14.0～30.0
TraDeS[56]	CVPR 2021	69.1	63.9	36.4	21.5	20 892	150 060	3 555	17.5
PermaTrack[46]	ICCV 2021	69.5	68.2	46.3	17.7	–	–	–	10.0
SOTMOT[50]	CVPR 2021	71.0	71.9	42.7	15.3	39 537	118 983	5 184	16.0
FairMOTv2[43]	IJCV 2021	73.7	72.3	43.2	17.3	27 507	117 477	3 303	25.9
TransTrack[48]	Arxiv 2021	74.5	63.9	46.8	11.3	28 323	112 137	3 663	10.0
CSTrack[20]	Arxiv 2020	74.9	72.6	41.5	17.5	23 847	114 303	3 567	15.8
FUFET[54]	Arxiv2020	76.2	68.0	51.1	13.6	32 796	98 475	3 237	–
OMC[57]	AAAI2022	76.3	73.8	44.7	13.6	28 894	101 022	3 858	12.8
CorrTrack[53]	CVPR 2021	76.5	73.6	47.6	12.7	29 808	99 510	3 369	14.8

根据不同的探究方向，本小节的对比分析可分为如下几个方面：

1) 多模型与一体化进行比较。本文在 MOT16 的基准上比较多模型堆叠方法和已有的一体化方

法，其中用于比较的多模型方法包括经典的SORT[4](使用 POI 检测结果的版本)、DeepSORT[6]、POI[7] 和 RAN[8]。从表 2 的数据可以分析得到，较早提出的一体化方法 (即 TubeTK[22]、JDE[19] 和 CTracker[21]) 虽然可以取得与多模型堆叠相近的 MOTA 分数，但是在匹配指标 IDF1 和 ID Sw.上依然有较大差距。随着进一步深入研究，从 2020 年开始，基于检测的一体化方法无论是跟踪精度还是匹配性能都获得了巨大提高，取得了绝对的“统治”地位。如现在性能最高的一体化方法 CorrTrack[53] 在多项多目标跟踪指标上已经远远超过了先前最先进的多模型跟踪方法 POI[7]。而在推理速度上，大多数一体化方法基本上都能保证 10～30 FPS 的运行速度，极大地缓解了多模型堆叠方法处理速度慢，不适应实际应用场景的问题。

表 4　MOT20 上基于私有检测的方法性能比较

方法	发布情况	MOTA↑	IDF1↑	MT↑	ML↓	FP↓	FN↓	IDS↓	Hz↑
FairMOTv2[43]	IJCV 2021	61.8	67.3	68.8	7.6	103 440	88 901	5 243	13.2
TransTrack[48]	Arxiv 2021	64.5	59.2	49.1	13.6	28 566	151 377	3 565	–
CorrTrack[54]	CVPR 2021	65.2	69.1	66.4	8.9	79 429	95 855	5 183	8.5
CSTrack[20]	Arxiv 2020	66.6	68.6	50.4	15.5	25 404	144 358	3 196	4.5
SOTMOT[50]	CVPR 2021	68.6	71.4	64.9	9.7	57 064	101 154	4 209	8.5
OMC[57]	AAAI2022	70.7	67.8	56.6	13.3	22 689	125 039	4 041	6.7

2) 不同一体化方法的性能分析。在 MOT16 和 MOT17 的基准上，评估了基于检测的不同一体化方法的性能。所比较的方法可分为 4 类：第一类是联合检测和外观建模的方法，包括 JDE[19]、QDTrack[44]、FairMOTv2[43] 和 CSTrack[20]；第二类是联合检测和运动建模的方法，包括 CenterTrack[24]、PermaTrack[46]、SOTMOT[50] 和 TransTrack[48]；第三类是联合检测、外观和运动建模的方法，包括 TraDeS[56]、FUFET[54]、OMC[57]和 CorrTrack[53]；第四类为基于视频输入的端到端方法，包括 TubeTK[22] 和 CTracker[21]。如表 2 和表 3 的结果所示，单独融合外观或运动信息均可构建出高性能的多目标跟踪器，无论在 MOTA 和 IDF1 上均可取得优异的性能。而对于同时进行外观和运动建模的第三类方法来说，其性能相比于单独考虑一种信息的方法获得进一步提高，取得当前最先进的水平。其性能优异的原因可总结为以下两点，一是有效融合了外观和运动信息以提高模型对物体的感知定位能力，极大地减少了漏检 (FN) 且增强了轨迹的连贯性 (均取得了极高的 MT 指标)；二是在匹配能力上的提高，由于同时考虑了外观和运动信息进行匹配，这类方法相较于其基准模型在 IDF1 指标上取得提升 (如 CSTrack[20] 对比于 OMC[57]，FairMOTv2[43] 对比于 CorrTrack[53])。第四类方法虽然取得了与多模型堆叠方法相近的 MOTA 分数，但是其 IDF1 指标与其他方法相比，依然存在较大差距，还有很大发展空间。

3) 模型对场景适应能力分析。为了分析一体化方法在不同场景的跟踪鲁棒性，本文在以目标密集著称的基准 MOT20 上进行进一步测试。如表 4 所示，联合检测、外观和运动建模的一体化方法 (OMC[57]) 依然取得了最先进的跟踪性能，即 MOTA 分数最高。而对于目标定位能力来说，基于无锚框检测器 CenterNet[35] 的方法，即 FairMOTv2[43]、CorrTrack[53] 和 SOTMOT[50] 可以在密集场景中生成更多的检测框，漏检 (FN) 更少，使得其 MT 的指标远高于其他方法。虽然无锚框检测相对于其他检测思路在密集场景可以获得更高的召回，但同样也带来了误检 (FP) 的急剧增加。大量误检会增加目标漂移发生的可能性，即 ID Sw.增加，同时也会损害多目标跟踪器的性能。此外，受到检测后处理及数据关联策略的影响，在 MOT20 上一体化方法的处理速度基本比 MOT17 要下降 50%。因此，一种针对密集场景的实时一体化方法有待被研究。

4　结束语

本文从多目标跟踪系统的组成出发，综述了近年来一体化多目标跟踪技术的研究进展，并从构建思路、框架结构及方法优缺点等方面对不同的一体化方法进行详细地分析。此外，也在权威的 MOT Challenge 基准上定量且公平地分析了各类方法的

优势和局限性。目前多目标跟踪在一体化方面的研究已经取得了极大突破，无论跟踪性能还是推理速度都取得了显著提高。但是对于可以可靠落地的一体化多目标跟踪方法来说，仍有许多关键性的问题需要深入细致地研究，包括以下几点：

1) 通用多目标跟踪。当前多目标跟踪的研究只围绕单一类别目标 (行人或车辆)，通用多目标跟踪技术的研究进展缓慢。由于人工标注难度大，当前还未有被研究者广泛接受的通用数据集提出，这也是限制该技术发展的主要原因。因此，一个大型通用多目标跟踪数据库的构建是打破现有多目标跟踪算法类别限制的关键。此外，通过无监督的方式迁移从单一类别学习到的知识以构建通用多目标跟踪器，也是值得研究的方向之一。

2) 处理速度实时性。虽然目前有一些高性能的一体化方法可以在 GPU 上达到实时推理的速度，但在移动端等边缘设备中，受到成本和功耗的影响，先进的多目标跟踪方法依然难以实时落地。目前解决该问题存在两种思路：一是通过工程加速和硬件优化加速算法推理；二是通过模型压缩、知识蒸馏等方式以更少的参数实时地实现先进的跟踪性能。

3) 场景泛化能力。受到拍摄角度、场景变化、天气条件等因素的干扰，当前的多目标跟踪方法往往需要在特定场景数据上微调模型参数才能取得优异的性能。然而，在如自动驾驶等实际应用场景中，所需考虑的场景覆盖范围广，很难实时采集数据来维护模型性能。针对该问题，当前一大研究趋势是融入如激光雷达、GPS 等多模态信息来提高模型的场景泛化能力。此外，实时的模型参数在线更新方法，也是该问题的解法之一。

参 考 文 献

[1] XIANG Y, ALAHI A, SAVARESE S. Learning to track: Online multi--object tracking by decision making [C]//Proceedings of the IEEE International Conference on Computer Vision. Santiago: IEEE Computer Society, 2015: 4705-4713.

[2] YAN X, WU X Q, KAKADIARIS I A, et al. To track or to detect? An ensemble framework for optimal selection [C]//European Conference on Computer Vision. Florence: Springer, 2012: 594-607.

[3] ZHANG L, VAN D M L. Structure preserving object tracking[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Portland: IEEE Computer Society, 2013: 1838-1845.

[4] BEWLEY A, GE Z Y, OTT L, et al. Simple online and realtime tracking[C]//2016 IEEE International Conference on Image Processing (ICIP). Phoenix: IEEE, 2016: 3464-3468.

[5] BOCHINSKI E, ISELEIN V, SIKORA T. High-speed tracking-by-detection without using image information [C]//2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). Lecce: IEEE, 2017: 1-6.

[6] WOJKE N, BEWLEY A, PAULUS D. Simple online and realtime tracking with a deep association metric[C]//2017 IEEE International Conference on Image Processing (ICIP). Beijing: IEEE, 2017: 3645-3649.

[7] YU F W, LI W B, LI Q Q, et al. Poi: Multiple object tracking with high performance detection and appearance feature[C]//European Conference on Computer Vision. Amsterdam: Springer, 2016: 36-42.

[8] FANG K, XIANG Y, LI X C, et al. Recurrent autoregressive networks for online multi-object tracking [C]//2018 IEEE Winter Conference on Applications of Computer Vision (WACV). Lake Tahoe: IEEE Computer Society, 2018: 466-475.

[9] ZHOU Z W, XING J L, ZHANG M D, et al. Online multi target tracking with tensor based high order graph matching[C]//2018 24th International Conference on Pattern Recognition (ICPR). Beijing: IEEE, 2018: 1809-1814.

[10] REN S Q, HE K M, GIRSHICK R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 39(6): 1137-1149.

[11] REDMON J, FARHADI A. Yolov3: An incremental improvement[EB/OL]. [2021-12-24]. https://arxiv.org/pdf/1804.02767.pdf.

[12] FELZENSZWALB P F, GIRSHICK R B, MCALLESTER D, et al. Object detection with discriminatively trained part-based models[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2009, 32(9): 1627-1645.

[13] YANG F, CHOI W, LIN Y Q. Exploit all the layers: Fast and accurate cnn object detector with scale dependent pooling and cascaded rejection classifiers[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE Computer Society, 2016: 2129-2137.

[14] ZAGORUYKO S, KOMODAKIS N. Wide residual networks[EB/OL]. [2021-12-24]. https://arxiv.org/pdf/1605.07146.pdf.

[15] LUO H, GU Y Z, LIAO X Y, et al. Bag of tricks and a strong baseline for deep person re-identification[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Long Beach: IEEE Computer Society, 2019: 1487-1495.

[16] WELCH G, BISHOP G. An introduction to the Kalman filter[EB/OL]. [2021-12-25]. https://www.cs.unc.edu/~welch/media/pdf/kalman_intro.pdf.

[17] BERTINETTO L, VALMADRE J, HENRIQUES J F, et al. Fully-convolutional siamese networks for object

tracking[C]//European Conference on Computer Vision. Amsterdam: Springer, 2016: 850-865.
[18] DANELLJAN M, BHAT G, SHAHBAZ K F, et al. Eco: Efficient convolution operators for tracking[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE Computer Society, 2017: 6638-6646.
[19] WANG Z D, ZHENG L, LIU Y X, et al. Towards real-time multi-object tracking[C]//European Conference on Computer Vision. Glasgow: Springer, 2020: 107-122.
[20] LIANG C, ZHANG Z P, LU Y, et al. Rethinking the competition between detection and reid in multi-object tracking[EB/OL]. [2021-12-25]. https://arxiv.org/pdf/2010.12138.pdf.
[21] PENG J L, WANG C G, WAN F B, et al. Chained-tracker: Chaining paired attentive regression results for end-to-end joint multiple-object detection and tracking[C]//European Conference on Computer Vision. Glasgow: Springer, 2020: 145-161.
[22] PANG B, LI Y Z, ZHANG Y F, et al. Tubetk: Adopting tubes to track multi-object in a one-step training model[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE Computer Society, 2020: 6307-6317.
[23] BERGMANN P, MEINHARDT T, LEAL-TAIXE L. Tracking without bells and whistles[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul: IEEE Computer Society, 2019: 941-951.
[24] ZHOU X Y, KOLTUN V, KRÄHENBÜHL P. Tracking objects as points[C]//European Conference on Computer Vision. Glasgow: Springer, 2020: 474-490.
[25] LUO W H, XING J L, MILAN A, et al. Multiple object tracking: A literature review[J]. Artificial Intelligence, 2021, 293: 103448.
[26] CIAPARRONE G, SÁNCHEZ F L, TABIK S, et al. Deep learning in video multi-object tracking: A survey[J]. Neurocomputing, 2020, 381: 61-88.
[27] 徐涛, 马克, 刘才华. 基于深度学习的行人多目标跟踪方法[J]. 吉林大学学报 (工学版), 2021, 51(1): 27-38.
XU T, MA K, LIU C H. Multi object pedestrian tracking based on deep learning[J]. Journal of Jilin University (Engineering and Technology Edition), 2021, 51(1): 27-38.
[28] 张瑶, 卢焕章, 张路平, 等. 基于深度学习的视觉多目标跟踪算法综述[J]. 计算机工程与应用, 2021, 57(13): 55-66.
ZHANG Y, LU H Z, ZHANG L P, et al. Overview of visual multi-object tracking algorithms with deep learning[J]. Computer Engineering and Applications, 2021, 57(13): 55-66.
[29] 李志华, 于杨. 基于检测的多目标跟踪算法综述[J]. 物联网技术, 2021, 11(4): 20-24.
LI Z H, YU Y. Overview of multi-object tracking algorithms with detection[J]. Internet of Things Technologies, 2021, 11(4): 20-24.
[30] 龚轩, 乐孜纯, 王慧, 等. 多目标跟踪中的数据关联技术综述[J]. 计算机科学, 2020, 47(10): 136-144.
GONG X, LE Z C, WANG H, et al. Survey of data association technology in multi-target tracking[J]. Computer Science, 2020, 47(10): 136-144.
[31] 蔡秀梅, 王妍, 卞静伟, 等. 多目标跟踪数据关联算法综述[J]. 西安邮电大学学报, 2021, 26(2): 77-86.
CAI X M, WANG Y, BIAN J W, et al. Overview of multi-target tracking data association algorithms[J]. Journal of Xi'an University of Posts and Telecommunications, 2021, 26(2): 77-86.
[32] LEAL-TAIXÉ L, MILAN A, REID I, et al. Motchallenge 2015: Towards a benchmark for multi-target tracking [EB/OL]. [2021-12-26]. https://arxiv.org/pdf/1504.01942.pdf.
[33] MILAN A, LEAL-TAIXÉ L, REID I, et al. MOT16: A benchmark for multi-object tracking[EB/OL]. [2021-12-26]. https://arxiv.org/pdf/1603.00831.pdf.
[34] DENDORFER P, REZATOFIGHI H, MILAN A, et al. Mot20: A benchmark for multi object tracking in crowded scenes[EB/OL]. [2021-12-26]. https://arxiv.org/pdf/2003.09 003.pdf.
[35] ZHOU X Y, WANG D Q, KRÄHENBÜHL P. Objects as points[EB/OL]. [2021-12-27]. https://arxiv.org/pdf/1904.07850.pdf.
[36] GE Z, LIU S T, WANG F, et al. Yolox: Exceeding yolo series in 2021[EB/OL]. [2021-12-27]. https://arxiv.org/pdf/2107.08430v2.pdf.
[37] MILAN A, REZATOFIGHI S H, DICK A, et al. Online multi-target tracking using recurrent neural networks [C]//Thirty-First AAAI Conference on Artificial Intelligence. San Francisco: AAAI, 2017: 4225-4232.
[38] CHU Q, OUYANG W L, LIU B, et al. Dasot: A unified framework integrating data association and single object tracking for online multi-object tracking[C]//Proceedings of the AAAI Conference on Artificial Intelligence. New York: AAAI, 2020: 10672-10679.
[39] KUHN H W. The Hungarian method for the assignment problem[J]. Naval Research Logistics Quarterly, 1955, 2(1-2): 83-97.
[40] SCARSELLI F, GORI M, TSOI A C, et al. The graph neural network model[J]. IEEE Transactions on Neural Networks, 2008, 20(1): 61-80.
[41] XIAO T, LI S, WANG B C, et al. Joint detection and identification feature learning for person search [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE Computer Society, 2017: 3376-3385.
[42] LU Z C, RATHOD V, VOTEL R, et al. Retinatrack: Online single stage joint detection and tracking[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE Computer Society, 2020: 14656-14666.
[43] ZHANG Y F, WANG C Y, WANG X G, et al. Fairmot: On the fairness of detection and re-identification in multiple object tracking[J]. International Journal of Computer Vision, 2021, 129(11): 3069-3087.
[44] PANG J M, QIU L L, LI X, et al. Quasi-dense similarity learning for multiple object tracking[C]//Proceedings of

the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021: 164-173.

[45] BALLAS N, YAO L, PAL C, et al. Delving deeper into convolutional networks for learning video representations [EB/OL]. [2021-12-27]. https://arxiv.org/pdf/1511.06432.pdf.

[46] TOKMAKOV P, LI J, BURGARD W, et al. Learning to track with object permanence[EB/OL]. [2021-12-26]. https://arxiv.org/pdf/2103.14258v1.pdf.

[47] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Advances in Neural Information Processing Systems. Long Beach: NIPS, 2017: 5999-6009.

[48] SUN P Z, CAO J K, JIANG Y, et al. Transtrack: Multiple-object tracking with transformer[EB/OL]. [2021-12-26]. https://arxiv.org/pdf/2012.15460v1.pdf.

[49] CARION N, MASSA F, SYNNAEVE G, et al. End-to-end object detection with transformers[C]//European Conference on Computer Vision. Glasgow: Springer, 2020: 213-229.

[50] ZHENG L Y, TANG M, CHEN Y Y, et al. Improving multiple object tracking with single object tracking[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021: 2453-2462.

[51] ZHENG L Y, TANG M, CHEN Y Y, et al. Learning feature embeddings for discriminant model based tracking[C]//European Conference on Computer Vision. Glasgow: Springer, 2020: 759-775.

[52] SHUAI B, BERNESHAWI A, LI X Y, et al. SiamMOT: Siamese multi-object tracking[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021: 12372-12382.

[53] WANG Q, ZHENG Y, PAN P, et al. Multiple object tracking with correlation learning[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021: 3876-3886.

[54] SHAN C B, WEI C B, DENG B, et al. Tracklets predicting based adaptive graph tracking[EB/OL]. [2021-12-26]. https://arxiv.org/pdf/2010.09015.pdf.

[55] BOUGUET J Y. Pyramidal implementation of the affine lucas kanade feature tracker description of the algorithm [EB/OL]. [2021-12-26]. http://robots.stanford.edu/cs223b04/algo_tracking.pdf.

[56] WU J L, CAO J L, SONG L C, et al. Track to detect and segment: An online multi-object tracker[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021: 12352-12361.

[57] LIANG C, ZHANG Z P, ZHOU X, et al. One more check: Making "fake background" be tracked again[EB/OL]. [2021-12-26]. https://arxiv.org/pdf/2104.09 441v1.pdf.

[58] BERNARDIN K, STIEFELHAGEN R. Evaluating multiple object tracking performance: The clear mot metrics[J]. EURASIP Journal on Image and Video Processing, 2008, 10: 246309.

[59] RISTANI E, SOLERA F, ZOU R, et al. Performance measures and a data set for multi-target, multi-camera tracking[C]//European Conference on Computer Vision. Amsterdam: Springer, 2016: 17-35.

编 辑 刘飞阳