多视角多目标跟踪

摘要

多视图聚合是解决多目标检测和跟踪中的遮挡和漏检的有效方法。最近的多视图检测和 3D 对象检测方法通过将所有视图投影到地平面并在鸟瞰图 (BEV) 中执行检测，实现了巨大的性能飞跃。在本次实验中，研究 BEV 中的跟踪是否也能为多目标多摄像头 (MTMC) 跟踪带来有效的性能飞越。我们采用和 Earlybird [7] 类似的框架，学习重新识别（re-ID）特征，并用 M-MVOT [10] 检测器获得 BEV 下的检测结果。最后的结果的对比实验表明，BEV 下的跟踪是有效的，更精确的检测可以显著提升跟踪的结果。

关键词：多视角聚合；多视角检测；多视角多目标跟踪

1 引言

行人检测与跟踪是一个重要的问题，广泛应用于视频监控、自动驾驶车辆和体育分析等领域。尽管单目多目标跟踪（MOT）取得了一定进展，遮挡仍然是该领域面临的最大挑战之一。遮挡导致检测丢失，跟踪被打断，从而限制了检测和跟踪的质量。然而，像体育分析这样的实际场景往往需要在高度杂乱或拥挤的环境中进行检测。在这些情况下，可能有多个视角重叠的摄像头可用。通过从多个视角观察场景，可以克服这些遮挡问题，因为在一个摄像头中隐藏的物体可能在另一个摄像头中可见。挑战在于如何聚合来自多个摄像头视角的信息。早期的方法通过后期融合技术来解决多视角检测的问题：首先，在一个视角中检测到行人，然后将该检测投影到 3D 空间或通常是地面平面上，并与其他视角的投影进行关联。较新的方法采用了早期融合策略，首先将所有视角的表示投影到公共地面平面或鸟瞰图中，然后进行检测。这些早期融合的检测方法比之前的后期融合方法显著提高了检测质量。后期融合方法的优势在于，它们所需的硬件较少，因为处理可以独立进行，并且投影到 3D 的图像比完整图像更稀疏。早期融合方法的优势在于它们可以端到端训练，而后期融合通常是分别优化检测和多视角关联。鸟瞰图（BEV）空间中检测的一个挑战是透视变换带来的失真。一些方法尝试解决这个问题。在本次实验中，我们属于 TBD 范式，也就是检测后跟踪。在检测部分使用 M-MVOT [10] 检测器，该检测器提出了一种新颖的基于马氏距离（Mahalanobis Distance）的多视角最优传输（M-MVOT）损失函数，专为多视角人群定位设计，M-MVOT 通过结合马氏距离、距离调节机制和多视角信息，提供了一种更精确、更鲁棒的多视角人群定位解决方案，实现了 SOTA 的性能。在跟踪部分，我们采用了 Earlybird [7] 中引入的思想，为鸟瞰图空间中的每个检测学习一个重新识别（re-ID）特征。基于外观的 re-ID 特征和卡尔曼滤波器作为运动模型进行关联，在 Wildtrack 数据集上 MODA 指标比 SOTA 模型高了 1.4，在 MultiviewX 数据集上高上 2.4。

本次实验贡献如下：1）我们将 M-MVOT 检测器和 Earlybird 结合，达到多视角多目标最好的性能，证明检测的结果很大程度影响了跟踪的结果；2）在 re-ID 部分做了消融实验，证明多视角融合的 re-ID 特征是有效的。

2 相关工作

2.1 基于检测的跟踪

如图 1所示, 许多检测与跟踪（TBD）方法使用独立的数据关联算法，可以与任何物体检测器结合使用。例如，SORT [1] 使用卡尔曼滤波器作为运动模型，并使用匈牙利匹配算法来关联物体。DeepSORT [8] 基于 SORT，在此基础上增加了外观信息提取，从而显著提高了 SORT 的性能。最近的工作，FairMOT [12]，将 Re-ID 特征提取和物体检测集成到一个网络中，然后将其输入到关联模块中，以生成跟踪结果。EarlyBird [7] 中的方法与 DeepSORT 相似，将多目标跟踪（MOT）扩展到多视角场景。它从多个视角聚合特征，生成鸟瞰图（BEV）特征，然后在 BEV 视角中执行多目标跟踪。

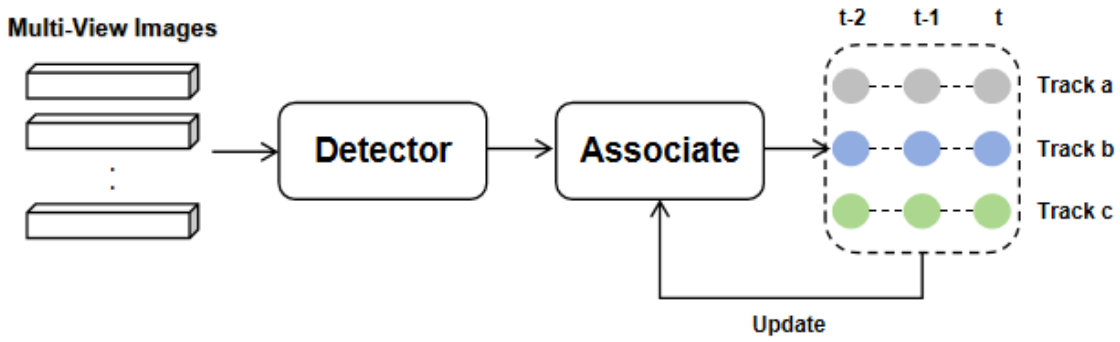

图 1. 基于检测的跟踪

2.2 基于查询的跟踪

如图 2所示, 最近的多目标跟踪（MOT）方法尝试通过 Transformer 查询来建模跟踪过程。TrackFormer [6] 和 MOTR [9] 将跟踪对象表示为一个查询，并通过自注意力机制和二分匹配实现隐式关联，用于目标分配。MUTR3D [11] 基于 DETR3D 检测器，将 MOTR 风格的注意力跟踪范式扩展到多摄像头 3D 多目标跟踪，并取得了令人鼓舞的成果。

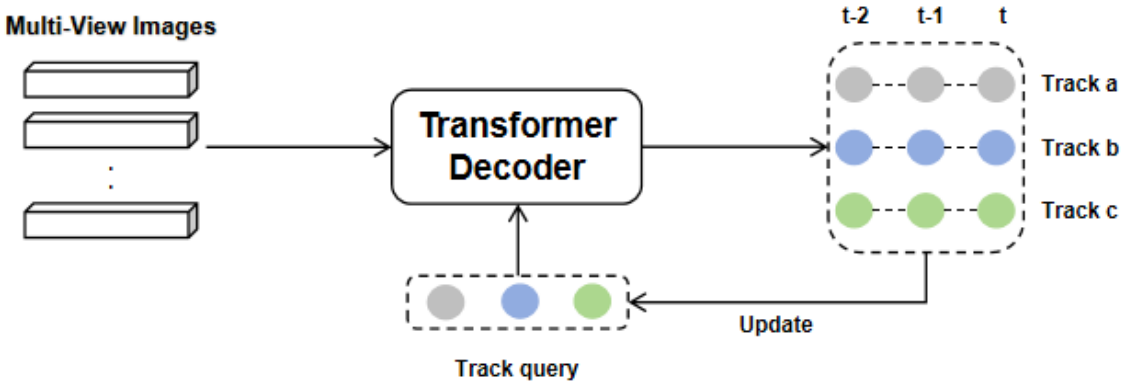

图 2. 基于查询的跟踪

2.3 基于解耦查询的跟踪

如图 3所示, 基于注意力的跟踪方法将检测和跟踪查询纠缠在一个嵌入中，导致性能不是最优。DQTrack [5] 使用解耦的跟踪和检测查询来检测物体并执行后续关联。3DMOTFormer [3] 与 DQTrack 具有相似的结构，但其后续关联模块基于图神经网络（GNN），在不同检测器之间展示了良好的泛化能力。ADA-Track [4] 改进了 3DMOTFormer，通过解耦检测和关联任务，同时利用这些任务之间的协同作用，实现了更高质量的跟踪结果。

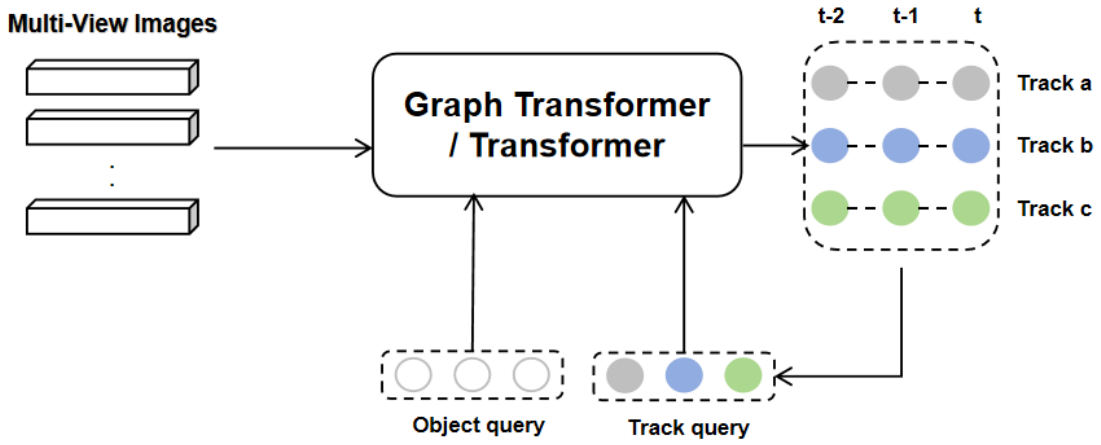

图 3. 基于解耦查询的跟踪

2.4 基于图神经网络（GNN）的跟踪

如图 4所示,ReST [2] 提出了一个新颖的可重构图模型，首先在空间上关联所有摄像头检测到的物体，然后将其重构为时间图，以沿时间维度进行时间关联。

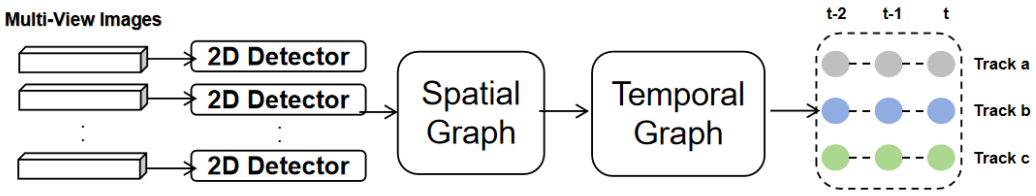

图 4. 基于图神经网络（GNN）的跟踪

3 本文方法

3.1 本文方法概述

在本次实验中，我们属于 TBD 范式，也就是检测后跟踪，模型架构如下图 5所示。在检测部分使用 M-MVOT 检测器，该检测器提出了一种新颖的基于马氏距离（Mahalanobis Distance）的多视角最优传输（M-MVOT）损失函数，专为多视角人群定位设计，M-MVOT 通过结合马氏距离、距离调节机制和多视角信息，提供了一种更精确、更鲁棒的多视角人群定位解决方案，实现了 SOTA 的性能。在跟踪部分，我们采用了中引入的思想，为鸟瞰图空间中的每个检测学习一个重新识别（re-ID）特征。基于外观的 re-ID 特征和卡尔曼滤波器作为运动模型进行关联, 得到最后的检测结果。

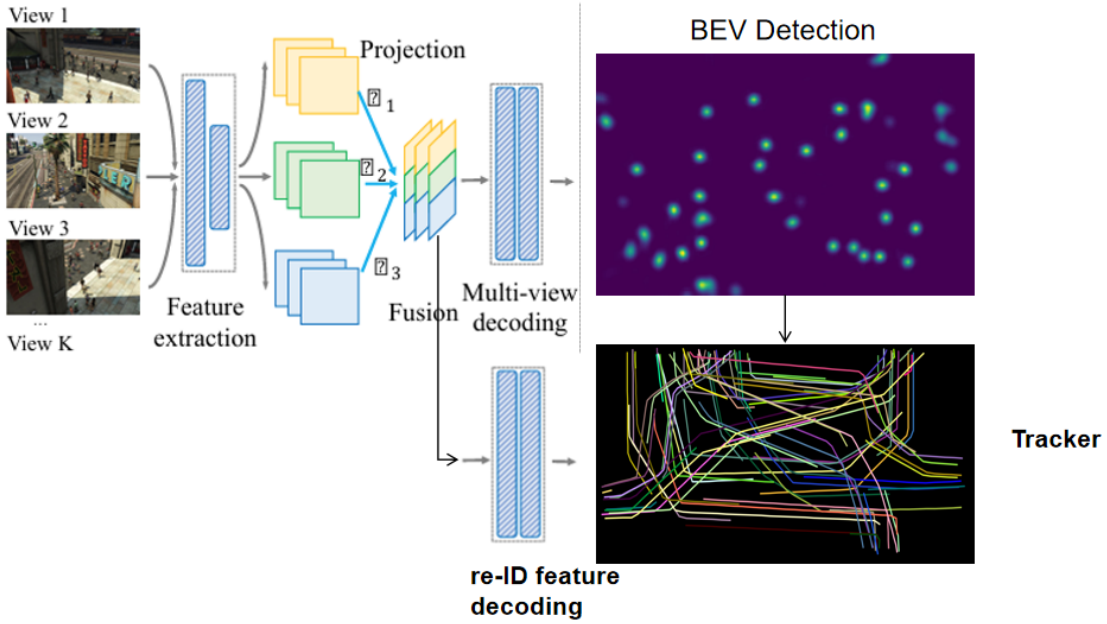

图 5. 多视角多目标跟踪架构

3.2 损失函数定义

re-ID 头的目的是生成可以区分个体行人的特征。理想情况下，不同的行人之间的相似度应该小于同一个行人之间的相似度。为了实现这一目标，我们通过分类任务和度量学习任务来学习 re-ID 特征。首先，我们在地面平面上应用一个头部，生成 re-ID 特征，其大小为

$C_{id,g} \times H_g \times W_g$，其中 $C_{id} = 64$，同时对图像特征应用一个头部，大小为 $C_{id,f} \times H_f \times W_f$。之后，我们在两个平面中提取中心检测位置的特征。我们使用一个线性层创建一个类身份分布，并通过交叉熵损失训练其与真实标签的匹配。正如之前所讨论的，透视变换会在地面平面上引入强烈的失真。因此，我们从图像视角来监督 re-ID 特征。除了交叉熵损失外，我们还应用了 SupConf 损失，该损失将同一类身份的特征拉近，同时将不同类别的特征推开。

4 复现细节

4.1 与已有开源代码对比

将两个模型进行结合后训练，具体方法如下：

1. 用 ECCV 24 的基于最优传输多视角人群检测模型替换掉 Earlybird 的 BEV detection head。

2. 冻结已经训练好的检测板块，只对 re-id 部分进行训练。

在虚拟数据集 MultiviewX 和真实数据集 Wildtrack 上进行训练，训练轮数为 50 次。

5 实验结果分析

5.1 数据集和评测指标

Wildtrack 数据集 Wildtrack 是一个真实世界的数据集，使用七台同步和标定的摄像机拍摄，覆盖一个 12 米 ×36 米的区域，摄像机的视野有重叠。行人的移动发生在公共环境中，且没有预设脚本。标注数据位于地面平面上，量化成一个 480 × 1440 的网格，每个网格单元的大小为 2.5 厘米 × 2.5 厘米。每帧的平均行人数为 20 人，每个位置覆盖 3.74 台摄像头。每个摄像头的图像分辨率为 1080 × 1920 像素，帧率为 2 帧每秒，总记录时间为 35 分钟。

MultiviewX 数据集 MultiviewX 是一个在游戏引擎中生成的合成数据集，旨在成为 Wildtrack 数据集的合成复制版本。MultiviewX 包含由 6 台虚拟摄像头生成的视角，具有重叠的视野。拍摄的区域为 16 米 ×25 米，稍小于 Wildtrack 数据集的区域。标注数据时，地面平面被量化为一个 640 × 1000 的网格，每个网格表示相同的 2.5 厘米 × 2.5 厘米的方格。每帧的平均行人数为 40 人，每个位置覆盖 4.41 台摄像头。摄像头的分辨率（1080×1920）、帧率（2 帧每秒）和长度（400 帧）与 Wildtrack 相同。

检测指标与单视角检测系统不同，后者评估预测的边界框，多视角检测系统评估的是投影到地面平面的占用图。因此，和地面真值的比较不再使用交并比（IoU），而是采用欧氏距离。如果检测结果在距离 $r = 0.5\,\mathrm{m}$ 范围内，则被视为真正检测，这个距离大致对应于人体的半径。检测任务一般使用多目标检测准确度（MODA）作为主要性能指标，因为它考虑了归一化的漏检和假阳性。此外还报告多目标检测精度（MODP）。

跟踪指标对于跟踪，指标也是在地面平面上计算的。我们报告了常见的 MOT 指标和基于身份的指标，其中正向分配的阈值设置为 $r = 1\,\mathrm{m}$ 来归一化多目标跟踪精度（MOTP）。主要考虑的指标是多目标跟踪准确度（MOTA）。MOTA 考虑了漏检、假检测和身份切换。

如下表1，2，我们在 wildtrack 和 multiviewx 数据集做了训练和测试，同时还将最近先进的多视角检测模型，Mvdet，Mv-detr，3drom，shot 加到跟踪模型上对比他们的结果。

Method	MODA	MODP	MOTA	MOTP
Earlybird	94.2	90.1	88.4	86.2
M-MVOT	96.7	86.1	90.8	84.1
MVDetr	94.1	91.5	85.74	85.5
MV-Det	82.3	77.7	69.00	81.2
3drom	94.6	83.5	86.4	83.1
Shot	89.2	82.3	80.65	83.1

表 1. 不同方法的检测与跟踪指标在 MultiviewX 数据集上比较

Method	MODA	MODP	MOTA	MOTP
Earlybird	91.2	81.8	89.5	86.6
M-MVOT	91.9	81.3	90.9	88.2
MVDetr	92.2	82.3	89.9	88.5
MV-Det	87.4	74.0	81.5	83.3
3drom	92.7	76.5	89.0	83.9
Shot	91.9	76.4	86.0	85.6

表 2. 不同方法的检测与跟踪指标在 Wildtrack 数据集上比较

5.2 消融实验

如下表3，我们还做了针对不同的 re-ID 特征提取方法做了消融实验，分别在图像特征，BEV 特征，融合特征，单个显示最好图像特征进行了对比，发现这些不同的 reid 特征提取方法对跟踪的结果没有影响。对于图像特征，我们采用的方法是，将 BEV 下检测的结果反投影到图像上，裁剪出该区域送入 reid 特征提取模型 OSnet 中，得到 reid 特征再送入跟踪模型中。对于融合特征，我们简单的取图像特征与 BEV 特征的平均值。

图像特征	BEV 特征	图像，BEV 融合特征	单个显示最好图像特征
MOTA	90.44	90.44	90.44

表 3. 不同 reid 特征提取方法的 MOTA 比较

6 总结与展望

本次实验我们将 M-MVOT 检测器和 Earlybird 结合，在 Wildtrack 数据集上 MODA 指标比 SOTA 模型高了 1.4，在 MultiviewX 数据集上高上 2.4, 证明检测的结果很大程度影响了跟踪的结果；同时在 re-ID 部分做了消融实验，证明多视角融合的 re-ID 特征是有效的。

在未来的工作中，我们会考虑用上不同的跟踪模型到不同的数据集上，希望在更大的更长时间的视频上达到精准的实时跟踪效果。

参考文献

[1] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple online and realtime tracking. In *2016 IEEE international conference on image processing (ICIP)*, pages 3464–3468. IEEE, 2016.

[2] Cheng-Che Cheng, Min-Xuan Qiu, Chen-Kuo Chiang, and Shang-Hong Lai. Rest: A reconfigurable spatial-temporal graph model for multi-camera multi-object tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10051–10060, 2023.

[3] Shuxiao Ding, Eike Rehder, Lukas Schneider, Marius Cordts, and Juergen Gall. 3dmot-former: Graph transformer for online 3d multi-object tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9784–9794, 2023.

[4] Shuxiao Ding, Lukas Schneider, Marius Cordts, and Juergen Gall. Ada-track: End-to-end multi-camera 3d multi-object tracking with alternating detection and association. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15184–15194, 2024.

[5] Yanwei Li, Zhiding Yu, Jonah Philion, Anima Anandkumar, Sanja Fidler, Jiaya Jia, and Jose Alvarez. End-to-end 3d tracking with decoupled queries. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18302–18311, 2023.

[6] Tim Meinhardt, Alexander Kirillov, Laura Leal-Taixe, and Christoph Feichtenhofer. Track-former: Multi-object tracking with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8844–8854, 2022.

[7] Torben Teepe, Philipp Wolters, Johannes Gilg, Fabian Herzog, and Gerhard Rigoll. Early-bird: Early-fusion for multi-view tracking in the bird's eye view. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 102–111, 2024.

[8] 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)*, pages 3645–3649. IEEE, 2017.

[9] Fangao Zeng, Bin Dong, Yuang Zhang, Tiancai Wang, Xiangyu Zhang, and Yichen Wei. Motr: End-to-end multiple-object tracking with transformer. In *European Conference on Computer Vision*, pages 659–675. Springer, 2022.

[10] Qi Zhang, Kaiyi Zhang, Antoni B Chan, and Hui Huang. Mahalanobis distance-based multi-view optimal transport for multi-view crowd localization. In *European Conference on Computer Vision*, pages 19–36. Springer, 2025.

[11] Tianyuan Zhang, Xuanyao Chen, Yue Wang, Yilun Wang, and Hang Zhao. Mutr3d: A multi-camera tracking framework via 3d-to-2d queries. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4537–4546, 2022.

[12] Yifu Zhang, Chunyu Wang, Xinggang Wang, Wenjun Zeng, and Wenyu Liu. Fairmot: On the fairness of detection and re-identification in multiple object tracking. *International journal of computer vision*, 129:3069–3087, 2021.