arXiv:2302.03802v2 [cs.CV] 3 Apr 2023

# 立足于过去与未来之间：面向多摄像头三维多目标跟踪的时空建模

Ziqi Pang[1]*, Jie Li[2], Pavel Tokmakov[2],Dian Chen[2],Sergey Zagoruyko[3],Yu-Xiong Wang[1]†伊利诺伊大学厄巴纳-香槟分校[1],丰田研究院[2],Woven Planet Level-5[3]

## 摘要

本文提出了一种端到端的多摄像头三维多目标跟踪（*MOT*）框架。该框架强调时空连续性，并将过去推理与未来推理相结合，以处理被跟踪物体。因此，我们将其命名为“用于跟踪的过去与未来推理”（*PF-Track*）。具体而言，我们的方法采用“基于注意力的跟踪”框架，并利用物体查询在时间维度上一致地表示被跟踪实例。为显式利用历史线索，我们的“过去推理”模块通过与先前帧中的查询及其他物体进行交叉注意，学习细化轨迹并增强目标特征。“未来推理”模块整合历史信息并预测稳健的未来轨迹。在长期遮挡的情况下，我们的方法通过整合运动预测来维持目标位置并实现重新关联。在 *nuScenes* 数据集上，与以往方法相比，我们的方法大幅提升了 *AMOTA*，并将身份切换减少了 *90%*，降低至原来的一个数量级，这是显著的改进。代码和模型已公开于 *https://github.com/TRI-ML/PF-Track*。

## 1. 引言

对三维物体轨迹进行推理是自主导航的基石。尽管已有许多基于激光雷达的方法 [43, 66, 71], 但传感器的成本和可靠性限制了它们的适用性。因此，仅依靠摄像机检测、跟踪和预测物体轨迹是一项至关重要的任务。这些任务分别取得了显著进展，但过去主要是孤立研究，并以临时拼凑的方式整合为全栈流水线。

具体而言，三维检测受到了广泛关注 [22,28,29,33,61], , 但这些检测结果随时间的关联，大多独立于定位过程进行。

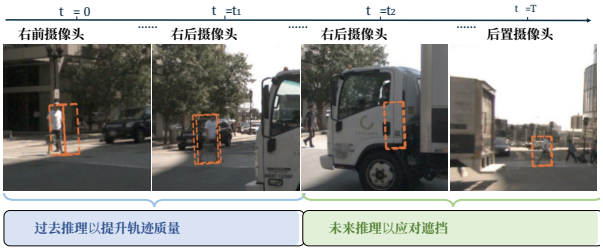


图 1. 我们将预测的三维边界框投影到图像上，以可视化模型的输出。起初，由于深度歧义，基于图像的检测可能不准确（$t = 0$）。借助“过去推理”，通过利用历史信息，边界框质量（$t = t_1$）逐步提升。借助“未来推理”，我们的 PF-Track 能预测目标的长期运动，并在遮挡（$t = t_2$）和摄像头切换情况下保持其状态。这使得无需显式重新识别（$t = T$）即可重新关联，因为物体标识不会切换。PF-Track 还在统一框架中结合过去与未来推理，以提升时空一致性。

过程 [21, 38, 50]。最近，已有少数端到端检测与跟踪方法被提出，但它们只处理相邻帧，未能整合长期时空线索 [7, 14, 40, 73]。另一方面，在预测领域，通常假设可以获得真实物体轨迹和高清地图 [4, 8, 13, 67]。也有少数尝试旨在进行更贴近现实的评估 [18, 23], , 但仅关注预测性能。

在本文中，我们认为，通过联合优化检测—跟踪—预测流程，尤其是在基于摄像头的系统中，可以显著提升多目标跟踪性能。图1展示了我们真实场景实验中的一个直观示例。起初，行人完全可见，但仅利用单帧信息的模型却产生了较大偏差的预测（图 $t = 0$ 中的第1帧）。随后，融入过去的时间信息，便可随着时间推移逐步纠正误差（图$t = t_1$ 中的第1帧），这是通过利用时空连续性实现的。此外，当行人完全被遮挡时（图 $t = t_2$ 中的第1帧），我们仍然可以利用聚合的过去信息来预测其位置

*在丰田研究院实习期间完成的工作。

†通讯作者：Ziqi Pang（ziqip2@illinois.edu）和 Yu-Xiong Wang（ytxw@illinois.edu）。

，从而估计未来轨迹。最后，通过长期预测，即使行人在另一台摄像机中重新出现，我们也能成功跟踪该行人，实现正确的重新关联（图 $t = T$ 中的第1帧）。上述稳健的时空推理得益于对过去和未来信息进行无缝的双向整合，这与当前主流的基于视觉的多摄像头三维多目标跟踪（三维多目标跟踪）流程形成了鲜明对比。

为此，我们提出一种端到端框架，用于同时完成三维多目标跟踪任务中的3D物体检测、跟踪和轨迹预测，如图所示。2，并采用“基于注意力的跟踪”[41,72,73]范式。与采用相同范式的最接近基线相比[73],，我们的不同之处在于显式地进行过去与未来推理：三维目标查询始终表示随时间变化的目标，在各帧间传递目标的时空信息，并生成相应的边界框和未来轨迹。为利用时空线索，我们的算法通过简单的注意力操作捕捉目标动态和交互，并将其用于轨迹细化以及稳健的长期轨迹预测。最后，我们将预测轨迹反馈至跟踪模块，以替代缺失的检测结果，从而形成闭环（例如，因遮挡而缺失的检测结果）。为凸显过去与未来联合推理的能力，我们将该方法命名为“用于跟踪的过去与未来推理”（PF-Track）。

我们在 nuScenes [4] 数据集上对 PF-Track 进行了全面评估，并证明联合建模过去与未来信息能够显著提升目标跟踪效果。具体而言，与以往的多摄像头三维多目标跟踪方法相比，PF-Track 将身份切换减少了超过 **90%** 。

总而言之，我们的贡献如下。1. 我们提出了一种端到端、仅依赖视觉的三维多目标跟踪框架，利用目标级时空推理处理过去和未来信息。2. 我们的框架通过交叉关注来自**“过去”**的特征，提升了轨迹质量。3. 我们提出了一个联合跟踪与预测流程，其组成部分包括“未来推理”，并证明跟踪能够明确受益于对**“未来”**的长期预测。4. 我们的方法在大规模 nuScenes 数据集 [4] 上取得了新的最先进水平，并显著提升了 AMOTA 和身份切换指标。

## 2. 相关工作

**基于激光雷达的三维多目标跟踪**。大多数先前的三维多目标跟踪工作都利用激光雷达模态。得益于近期基于激光雷达的三维检测取得的进展， [26, 69], 尤其是可靠的距离信息，大多数最先进的三维多目标跟踪算法都采用“基于检测的跟踪”范式。 [64]针对单帧检测输出，已有多种方法被提出，以改进数据关联、 [43, 66, 71], 运动传播 [10, 76], 和生命周期管理。 [43, 59]然而，这些工作大多假设检测输出具有较高的定位精度。因此，数据关联通常基于位置进行，并可选择结合抽象的目标属性（例如，三维交并比（3D IoU）、 [64], 3三维广义交并比（GIoU）[43], 以及 L2距离 [69]）。当这些系统迁移到摄像机模态时，这种偏向会使其变得脆弱，因为三维检测的定位不确定性更高。尽管最新方法引入了基于学习的算法，利用高保真特征改进关联，例如点云中的低级特征 [53]或摄像机的中间特征， [10], 但这些方法仍建立在基于激光雷达的框架之上，并且同样依赖定位质量。

**基于摄像机的二维多目标跟踪**。基于摄像机的二维多目标跟踪是计算机视觉中的经典任务。受“检测式跟踪”范式 [3], 2主导，三维多目标跟踪在利用高保真特征 [45,56, 68, 74]方面取得了更多成功。较早的工作，如 DeepSORT [68] ，利用深度网络中的中间特征来衡量外观相似度。FairMOT [74] 采用了额外的重识别分支，在检测网络中学习判别性特征。TransMOT [45] 提出利用图网络融合时空特征。

**基于相机的三维多目标跟踪**。 近年来，随着单目深度估计 [15, 17, 19] 和基于图像的三维目标检测 [22, 28, 29, 33, 34, 44, 48, 60, 61]技术的发展，基于相机的三维多目标跟踪在自动驾驶应用中日益受到关注。早期方法改造二维多目标跟踪算法，并利用单目深度提升二维跟踪结果 [57, 76]。较新的方法则在数据关联中引入额外的三维信息 [21, 38, 50]。[38] 提出利用三维重建， [21] 则通过三维属性（例如 深度和朝向）增强二维重识别特征。CC-3DT [14]融合多视图摄像机中同一目标的特征，以改进跨时间跨视角关联。然而，针对基于相机的三维检测中的高不确定性和偏差进行考量或校正，相关研究仍较少。在本研究中，我们利用长期目标推理，尤其是过去推理，来提升三维边界框的质量。

**基于注意力的跟踪**。 多目标跟踪领域的一种新兴趋势是“基于注意力的跟踪”范式 [41, 54, 72, 73], , 该范式受到了新颖的、基于 Transformer 的检测架构 DETR [5] 的启发。MOTR [72] 和 Trackformer [41] 通过在不同帧之间传递查询，扩展了 DETR [5] 中的基于查询的检测框架。在这一范式中，

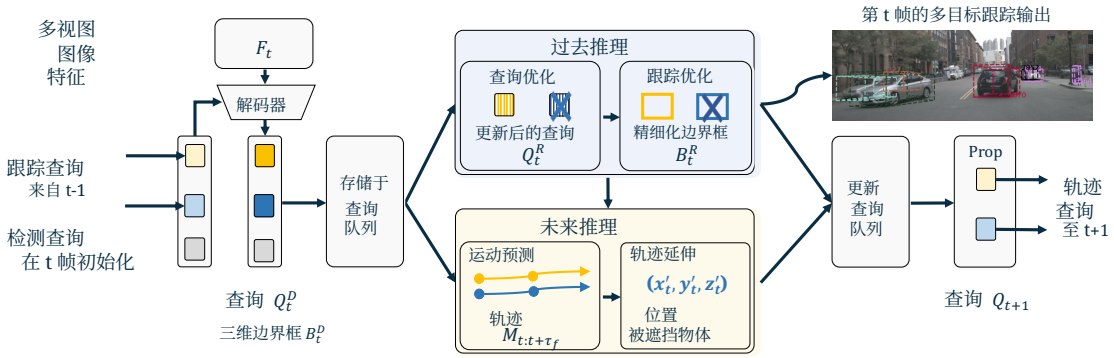


图 2. **PF-Track 框架**。 PF-Track 将物体表示为查询，对图像特征进行解码，并预测边界框。为提升时空一致性，我们引入了新颖的“过去推理”和“未来推理”模块。(1)“过去推理”利用查询队列中的历史信息，优化查询特征和轨迹的边界框。(2)“未来推理”通过估计长期未来轨迹，改善查询在帧间的传播。此外，如果目标因置信度低或受到遮挡而丢失（蓝色方块 ×），则“轨迹延伸”模块可以利用长期轨迹维持其位置。最后，PF-Track 将过去与未来推理结合起来，用于三维多目标跟踪。（建议彩色查看，详情见第 3.1 节。）

数据关联被替换为在当前帧中使用一组跟踪查询进行“检测”。MUTR3D [73] 提出了首个将这一范式应用于三维多目标跟踪领域的框架。它使用三维轨迹查询，在多个时间戳和多视角之间联合建模目标特征。尽管 MUTR3D 在当时取得了改进，但它总体上沿用了二维多目标跟踪方法的设计，未针对提升轨迹的定位质量或更好地将查询传递至未来帧进行特别处理。我们提出的算法同样采用“基于注意力的跟踪”范式，但扩展了现有方法的时间范围。具体而言，我们证明，过去与未来联合推理能够提供强大的时空目标表征，从而改进跟踪框架。

**运动预测**。 预测智能体轨迹对自动驾驶至关重要 [16, 24, 35, 42, 49, 52, 70]。最常见的做法是根据人工标注或自动标注得到的干净轨迹进行预测 [8, 13, 55, 67]。许多研究关注从感知结果进行端到端预测 [1, 6, 20, 30, 39, 46, 47, 51, 62, 63, 65, 75],，尤其是如何直接基于感知结果改进运动预测。然而，我们的目标有所不同：运动预测模型能否使三维多目标跟踪受益？在二维场景中，近期只有少量研究关注这一问题 [12]。我们的算法将这项研究拓展到更具挑战性的多摄像头三维场景，且无需显式重新识别。

## 3. 方法：PF-Track

本节介绍我们新颖的 3D 多目标跟踪框架，如图 2所示。该框架以端到端方式对物体轨迹的显式过去与未来建模为核心。我们首先在第3.1节概述整个流程，随后说明如何高效利用“过去”（第3.2节）和“未来”（第3.3节）的信息。

最后，我们总结第3.4节中介绍的框架所用损失函数。

### 3.1. PF-Track 流程

我们提出的 PF-Track 迭代使用一组物体查询 [41, 72, 73] 来处理多视图、多目标三维跟踪。在每个时间戳 $t$，给定 $K$ 来自周围摄像机的图像$\mathbf{I}_t^k$，三维多目标跟踪的目标是在各帧中生成具有一致 ID 的物体检测结果，记作$\mathbf{B}_t = \{\mathbf{b}_t^i\}$，其中 $i$ 是物体标识。

**3D物体查询**。我们框架的入口是接收$\mathbf{Q}_t = \{\mathbf{q}_t^i\}$ 从前一帧传播而来的物体查询 $t-1$（图2中的黄色和蓝色方块），它们表示被跟踪的物体：

$$\mathbf{Q}_t \leftarrow \mathbf{Prop}(\mathbf{Q}_{t-1}). \tag{1}$$

这种基于查询的设计能够自然地处理跟踪任务，因为查询会随时间携带物体的身份。除了来自前一帧、代表被跟踪实例的查询外，我们还添加固定数量的检测查询（图2中的灰色方块），用于发现新物体。实际应用中，我们使用 500 个检测查询，并将其初始化为可学习嵌入。

每个查询 $\mathbf{q}_t^i \in \mathbf{Q}_t$ 都代表一个独特的 3D物体，包含一个特征向量 $\mathbf{f}_t^i$ 和一个 3D位置 $\mathbf{c}_t^i$：$\mathbf{q}_t^i = \{\mathbf{f}_t^i, \mathbf{c}_t^i\}$。这里我们强调，查询位置会主动参与下方物体边界框的解码。

**解码器**。为预测三维边界框，并利用最新图像输入更新查询，PF-Track采用了基于注意力的检测架构 [5, 77] 来解码图像特征$\mathbf{F}_t$ 和物体查询：

$$\mathbf{B}_t^D, \mathbf{Q}_t^D \leftarrow \mathbf{Decoder}(\mathbf{F}_t, \mathbf{Q}_t), \tag{2}$$

其中$\mathbf{B}_t^D$和$\mathbf{Q}_t^D$分别表示检测到的三维边界框和更新后的查询特征。在解码过程中，解码器将查询的三维位置$\mathbf{c}_t^i$提升为位置嵌入，以聚焦于与物体空间位置相关的图像区域。尽管 PF-Track 的设计不依赖于基于查询的检测算法，但在实验中我们主要采用当前最先进水平的三维检测器 PETR [33], 。

**用于细化与传播的过去与未来推理**。从单帧图像特征中解码出查询和边界框后，PF-Track 依次进行过去与未来推理，以（1）细化当前检测结果$\mathbf{B}_t^D$，得到$\mathbf{B}_t^R$，并将查询$\mathbf{Q}_t^D$细化为$\mathbf{Q}_t^R$。（$R$ 是“细化”的缩写）；（2）根据预测运动将查询传播至下一个时间戳。

“过去推理”**PR**（·）是一个组件，它聚合来自先前帧的信息，生成细化后的查询$\mathbf{Q}_t^R$和细化后的边界框$\mathbf{B}_t^R$:

$$\mathbf{Q}_t^R, \mathbf{B}_t^R \leftarrow \mathbf{PR}(\mathbf{Q}_t^D, \mathbf{B}_t^D, \mathbf{Q}_{t-\tau_h:t-1}, ). \tag{3}$$

实际应用中，历史查询$\mathbf{Q}_{t-\tau_h:t-1}$ 来自一个**查询队列**，该队列维护来自过去 $\tau_h$ 帧的查询（$h$是“历史”的缩写）。

经过过去推理后，“未来推理”模块**FR**（·）从查询传播的角度提升目标位置的连贯性。该模块通过预测未来至多$\tau_f$ 帧的运动（$f$ 是“未来”的缩写），并相应地变换查询的位置来实现这一点:

$$\mathbf{Q}_{t+1}, \mathbf{M}_{t:t+\tau_f} \leftarrow \mathbf{FR}(\mathbf{Q}_t^R, \mathbf{Q}_{t-\tau_h:t-1}). \tag{4}$$

具体而言，未来推理从历史查询特征中提取目标运动$\mathbf{M}^{t:t+\tau_f}$，以预测轨迹$\mathbf{M}^{t:t+1}$ 。利用单步运动$\mathbf{M}_t^R$将当前查询$\mathbf{Q}^{t+1:t+\tau_f}$ 传播至下一个时间戳，并利用长期轨迹$\mathbf{M}^{t+1:t+\tau_f}$ 处理遮挡。图 2 中的“轨迹延伸”指的是通过预测轨迹进行遮挡推理。

PF-Track 迭代执行上述过程。细化后的三维边界框$\mathbf{B}_t^R$是三维多目标跟踪的输出。

### 3.2. 过去推理

为解决仅依赖视觉的三维定位中检测的不确定性，过去推理主要关注两个方面：(1) 通过关注历史嵌入来增强查询特征；(2) 利用改进后的查询特征调整边界框，从而细化轨迹。

**查询优化：从 $\mathbf{Q}_t^D$到 $\mathbf{Q}_t^R$**。如图所示，我们首先沿时间轴和实例轴执行注意力操作，利用历史信息显式更新查询特征

见图 3。“跨帧”注意力促进了特征之间的交互

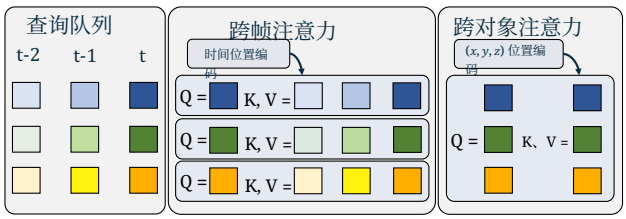


图3. **查询优化**。“跨帧”和“跨目标”注意力模块分别处理查询队列，以捕获时间关系和对象间关系。它们分别对时间 $t$ 和空间位置 $(x, y, z)$ 应用位置编码。（彩色观看效果最佳。）

每个目标在包含 $\tau_h$ 帧的历史窗口内:

$$\begin{aligned}\mathbf{f}_t^i \leftarrow \mathbf{CrossFrameAttn}(&\mathrm{Q} = \mathbf{f}_t^i,\\ &\mathrm{K} = \mathbf{f}_{t-\tau_h:t}^i, \mathrm{V} = \mathbf{f}_{t-\tau_h:t}^i,\\ &\mathrm{PE} = \mathbf{Pos}(t-\tau_h:t)),\end{aligned} \tag{5}$$

其中，$\mathbf{Pos}(t-\tau_h:\ t)$ 将时间戳转换为位置嵌入，而特征为空的历史帧在注意力计算中会被忽略。

接下来，过去推理应用“跨目标”注意力来整合上下文信息，并为每个目标生成更具区分性的特征表示。具体而言，跨目标注意力（图 3，右）通过以下方式进一步更新查询特征

$$\begin{aligned}\mathbf{f}_t^{1:N_t} \leftarrow \mathbf{CrossObjectAttn}(&\mathrm{Q,K,V} = \mathbf{f}_t^{1:N_t},\\ &\mathrm{PE} = \mathbf{Pos}(\mathbf{c}^{1:N_t})),\end{aligned} \tag{6}$$

其中，跨对象注意力在三维位置嵌入$\mathbf{Pos}(\mathbf{c}^{1:N_t})$的引导下交换$N_t$ 个目标的特征。最终输出$\mathbf{f}_t^{1:N_t}$成为查询$\mathbf{Q}_t^R$中的精炼特征向量。

简而言之，将跨帧注意力与跨对象注意力解耦有两项优势。首先，将帧间注意力（跨帧）与目标间注意力（跨对象）分开，使我们能够针对两者分别设计专门的时间和位置编码。其次，这将全局交叉注意力的计算复杂度从 $\mathcal{O}(N_t^2\tau_h^2)$降低至 $\mathcal{O}(N_t^2 + N_t\tau_b^2)$，显著降低了计算复杂度。我们的设计也与运动预测方法 [16, 42] 建模时空关系的方式密切相关。

**轨迹细化：从 $\mathbf{B}_t^D$到 $\mathbf{B}_t^R$**。利用历史信息细化查询后，过去推理进一步通过轨迹细化来提升三维边界框质量。如公式所示7，我们使用多层感知机 (MLP) 预测目标更新后的属性，包括中心残差 $(\Delta x,\ \Delta y,\ \Delta z)$、尺寸 $(l,\ w,\ h)$、方向 $(\theta)$、速度 $(\mathbf{v})$ 和分数 $(s)$:

$$(\Delta x, \Delta y, \Delta z, l, w, h, \theta, \mathbf{v}, s)^i = \mathbf{MLP}(\mathbf{f}_t^i). \tag{7}$$

随后使用这些属性对原始边界框进行如下调整：

$$\mathbf{b}_t^i = (\Delta x + x_t^i, \Delta y + y_t^i, \Delta z + z_t^i, l, w, h, \theta, \mathbf{v}, s), \quad (8)$$

从而得到$\mathbf{B}_t^R$，这就是帧 $t$的最终模型输出。

### 3.3. 未来推理

“未来推理”着重于改进查询在帧间的传播，以提升时空一致性。它首先学习轨迹预测，用于在相邻帧之间移动查询。随后，未来推理利用预测得到的长期轨迹，维持被遮挡或含噪轨迹的位置。

**运动预测**。 轨迹预测监督模型捕捉物体运动的能力，并进一步有助于在不同时间戳间传播查询位置。与过去推理类似，我们的未来推理模型采用简单的基于注意力的架构。首先，我们通过跨帧注意力为 $\tau_f$ 个时间戳生成运动嵌入$\mathbf{mf}_{t:t+\tau_f}^i$：

$$\begin{aligned}\mathbf{mf}_{t:t+\tau_f}^i \leftarrow \mathbf{CrossFrameAttn}(&\\ \mathrm{Q} &= \mathbf{mf}_{t:t+\tau_f}^i,\\ \mathrm{K}, \mathrm{V} &= \mathbf{f}_{t-\tau_h:t}^i, \mathbf{f}_{t-\tau_h:t}^i,\\ \mathrm{PE} &= \mathbf{Pos}(t-\tau_h : t+\tau_f)),\end{aligned} \quad (9)$$

其中$\mathbf{mf}_{t:t+\tau_f}^i$ 初始化为零，历史特征$\mathbf{f}_{t-\tau_h:t}^i$ 则作为信息来源。随后，通过多层感知机解码每个时间戳的运动：

$$\mathbf{m}_{t:t+\tau_f}^i = \mathbf{MLP}(\mathbf{mf}_{t:T+\tau_f}^i), \quad (10)$$

通过结合这些帧级输出，可以恢复三维空间中的目标轨迹。我们的架构受到 SceneTransformer 的启发 [42], 它同样采用了完全基于注意力的架构。

预测轨迹 $\mathbf{m}_{t:t+\tau_f}^i$ 的保真度优于解码器在 $\mathbf{B}_t^R$ 和 $\mathbf{B}_t^D$ 中预测的速度 $\mathbf{v}$。因此，我们可以通过加上单步轨迹来传播查询的位置：

$$\mathbf{c}_{t+1}^i = \mathbf{c}_t^i + \mathbf{m}_{t:t+1}^i. \quad (11)$$

**轨迹延伸**。 为应对遮挡或观测噪声，我们提出利用预测的轨迹来延伸轨迹。具体而言，我们用运动预测模块的输出替换缺失或置信度较低的检测结果，该模块由高置信度观测初始化。以往的三维多目标跟踪方法在此类情况下，要么终止轨迹，要么借助启发式运动模型（例如卡尔曼滤波器）延长轨迹。然而，这两种解决方案都可能因“过早终止” [43] 或错误关联而导致身份切换。相比之下，我们可学习的

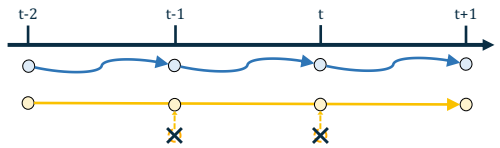


图4。**轨迹延伸**。 PF-Track 会在每个时间戳更新目标位置并预测未来轨迹（上排）。然而，如果无法可靠地定位目标（例如 由于遮挡或噪声观测，如帧 $t-1$ 和 $t$ 的下排所示），我们的方法将依靠从可靠时间戳（帧 $t-2$）预测的长期轨迹来推断该目标的位置，并忽略噪声观测（划掉的圆圈）。

运动预测模块和轨迹延伸策略更加准确、稳健。

我们在图 4中展示了高层直觉，并在第 B.2 节（附录）中提供了更多细节。在图 4中，长期轨迹有助于传播黄色实例的位置（下排）。当 PF-Track 遇到噪声观测或遮挡情况时，会依靠先前可靠帧中的运动预测来模拟被遮挡物体的移动。在极端情况下，我们的模型能够处理长达 $\tau_f - 1$ 帧的遮挡。第 4.4节中的消融研究表明，轨迹延伸可以大幅减少身份切换。据我们所知，我们是首个将长期预测纳入基于查询的框架，并在不进行显式重新识别的情况下处理遮挡的方法。

### 3.4. 损失函数

最终的损失函数定义如下：

$$\begin{aligned}\mathcal{L} = &\lambda_{\text{cls}}^D \mathcal{L}_{\text{cls}}^D + \lambda_{\text{box}}^D \mathcal{L}_{\text{box}}^D + \\ &\lambda_{\text{cls}}^R \mathcal{L}_{\text{cls}}^R + \lambda_{\text{box}}^R \mathcal{L}_{\text{box}}^R + \lambda_f \mathcal{L}_f\end{aligned} \quad (12)$$

其中 $\mathcal{L}_{\text{cls}}$ 和 $\mathcal{L}_{\text{cls}}^R$ 是焦点损失 [32]，其系数分别为 $\lambda_{\text{cls}}^D$ 和 $\lambda_{\text{cls}}^R$。它们分别监督$\mathbf{B}_t^D$和$\mathbf{B}_t^R$的分类分数。$\mathcal{L}_{\text{box}}^D$ 和 $\mathcal{L}_{\text{box}}^R$ 均为作用于$\mathbf{B}_t^D$和$\mathbf{B}_t^R$的 L1 损失，用于边界框回归。其系数为 $\lambda_{\text{box}}^D$ 和 $\lambda_{\text{box}}^R$。运动预测损失$\mathcal{L}_f$ 是预测轨迹与真实轨迹的运动之间的 L1 损失，其权重为 $\lambda_f$。真实标注分配会将一个查询与随时间保持一致的真实标注实例进行匹配，以鼓励身份一致性。更多细节请参见第 B 节（附录）。

## 4. 实验

### 4.1. 数据集与指标

**数据集**。 我们在大规模自动驾驶数据集 nuScenes 上进行了实验 [4]。该数据集包含 1,000 个视频序列，涵盖多种模态，包括来自 6 个周围摄像机的 RGB图像，以及来自激光雷达的点云

| | AMOTA ↑ | 平均多目标跟踪精度 ↓ | 召回率 ↑ | 多目标跟踪准确度 ↑ | IDS ↓ |
|---|---|---|---|---|---|
| 验证集 | | | | | |
| DEFT [7] | 0.201 | N/A | N/A | 0.171 | N/A |
| QD3DT [21] | 0.242 | 1.518 | 39.9% | 0.218 | 5646 |
| MUTR3D [73] | 0.294 | 1.498 | 42.7% | 0.267 | 3822 |
| TripletTrack [40] | 0.285 | 1.485 | N/A | N/A | N/A |
| CC-3DT* [14] | 0.429 | 1.257 | 53.4% | 0.385 | 2219 |
| PF-Track-S（我们的方法） | 0.408 | 1.343 | 50.7% | 0.376 | **166** |
| PF-Track-F（我们的方法） | **0.479** | **1.227** | **59.0%** | **0.435** | 181 |
| 测试集 | | | | | |
| CenterTrack [76] | 0.046 | 1.543 | 23.3% | 0.043 | 3807 |
| PermaTrack [57] | 0.066 | 1.491 | 18.9% | 0.060 | 3598 |
| DEFT [7] | 0.177 | 1.564 | 33.8% | 0.156 | 6901 |
| QD3DT [21] | 0.217 | 1.550 | 37.5% | 0.198 | 6856 |
| MUTR3D [73] | 0.270 | 1.494 | 41.1% | 0.245 | 6018 |
| TripletTrack [40] | 0.268 | 1.504 | 40.0% | 0.245 | 1144 |
| CC-3DT* [14] | 0.410 | 1.274 | **53.8%** | 0.357 | 3334 |
| PF-Track-F（我们的方法） | **0.434** | **1.252** | **53.8%** | **0.378** | **249** |

表1. 与 nuScenes 上最先进的基于摄像头的三维多目标跟踪算法进行比较 [4]。“**S**”和“**F**”分别表示采用低分辨率和全分辨率设置训练的模型（详见第 4.2节）。我们的方法在 AMOTA 和身份切换次数（全分辨率）方面均具有显著优势，其身份切换次数几乎减少了 90%，且比其他方法 低一个数量级 。（*）表示同期工作。

和雷达。本文仅使用摄像头传感器。每个序列大约持续 20 秒，关键帧以 2Hz 的频率进行标注。该数据集提供了 140 万个三维边界框，涵盖道路上 10 类常见物体。对于跟踪任务，nuScenes 选取了 7 个移动类别的子集，例如汽车、行人和摩托车，并排除了交通锥等静态物体。

**指标**。 我们严格遵循 nuScenes 针对多目标跟踪任务的官方评估指标。该指标通过考虑多个召回率阈值，对 CLEAR MOT 指标进行了修改 [2] 。主要指标为“平均多目标跟踪准确度”（AMOTA） [64]。同时，我们还考虑其他分析指标，例如“身份切换次数”（IDS）和“平均多目标跟踪精度”（AMOTP）。

## 4.2. 实现细节

由于篇幅有限，我们在此说明两种训练设置，并在第 B 节（附录）中介绍更多实现细节。在我们的实现中，每个训练样本都包含来自不同时间戳的三个相邻帧。然而，由于每帧包含六张高分辨率图像，因此计算开销很大。因此，受 PETR 的启发，我们采用两种设置，将图像下采样至不同分辨率 [33]。

**全分辨率**。在每个时间点，我们将原始分辨率的图像裁剪为 $1600 \times 900$ 到 $1600 \times 640$，排除天空区域。然而，以此分辨率训练多帧跟踪器无法在单张 A100 GPU 上完成。因此，我们

| 索引 | Past QR | Past TR | 未来 预测 | 未来 扩展 | AMOTA↑平均多目标跟踪精度↓ | | IDS↓ |
|---|---|---|---|---|---|---|---|
| 1 | | | | | 0.368 | 1.421 | 507 |
| 2 | ✓ | | | | 0.378 | 1.414 | 453 |
| 3 | ✓ | ✓ | | | 0.380 | 1.408 | 400 |
| 4 | | | ✓ | | 0.374 | 1.402 | 469 |
| 5 | | | ✓ | ✓ | 0.391 | 1.360 | **155** |
| 6 | ✓ | ✓ | ✓ | ✓ | **0.408** | **1.343** | 166 |

表 2. **PF-Track 模块消融实验**。 对于过去推理，“QR”和“TR”分别表示第 3.2 节中的“查询优化”和“轨迹细化”。对于未来推理，“Pred”和“Ext”分别表示第 3.3 节中的“运动预测”和“轨迹延伸”。过去推理和未来推理能够独立提升三维多目标跟踪性能，而 PF-Track 通过将二者结合到端到端框架中取得了最佳结果。

首先按照一些先前工作的做法，使用单帧检测对骨干网络进行 24 个 epoch 的预训练 [57]。然后固定骨干网络，并使用三帧样本训练跟踪器，再训练 24 个 epoch。我们仅在表中以“-F”标示的完整模型结果中采用这一设置。1。

**低分辨率**。除非另有说明，我们的所有消融分析均采用低分辨率设置。在此设置下，我们将裁剪图像下采样至 $800 \times 320$ 的分辨率。我们先训练单帧检测模型 12 个 epoch，然后使用三帧样本训练跟踪器，再训练 12 个 epoch。

## 4.3. nuScenes 上的最先进方法比较

在表1中，我们将模型性能与 nuScenes 上已发表的其他基于摄像头的三维多目标跟踪算法进行比较。我们的方法在各项指标上均取得显著提升，达到了新的最先进水平。与先前方法相比，我们的 AMOTA 在测试集上提升了 7% 以上，在验证集上提升了12%，其中包括一项非常强劲的同期工作 [14]。值得注意的是，在建立了更多轨迹（召回率更高）的同时，我们的身份切换数量仅为先前方法的10%，消除了超过90 % 的身份切换错误。这一结果表明，我们的算法具有很强的关联能力，这得益于对过去与未来信息的综合利用。即使在低分辨率设置下进行训练，我们模型的优势依然明显，而此前大多数工作都采用全分辨率。

## 4.4. 消融研究

**过去与未来推理的有效性**。在表2中，我们使用 nuScenes 验证集分析了各个模块对模型性能的重要性。具体而言，我们评估了以下变体。**(1) 基线**。我们的基线是不进行显式时空推理的“基于注意力的跟踪”模型（第 1 行）。这是一个强基线，在表1中优于先前工作。**(2) 过去推理**。我们首先分析查询优化的效果（第

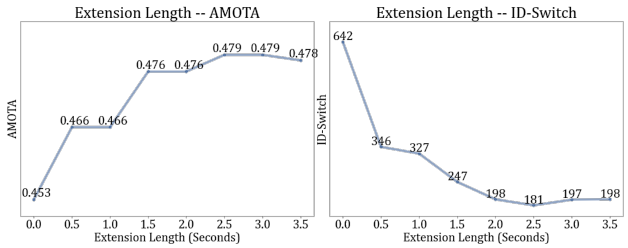


图 5. **轨迹延伸有助于多目标跟踪**。 利用预测轨迹维持低置信度轨迹的状态，可显著提升 AMOTA 并减少身份切换。

2 行），该方法通过交叉注意力显式引入历史查询。如图所示，这提升了整体跟踪质量。随后，我们利用增强后的特征细化轨迹的三维边界框（第 3 行）。这降低了身份切换和平均多目标跟踪精度，表明轨迹细化有助于三维多目标跟踪。**(3) 未来推理**。接下来，我们证明，学习预测目标运动并传播位置（第 4 行）有助于建模目标动态，并提升跟踪性能。此外，使用长期轨迹预测替换低置信度定位结果（第 5 行），可使身份切换减少 67%。**(4)过去与未来联合推理**。最后，将过去与未来推理结合到图2所示的端到端框架中（第 6 行），使模型取得最佳性能。这一结果证实，过去与未来推理对三维多目标跟踪具有相互促进的作用。

**轨迹延伸长度**。 在图 5中，我们使用在 nuScenes 验证集上表现最佳的全分辨率模型，分析轨迹延伸长度对 AMOTA 和身份切换的影响。与不采用延伸策略（0.0 秒）相比，延长轨迹可显著提升性能，最长可达 2 秒。之后，各项指标趋于饱和，因为经过如此长时间后重新出现的物体很少。 请注意，这些改进是在未进行显式重新识别的情况下取得的。

**未来推理中的预测长度**。 接下来，我们在表 3中分析预测长度对 nuScenes 验证集上三维多目标跟踪性能的影响。具体而言，我们训练了三个完整模型，其预测长度分别为 2.0、3.0 和 4.0 秒。AMOTA 和身份切换指标表明，4.0 秒（8 帧）略优于 2.0 秒（4 帧）和 3.0 秒（6 帧）。这一结果说明，学习更长时间范围的轨迹预测有助于提升我们的三维多目标跟踪框架。

**与“检测式跟踪”基线方法的比较**。在表 4中，我们比较了端到端框架与以往“检测式跟踪”算法的性能 [43,64,69],，这些算法是基于激光雷达的三维多目标跟踪的强基线方法。在这些实验中，我们也使用了验证集

| 运动预测时长扩展 | | AMOTA ↑ | 平均多目标跟踪精度 ↓ | IDS ↓ |
|---|---|---|---|---|
| 2.0s | ✗ | 0.392 | 1.376 | 604 |
| 2.0s | ✓ | 0.402 | 1.342 | 217 |
| 3.0s | ✗ | 0.392 | 1.372 | 540 |
| 3.0s | ✓ | 0.402 | **1.340** | 208 |
| 4.0s | ✗ | 0.391 | 1.387 | 471 |
| 4.0s | ✓ | **0.408** | 1.343 | **166** |

表 3. **运动预测时长**。“扩展”指使用“轨迹延伸”。我们训练了三个模型，其预测时域分别为 2.0 秒、3.0 秒和 4.0 秒。根据 AMOTA 和 IDS，学习更长的预测时域有利于跟踪。

| | AMOTA ↑ | 平均多目标跟踪精度 ↓ | IDS ↓ |
|---|---|---|---|
| AB3DMOT [64] | 0.292 | 1.333 | 2419 |
| AB3DMOT [64]$^{\Psi}$ | 0.329 | 1.388 | 2677 |
| CenterPoint [69] | 0.233 | **1.270** | 2715 |
| CenterPoint [69]$^{\Psi}$ | 0.383 | 1.329 | 3082 |
| SimpleTrack [43] | 0.320 | 1.295 | 1606 |
| SimpleTrack [43]$^{\Psi}$ | 0.402 | 1.324 | 2053 |
| PF-Track（我们的方法） | **0.408** | 1.343 | **166** |

表 4. **与“基于检测的跟踪”进行比较**。我们将三维多目标跟踪中的强基线应用于 PETR [33]: AB3DMOT [64], CenterPoint [69], 和 SimpleTrack [43]。“Ψ”表示我们调整这些方法的超参数以适配 PETR 检测结果，而非使用其原始配置。我们的端到端方法具有显著优势。

| 方法 | 平均位移误差 ↓（@4.0秒） | 最终位移误差 ↓（@4.0秒） |
|---|---|---|
| 长短期记忆网络 [8] | 2.32 | 2.87 |
| VectorNet [16] | 2.01 | 2.48 |
| 速度 | 2.10 | 2.64 |
| PF-Track（我们的方法） | **1.88** | **2.38** |

表 5. **基于特征或抽象状态的运动预测**。我们基于 PF-Track 在 nuScenes 验证集上的真阳性轨迹构建了一个运动预测基准，并使用轨迹的三维位置训练长短期记忆网络 [8]和 VectorNet [16]。“速度”一行是在恒定速度运动模型假设下得到的结果。结果表明，基于特征进行预测能够提供更丰富的信息，从而提升轨迹质量。

划分的 nuScenes 数据集。为公平比较，我们使用 PETR [33]检测结果评估这些方法，并针对 AMOTA 调整其超参数（表中以“Ψ”标示，详情见第B.4 节（附录）。结果清楚地表明，与更传统的模块化框架相比，我们的端到端方法具有优势，在 ID 切换指标上的提升尤为显著。

**对查询特征进行预测分析**。尽管本文重点关注多目标跟踪，我们还分析了预测性能。我们表明，直接基于目标特征进行端到端预测，优于

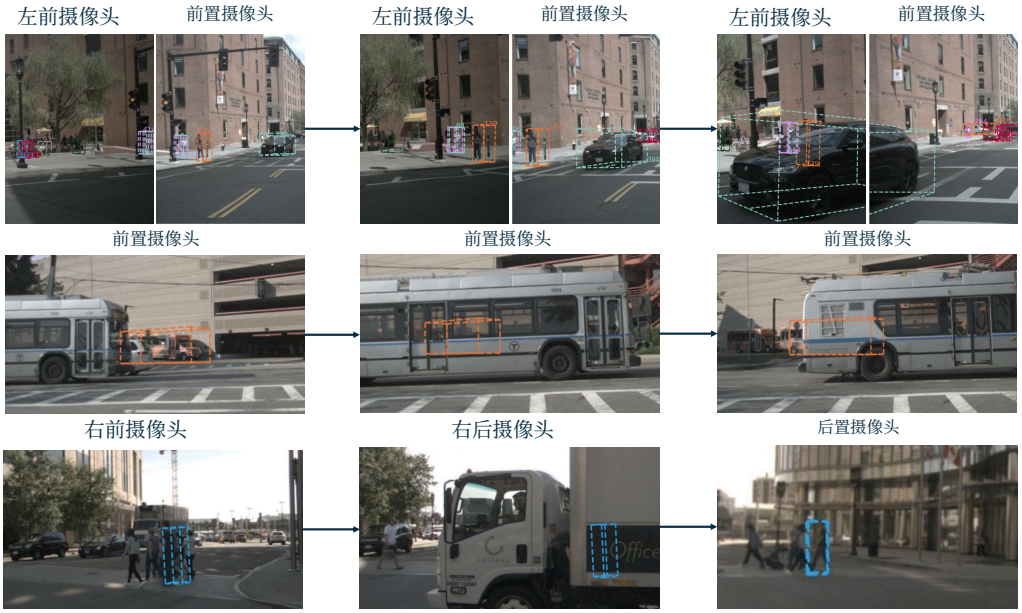


图 6。**三维多目标跟踪的定性结果**。（1）在顶部行中，我们展示了图像级三维多目标跟踪结果。这些图突出了不同图像之间的一致性，例如穿过前置摄像头左侧和前置摄像头的车辆。（2）在中间行和底部行中，我们分别展示了两个专门用于处理大物体和小物体遮挡的示例。

基于低层目标状态（如中心位置）进行预测。具体而言，我们使用 PF-Track 的真阳性轨迹训练了两个运动预测基线：VectorNet [16] 和 LSTM [8], ，遵循以往研究的做法 [39], ，并在表5的前几行报告其结果。由于我们的方法不使用高清地图，为了公平比较，我们在这些实验中排除了运动预测算法中处理高清地图的部分。此外，我们还报告了另一个基线：假设运动模型的速度恒定，使用模型解码器预测的速度进行轨迹预测（表5第三行）。评估指标为“平均位移误差”（ADE）和“最终位移误差”（FDE），数值越低越好。更多详情请参见B.5（附录）。

表 5 比较了端到端 PF-Track 与上述基线方法在 nuScenes 验证集上的性能。ADE 和 FDE 越低，PF-Track 的轨迹质量越好。我们的结论与先前研究一致 [18, 39, 63]。具体而言，LSTM 是一种浅层模型，无法从噪声轨迹中捕捉有意义的动态特征；更强的 VectorNet 模型优于其他基线方法，但其表现仍不如我们方法所提出的在端到端框架中进行轨迹预测。

### 4.5. 定性结果

我们在图中通过将三维边界框投影到图像上，展示三维多目标跟踪结果。6 边界框的颜色从七种颜色组成的调色板中随机选取，并根据其 ID 分配，因此每个目标的颜色始终一致随时间变化的颜色。

第一行展示了多摄像头三维多目标跟踪的整体可视化结果，重点关注左前方和前置摄像头。如图所示，PF-Track 能够连贯地跟踪目标，尤其是出现在两个独立摄像头画面中的行人和车辆。图6的下两行展示了两个解决遮挡问题的示例。对于大型目标（公交车）和小型目标（行人），我们的方法会在目标被遮挡的帧中传播其位置，并在遮挡解除后的帧中成功重新关联目标，即使此时目标出现在不同的摄像头画面中。值得强调的是，这一过程无需显式的重识别模块。

## 5. 结论

本文提出了一种用于多摄像头三维多目标跟踪的基于查询的端到端方法，以增强时空一致性。通过过去推理，我们的框架利用历史信息增强查询特征和轨迹质量。通过未来推理，预测轨迹能够在相邻帧之间以及被遮挡的长时间段内更好地传播查询。我们还证明，过去与未来联合推理进一步增强了跟踪器的能力。在大规模 nuScenes 数据集上进行的大量评估表明，我们的方法能够有效地生成连贯轨迹。

**致谢**。 本研究部分得到了丰田研究院、美国国家科学基金会资助项目 2106825、美国国家食品与农业研究所资助项目 2020-67021-32799 以及 NCSA Fellows 计划的支持。

# 附录

我们的附录介绍了额外的实验分析和实现细节。目录如下：(A) **视频演示**。 我们提供了一个多摄像头 3D 多目标跟踪（多目标跟踪）的演示，如第 A 节所述。(B) **实现细节**。我们在第 B 节介绍详细的模型架构、训练和推理流程，以及消融研究的设置。(C) **额外消融实验**。 我们在第 C 节提供更多分析和实验结果。(D) **性能验证**。 为核验结果，我们在第 E节提供测试集结果的截图以供验证。

## A. 多摄像头跟踪视频演示

我们的演示视频位于https://youtu.be/eJghONb2AGg。内容包括：
• 在环视图像和鸟瞰图上展示三维多目标跟踪结果。• 解决遮挡问题的示例。• 预测轨迹的定性结果。

## B. 实现细节

### B.1. 模型架构

我们将在以下各节中解释 PF-Track 的设计选择。为了从整体上展示模型，我们在本补充材料中附上 mmdetection3d [11]格式的配置文件。

**骨干网络**。 我们使用 VoVNetV2 [27] 作为骨干网络。对于特征金字塔 [31],，C5特征（第 5 个阶段的输出）经过上采样后，与 C4特征（第 4 个阶段的输出）进行融合。为了在训练期间节省 GPU显存，我们默认采用检查点技术 [9]。

**检测头**。 我们遵循 PETR [33]的设计，将区域设置为 XY 轴上的 $[-51.2m, 51.2m]$ 和 Z 轴上的 $[-5m, 3m]$。边界框的中心分别归一化为$[0, 1]$,。检测头由 6 个 Transformer 解码器层 [58] 以及用于边界框回归和分类的 2 个 MLP 头组成。每个 Transformer 解码器层的嵌入维度为 256，前馈维度为 2048。丢弃概率为 0.1。两个 MLP 头均为两层 MLP。边界框回归头预测物体的中心（$x$, $y$, $z$）、尺寸（$l$, $w$, $h$）、朝向和速度，而分类头则返回每个类别的逻辑值。

**PF-Track 配置**。我们的 PF-Track 在每帧使用固定数量的 500 个检测查询。跟踪查询的训练和推理流程则有所不同。训练方案会在查询与真实标注（<2.0m）形成正匹配后，将其加入跟踪查询集合，并使跟踪查询始终与同一目标的真实标注相关联。推理方案则会在相应的置信度分数大于 0.4 时初始化跟踪查询。在推理时，我们的模型最多输出 300 个物体（与 [33, 61, 73] 相同），并设置 0.2 的最低分数阈值。

**算法A**“查询传播”算法。

**输入：**
$\mathbf{Q}_{t-1}$、$\mathbf{C}_{t-1}$、$\mathbf{M}_{t-1:t+\tau_f-1}^{t-1}$：来自帧 $t-1$ 中目标的查询、中心位置和运动预测；$N$：查询数量。

**输出：**
$\hat{\mathbf{Q}}_{t-1}$、$\hat{\mathbf{C}}_{t-1}$、$\hat{\mathbf{M}}_{t:t+\tau_f}^{t-1}$：从帧$t-1$ 到帧 $t$传播的查询、中心位置和目标运动预测。1：对于 $i=0$ 到 $N-1$，执行 2：$\hat{\mathbf{C}}_{t-1}^{i} \leftarrow \mathbf{C}_{t-1}^{i} + \mathbf{M}_{t-1\,t}^{t-1\,i}$，：3：$\hat{\mathbf{Q}}_{t-1}^{i} \leftarrow \mathbf{Q}_{t-1}^{i}$4：$\hat{\mathbf{M}}_{t:t+\tau_f}^{t-1} \leftarrow$ 填充$(\mathbf{M}_{t:t+\tau_f-1}^{t-1})$ 5：结束循环 6：返回 $\hat{\mathbf{Q}}_{t-1}$、$\hat{\mathbf{C}}_{t-1}$、$\hat{\mathbf{M}}_{t:t+\tau_f}^{t-1}$

**过去推理**。 查询队列的长度为 1.5 秒（$\tau_h = 3$ 帧），因为我们的每个训练样本都包含 3 帧。不过，我们强调，由于过去推理模块具有循环结构，PF-Track 能够在推理 期间聚合整个视频中的历史信息。具体而言，帧 $t$ 中的查询会关注帧 $[t-2, t]$，而帧 $t-2$中的查询则会进一步关注帧 $[t-4, t-2]$，以此类推。因此，帧 $t$ 中的查询可以访问所有先前帧的信息。

对于跨帧处理和跨帧注意力，我们分别为二者采用两个 Transformer 解码器层。每个 Transformer 解码器层的嵌入维度为 256，前馈维度为 2048。与检测头中的 Transformer 层相同，这两个层的丢弃概率也为 0.1。轨迹细化模块使用两个独立的 2 层 MLP 头，分别用于回归和分类。

**未来推理**。未来推理模块预测未来 4.0 秒（$\tau_f = 8$帧）内的运动。首先，它使用相同的跨帧注意力生成未来特征，然后应用一个 2 层多层感知机，将特征转换为 XY 平面上的运动。对于轨迹延伸，我们选择能使验证集上的 AMOTA 达到最大的延伸长度：低分辨率模型为 $2.0s$（4 帧），全分辨率模型为 $2.5s$（5 帧）。

**损失权重**。在主论文的3.4 节中，我们使用系数来平衡各损失项。边界框回归损失为$\lambda_{box}^{D} = 0.25$，分类焦点损失为 $\lambda_{cls}^{D} = 2.0$，与 [33]相同。对于轨迹细化部分，我们分别采用相同的边界框回归和分类损失权重： $\lambda_{box}^{R} = 0.25$ 和 $\lambda_{cls}^{R} = 2.0$。未来推理中运动预测的权重为 $\lambda_f = 0.5$。

### B.2. “轨迹延伸”算法

我们详细说明第 3.3 节（主论文）中所述轨迹扩展算法的具体步骤。为清晰起见，我们严格地

**算法B**“轨迹延伸”算法。**输入：** $\hat{\mathbf{Q}}_{t-1}$, $\hat{\mathbf{C}}_{t-1}$, $\hat{\mathbf{M}}_{t}^{t-1}$ : $_{t+\tau_f}$：来自帧 $t-1$ 的传播查询、中心位置和运动预测；$\mathbf{F}_t$：帧 $t$ 上的图像特征；$\mathbf{L}_t$：查询已延伸的帧数

d

连续地；$\tau_e$：扩展的最大帧数；

$N$：查询数量。

**输出：**

$\hat{\mathbf{Q}}_t$, $\hat{\mathbf{C}}_t$, $\hat{\mathbf{M}}_{t+1}^{t}$:$_{t+\tau_f+1}$：从帧 $t$ 到帧 $t+1$ 被跟踪物体的传播查询、中心位置和运动预测。1：$\mathbf{Q}_t^R$, $\mathbf{C}_t^R$, $\mathbf{M}_{t:t+\tau_f}^{t}$ ← PF-Track($\mathbf{F}_t$, $\mathbf{Q}_t$, $\mathbf{C}_t$) 2：对于 $i=0$ 到 $N-1$，执行以下操作 3：如果置信度分数 $S^{R}$,$_t$ $^i$ 低于阈值，则 4：如果 $i$ 是活动目标 $L_t^i < \tau_e$，则 5：然后使用第 6-8 行中的先前信息：6：$\hat{\mathbf{C}}_t^i \leftarrow \hat{\mathbf{C}}_{t-1}^i +$ $\hat{\mathbf{M}}_{t\ t+1}^{t-1\ i}$, : 7：$\hat{\mathbf{Q}}_t^i \leftarrow \hat{\mathbf{Q}}_{t-1}^i$8：$\hat{\mathbf{M}}_{t+1}^{t}$:$_{t+\tau_f+1}$ ← 填充($\hat{\mathbf{M}}_{t+1}^{t-1}$:$_{t+\tau_f}$) 9：记录扩展：$L_t^i \leftarrow L_t^i + 1$10：否则 11：终止第 $i$ 个目标的轨迹。12：将其从记忆中移除。13：结束 if 14：否则 15：按照第 15-18 行正常传播 16：$\hat{\mathbf{C}}_t^i \leftarrow \mathbf{C}_t^{R\ i}$ +, $\mathbf{M}_t^{t\ i}$, :$_{t+1}$17：$\hat{\mathbf{Q}}_t^i \leftarrow \mathbf{Q}_t^{R\ i}$, 18：$\hat{\mathbf{M}}_{t+1}^{t}$:$_{t+\tau_f+1}$ ← 填充($\mathbf{M}_{t+1:t+\tau_f}^{t}$) 19：将连续扩展置零：$L_t^i \leftarrow 0$20：结束 if 21：结束 for 22：返回 $\hat{\mathbf{Q}}_t$, $\hat{\mathbf{C}}_t$, $\hat{\mathbf{M}}_{t+1}^{t}$:$_{t+\tau_f+1}$

先在算法 A 中描述查询传播，然后在算法 B中介绍轨迹延伸。

ˆ 如算法 A所示，传播后的中心位置 $\mathbf{C}_{t-1}$ 由当前状态 $\mathbf{C}_{t-1}$ 与运动预测 $\mathbf{M}_{t-1\ :t}^{t-1}$相加得到。ˆ 查询 $\mathbf{Q}_{t-1}$ 的传播直接复用最新结果 $\mathbf{Q}_{t-1}$。运动预测的传播可以使用简单的填充函数，因为我们的轨迹延伸不会超过运动预测时长。

接下来，我们在算法B中展示加入轨迹延伸后的算法。主要区别在于如何处理低置信度目标（第 3–14 行），这类目标可能存在噪声。轨迹延伸并非总是根据最新结果进行更新，而是更多地依赖先前帧的结果，以提高保真度（第 5–8 行）。值得注意的是，如果某个目标在 $\tau_e$ 连续帧中都没有置信度较高的结果，我们会按照通行做法终止该轨迹（第 10–13 行）。请注意，默认情况下，我们将$\tau_e$ 设为小于预测时域 $\tau_f = 8$ 帧（4.0 秒）。如表2所示，轨迹延伸有效提升了 AMOTA，并减少了 IDS。

## B.3. 模型训练

每个训练样本由三个相邻帧组成。PF- Track 使用 AdamW 优化器 [25, 37] 进行训练，权重衰减为 0.01。学习率初始为 $2.0 \times 10^{-4}$，并按照余弦退火进行调度 [36]。上述过程与基于图像的 3D 检测方法 [33, 61]相同。不过，由于图像数据增强可能会破坏物体的运动模型，我们在跟踪器训练期间禁用数据增强。训练轮数遵循第 4.2 节（主论文）中讨论的设置。全分辨率设置在 8×A100 GPU 上需要 3 天，小分辨率设置在 8×A100 GPU 上需要 1 天。

## B.4. “检测式跟踪”实验

第 4.4 节（主论文）中“检测式跟踪”基线方法的实验根据验证集上的 AMOTA 调整检测分数阈值。最终，我们将输出边界框的最低分数阈值设为 0.2，这与 PF-Track 相同。根据表 4 （主论文），我们的超参数调优显著提升了“检测式跟踪”基线方法的性能，并实现了公平比较。原因在于，AB3DMOT [64], CenterPoint [69], 和 SimpleTrack [43] 针对基于激光雷达的 3D 检测设计了跟踪器，并采用较低的（例如，SimpleTrack 中为 0.01）或不设分数阈值。然而，基于图像的检测包含更多假阳性，因此需要更强的过滤。

## B.5. 基于抽象目标状态的运动预测设置

本节将介绍第 4.4 节（主论文）中“基于查询特征的预测”的详细内容。

**数据集构建。** 我们从 PF-Track 的输出中召回真阳性轨迹，以构建运动预测数据集。具体而言，我们对预测边界框与真实标注进行匈牙利匹配，并将与真实标注距离小于 2.0 米的匹配判定为正匹配。我们采用 nuScenes [4] 数据集的训练集/验证集划分，并移除没有正匹配的帧。最终，运动预测训练集和验证集中分别包含 27,960 帧和 5,879 帧，以及 422,167 条和 69,804 条轨迹。

**模型训练与推理。** 我们的模型架构与 Argoverse [8] 和 Vector-Net [16]中的长短期记忆网络模型相同。为进行公平比较，我们提供 2.0 秒的历史轨迹，并要求模型预测最长 4.0 秒，这与我们的 PF-Track 相同。由于每帧包含多条轨迹，我们的实现参照 [13]，根据自车的位置对坐标进行归一化。模型的输入包括 XY 平面上的边界框中心和速度。对于缺失的观测，我们将输入补齐至完整长度，并使用

| 模型 | 扩展 | AMOTA ↑ | 平均多目标跟踪精度 ↓ | IDS ↓ |
|---|---|---|---|---|
| PF-Track | | 0.391 | 1.387 | 471 |
| | ✓ | **0.408** | **1.343** | 166 |
| 不含跨帧 | | 0.385 | 1.402 | 410 |
| | ✓ | 0.395 | 1.373 | 171 |
| 无/ 跨目标 | | 0.383 | 1.390 | 481 |
| | ✓ | 0.399 | 1.352 | **155** |

表 A. **跨帧和跨对象注意力**。“扩展”表示是否采用轨迹延伸。在过去推理的“查询优化”模块（主论文第 3.2 节）中移除跨帧或跨对象机制，都会对模型性能产生负面影响。

| 模型 | 扩展 | AMOTA ↑ | 平均多目标跟踪精度 ↓ | IDS ↓ |
|---|---|---|---|---|
| PF-Track | | 0.391 | 1.387 | 471 |
| | ✓ | **0.408** | **1.343** | **166** |
| 速度传播 | | 0.374 | 1.389 | 478 |
| | ✓ | 0.382 | 1.359 | 192 |

表 B. **用于查询传播的速度**。“扩展”表示使用轨迹延伸。在未来推理（主论文第 3.3 节）中使用速度而非学习得到的轨迹进行查询传播，会对模型性能产生负面影响。

最近的观测，并用 0-1 掩码表示填充。我们使用 AdamW 在自定义运动预测数据集上训练模型 24 个周期 [25, 37]优化器，初始学习率为 $1.0 \times 10^{-3}$。学习率在第0。1 个和第 20 个周期下降。

# C. 补充消融研究

## C.1. 过去推理中的跨帧和跨对象注意力。

我们分析了过去推理（主论文第3.2 节）中跨帧和跨对象注意力对“查询优化”的影响。如表A所示，如果从最终的 PF-Track 中移除其中任一项，性能都会下降，尤其是 AMOTA。尽管 PF-Track 的身份切换略多，但我们认为这是由于其 AMOTA 较高所致。因此，跨帧和跨对象注意力对于多摄像头三维多目标跟踪都很有用。

## C.2. 查询传播使用轨迹还是速度

为验证学习轨迹预测对于查询传播的必要性，我们在表B中进行了使用速度进行传播的实验。具体而言，我们训练了一个与 PF-Track 完全相同的模型，不同之处在于使用速度变换查询在不同帧间的位置。如表B所示，与使用速度的变体相比，使用学习到的轨迹取得了更好的性能。因此，我们得出结论：学习长期轨迹预测对于实现稳健、准确的三维多目标跟踪是必要的。

## C.3. 未来推理中的运动预测权重

我们分析了运动预测损失权重对跟踪性能的影响。我们发现，跟踪性能对运动预测损失权重非常敏感。

tion $\lambda_f$（主论文第 3.4 节）。在表 C 中，我们改变

| $\lambda_f$ | 扩展 | AMOTA ↑ | 平均多目标跟踪精度 ↓ | IDS ↓ |
|---|---|---|---|---|
| 0.25 | | 0.389 | 1.382 | 539 |
| | ✓ | 0.396 | 1.353 | 185 |
| 0.50 | | 0.391 | 1.387 | 471 |
| | ✓ | **0.408** | **1.343** | **166** |
| 1.00 | | 0.377 | 1.395 | 489 |
| | ✓ | 0.397 | 1.346 | 183 |

表 C。 **运动预测的损失权重**。“$\lambda_f$”表示运动预测的损失权重 $\lambda_f$（主论文第 3.4 节），“扩展”表示是否使用轨迹延伸。该表展示了运动预测的敏感性，并呼吁未来工作关注更好的多任务学习策略。

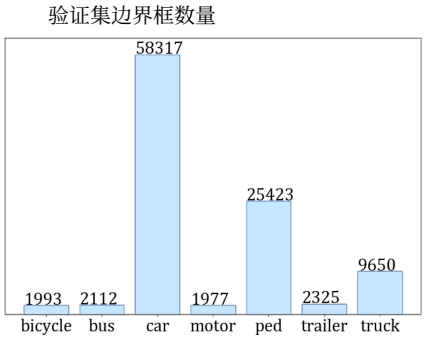


图 A. nuScenes 验证集中各类别实例数量的分布。“motor”表示“摩托车”，“ped”表示“行人”。如图所示，nuScenes 上的数据分布并不均衡。

使用 $\lambda_f = [0.25, 0.50, 1.00]$ 进行运动预测，它们可能会导致跟踪性能出现差异。这需要未来研究予以关注，或采用更好的多任务学习策略

## C.4. 类别级分析

我们比较了使用过去与未来推理前后的类别级性能。为了介绍 nuScenes 数据集的背景 [4], 我们首先在图 A 中直观展示各类别的数据分布。如图所示，nuScenes 的类别分布并不均衡，“汽车”和“行人”占据了大多数物体。

接着，在表 D 中，我们比较了三种主要类型的类别级性能：汽车、行人和卡车。与不使用任何过去或未来推理的基线（与主论文表 1 中的基线相同）相比，我们提出的方法在各个类别上均显著优于基线。

### C.5. 使用 DETR3D 的 PF-Track

我们补充分析了将 PF-Track 应用于另一种三维检测器：DETR3D [61]。

**实验设置**。 遵循 DETR3D [61]，，我们使用 VoVNetV 2 [27] 作为骨干网络，并通过 FPN [31] 融合 C2-C5 的特征，作为图像特征。由于 DETR3D 仅支持全分辨率设置，我们首先使用单帧检测训练骨干网络 24 个 Epoch，然后固定骨干网络，使用三帧跟踪训练整个跟踪器，再训练 24 个 Epoch。一轮训练在 8 × A100 GPU 上需要 3 天。

```
Calculating metrics...
Saving metrics to: /tmp/tmp_bimu3cd/nuscenes-metrics

### Final results ###

Per-class results:
                AMOTA   AMOTP   RECALL  MOTAR   GT      MOTA    MOTP    MT      ML      FAF     TP      FP      FN      IDS     FRAG    TID     LGD
bicycle         0.322   1.442   0.387   0.762   2186    0.295   0.635   46      102     14.7    845     201     1341    0       8       1.48    1.87
bus             0.408   1.257   0.544   0.605   1701    0.329   0.720   36      33      25.3    924     365     776     1       21      1.92    2.48
car             0.622   0.916   0.719   0.777   68518   0.558   0.520   2443    1371    186.6   49194   10956   19225   99      415     0.87    1.19
motorcy         0.448   1.245   0.457   0.840   1945    0.384   0.561   41      87      11.5    888     142     1057    0       5       1.31    1.61
pedestr         0.451   1.279   0.541   0.736   34010   0.395   0.738   734     873     96.9    18250   4827    15618   142     273     1.52    2.16
trailer         0.380   1.421   0.528   0.636   2566    0.334   0.884   57      64      48.7    1350    492     1212    4       23      0.99    1.61
truck           0.405   1.206   0.592   0.596   8639    0.352   0.658   191     178     51.1    5107    2065    3529    3       94      1.11    2.07

Aggregated results:
AMOTA   0.434
AMOTP   1.252
RECALL  0.538
MOTAR   0.707
GT      17080
MOTA    0.378
MOTP    0.674
MT      3548
ML      2708
FAF     62.1
TP      76558
FP      19048
FN      42758
IDS     249
FRAG    839
TID     1.32
LGD     1.86
Eval time: 5146.9s

Completed evaluation for test phase
```

图 B。**测试集结果截图**。这是主论文表格中结果的补充材料。1（主论文）。

| 指标 | PF-Track | Car | 行人 | 卡车 | All |
|---|---|---|---|---|---|
| AMOTA ↑ | | 0.562 | 0.372 | 0.374 | 0.358 |
| | ✓ | **0.579** | **0.415** | **0.403** | **0.408** |
| 平均多目标跟踪精度 ↓ | | 1.058 | 1.431 | 1.341 | 1.419 |
| | ✓ | **1.021** | **1.362** | **1.288** | **1.343** |
| IDS ↓ | | 368 | 154 | 25 | 507 |
| | ✓ | **67** | **83** | **7** | **166** |

表 D. **各类别跟踪指标分析**。“全部”表示所有七个类别的指标平均值。在 nuScenes 的三个主要类别上，与不涉及过去或未来推理的基线相比，我们的 PF-Track 实现了显著且稳定的提升。

| 指标 | PF-Track | Car | 行人 | 卡车 | All |
|---|---|---|---|---|---|
| AMOTA ↑ | | 0.583 | 0.370 | 0.305 | 0.344 |
| | ✓ | **0.600** | **0.431** | **0.310** | **0.362** |
| 平均多目标跟踪精度 ↓ | | 1.038 | 1.436 | 1.399 | 1.419 |
| | ✓ | **1.001** | **1.338** | **1.366** | **1.363** |
| IDS ↓ | | 394 | 222 | 32 | 680 |
| | ✓ | **172** | **116** | **5** | **300** |

表 E. **适用于 DETR3D 检测头的 PF-Track**。“全部”表示七个类别的平均指标数值，“PF-Track”表示使用 PF-Track 的过去与未来推理。在 nuScenes 的三个主要类别及平均指标上，PF-Track 能够提升跟踪性能。

**PF-Track 的有效性**。我们在表 E 中分析了 PF-Track 不采用或采用过去与未来联合推理时的性能。我们列出了排名前三的主要类别以及全部类别的平均结果。我们的过去与未来推理还显著提升了 AMOTA，并减少了身份切换。这一结果表明，PF-Track 对其他基于查询的检测器也具有良好的泛化能力。

## D. 局限性与未来工作

**处理高清地图以实现端到端预测**。PF-Track 的主要目标是三维多目标跟踪，由于未考虑高清地图，它并不是一个完整的运动预测流程。然而，随着轨迹质量的提升，我们对过去与未来的推理可能有益于下游运动预测。因此，未来可以考虑将高清地图作为我们方法的可选输入，并探索端到端运动预测。

**传感器模态**。当前的 PF-Track 是一个通用的基于查询的框架。因此，我们可以将 PF-Track 扩展到多摄像头设置之外，并将激光雷达或雷达纳入我们的框架。

## E. 测试集划分截图

我们在图 B 中提供了提交至 nuScenes 测试集的截图，作为表 1（主论文）中测试集结果的补充。

## 参考文献

[1] Adil Kaan Akan 和 Fatma G¨uney. Stretchbev：在空间和时间维度上拓展未来实例预测。载于 *ECCV*，2022年。 3[2] Keni Bernardin 和 Rainer Stiefelhagen。评估多目标跟踪性能：CLEAR MOT 指标。 *EURASIP* 图像与视频处理期刊，2008年。 6[3] Alex Bewley、Zongyuan Ge、Lionel Ott、Fabio Ramos 和 Ben Upcroft。简单的在线实时跟踪。载于 *ICIP*，2016年。 2[4] Holger Caesar、Varun Bankiti、Alex H Lang、Sourabh Vora、Venice Erin Liong、Qiang Xu、Anush Krishnan、Yu Pan、Gi-

ancarlo Baldan 和 Oscar Beijbom。nuScenes：用于自动驾驶的多模态数据集。载于 CVPR，2020年。1、2、5、6、10、11[5] Nicolas Carion、Francisco Massa、Gabriel Synnaeve、Nicolas Usunier、Alexander Kirillov 和 Sergey Zagoruyko。使用 Transformer 进行端到端目标检测。载于 ECCV，2020年。2、3[6] Sergio Casas、Abbas Sadat 和 Raquel Urtasun。MP3：用于地图构建、感知、预测和规划的统一模型。载于 CVPR，2021年。3[7] Mohamed Chaabane、Peter Zhang、J. Ross Beveridge 和 Stephen O'Hara。DEFT：用于跟踪的检测嵌入。arXiv 预印本 arXiv:2102.02267，2021年。1、6[8] Ming-Fang Chang、John Lambert、Patsorn Sangkloy、Jag- jeet Singh、Slawomir Bak、Andrew Hartnett、De Wang、Peter Carr、Simon Lucey、Deva Ramanan 和 James Hays。Argo- verse：利用丰富地图进行3D跟踪与预测。载于 CVPR，2019年。1、3、7、8、10[9] Tianqi Chen、Bing Xu、Chiyuan Zhang 和 Carlos Guestrin。以次线性内存开销训练深度神经网络。arXiv 预印本 arXiv:1604.06174，2016年。9[10] Hsu-kuang Chiu、Jie Li、Rares. Ambrus. 和 Jeannette Bohg。面向自动驾驶的概率式3D多模态、多目标跟踪。载于 ICRA，2021年。2[11] MMDetection3D 贡献者。MMDetection3D：用于通用3D目标检测的 OpenMMLab 新一代平台。https://github.com/open- mmlab/ mmdetection3d，2020年。9[12] Patrick Dendorfer、Vladimir Yugay、Aljo ˇ sa O ˇsep 和 Laura Leal-Taix´e。Quo Vadis：轨迹预测是实现长期多目标跟踪的关键吗？载于 NeurIPS，2022年。3[13] Scott Ettinger、Shuyang Cheng、Benjamin Caine、Chenxi Liu、Hang Zhao、Sabeek Pradhan、Yuning Chai、Ben Sapp、Charles R Qi、Yin Zhou 等。面向自动驾驶的大规模交互式运动预测：Waymo 开放运动数据集。载于 ICCV，2021年。1、3、10[14] Tobias Fischer、Yung-Hsu Yang、Suryansh Kumar、Min Sun 和 Fisher Yu。CC-3DT：通过跨摄像头融合实现全景三维目标跟踪。载于 CoRL，2022年。1、2、6[15] Huan Fu、Mingming Gong、Chaohui Wang、Kayhan Bat- manghelich 和 Dacheng Tao。用于单目深度估计的深度序数回归网络。载于 CVPR，2018年。2[16] Jiyang Gao、Chen Sun、Hang Zhao、Yi Shen、Dragomir Anguelov、Congcong Li 和 Cordelia Schmid。VectorNet：从向量化表示中编码高精地图和智能体动态。载于 CVPR，2020年。3、4、7、8、10[17] Cl´ement Godard、Oisin Mac Aodha、Michael Firman 和 Gabriel J. Brostow。深入探究自监督单目深度估计。载于 ICCV，2019年。2[18] Junru Gu、Chenxu Hu、Tianyuan Zhang、Xuanyao Chen、Yilun Wang、Yue Wang 和 Hang Zhao。ViP3D：基于三维智能体查询的端到端视觉轨迹预测。arXiv 预印本 arXiv:2208.01582，2022年。1、8[19] Vitor Guizilini、Rares Ambrus、Sudeep Pillai、Allan Raven- tos 和 Adrien Gaidon。用于自监督单目深度估计的三维打包。载于 CVPR，2020年。2

[20] Anthony Hu、Zak Murez、Nikhil Mohan、Sof´ıa Dudas、Jef- frey Hawke、Vijay Badrinarayanan、Roberto Cipolla 和 Alex Kendall。FIERY：利用环视单目摄像头在鸟瞰视角下进行未来实例预测。载于 ICCV，2021年。3

[21] Hou-Ning Hu、Yung-Hsu Yang、Tobias Fischer、Trevor Dar- rell、Fisher Yu 和 Min Sun。单目准密集三维目标跟踪。《IEEE模式分析与机器智能汇刊》，2022年。1、2、6[22] Junjie Huang、Guan Huang、Zheng Zhu 和 Dalong Du。BEVDet：鸟瞰视角下的高性能多摄像头三维目标检测。arXiv 预印本 arXiv:2112 11790，2021年 1、2 . . [23]Boris Ivanovic、Kuan-Hui Lee、Pavel Tokmakov、Blake Wulfe、Rowan McAllister、Adrien Gaidon 和 Marco Pavone。结合类别不确定性的异构智能体轨迹预测。arXiv 预印本 arXiv:2104 12446，2021年 1 . . [24] Boris Ivanovic 和 Marco Pavone。The Trajectron：利用动态时空图进行概率多智能体轨迹建模。载于 ICCV，2019年。3[25] Diederik P Kingma 和 Jimmy Ba。Adam：一种随机优化方法。arXiv 预印本 arXiv:1412 6980，2014年 10、11 . . [26] Alex H Lang、Sourabh Vora、Holger Caesar、Lubing Zhou、Jiong Yang 和 Oscar Beijbom。PointPillars：用于从点云中检测目标的快速编码器。载于 CVPR，2019年。2[27] Youngwan Lee、Joong-won Hwang、Sangrok Lee、Yuseok Bae 和 Jongyoul Park。一种能效和 GPU 计算效率俱佳的实时目标检测骨干网络。载于 CVPRW，2019年。9、11[28] Yinhao Li、Zheng Ge、Guanyi Yu、Jinrong Yang、Zengran Wang、Yukang Shi、Jianjian Sun 和 Zeming Li。BEVDepth：为多视角三维目标检测获取可靠深度。arXiv 预印本 arXiv:2206.10092，2022年。1、2[29] Zhiqi Li、Wenhai Wang、Hongyang Li、Enze Xie、Chong- hao Sima、Tong Lu、Qiao Yu 和 Jifeng Dai。BEVFormer：通过时空变换器从多摄像头图像中学习鸟瞰视角表示。载于 ECCV，2022年。1、2[30] Ming Liang、Bin Yang、Wenyuan Zeng、Yun Chen、Rui Hu、Sergio Casas 和 Raquel Urtasun。PnPNet：将跟踪纳入循环的端到端感知与预测。载于 CVPR，2020年。3[31] Tsung-Yi Lin、Piotr Doll´ar、Ross Girshick、Kaiming He、Bharath Hariharan 和 Serge Belongie。用于目标检测的特征金字塔网络。载于 CVPR，2017年。9、11[32] Tsung-Yi Lin、Priya Goyal、Ross Girshick、Kaiming He 和 Piotr Doll´ar。用于密集目标检测的焦点损失。载于 ICCV，2017年。5[33] Yingfei Liu、Tiancai Wang、Xiangyu Zhang 和 Jian Sun。PETR：用于多视角三维目标检测的位置嵌入变换。载于 ECCV，2022年。1、2、4、6、7、9、10[34] Yingfei Liu、Junjie Yan、Fan Jia、Shuailin Li、Qi Gao、Tian- cai Wang、Xiangyu Zhang 和 Jian Sun。PETRv2：用于多摄像头图像三维感知的统一框架。arXiv 预印本 arXiv:2206.01256，2022年。2

[35] Yicheng Liu、Jinghuai Zhang、Liangji Fang、Qinhong Jiang 和 Bolei Zhou。利用堆叠式 Transformer 进行多模态运动预测。载于 CVPR，2021年。3[36] Ilya Loshchilov 和 Frank Hutter。SGDR：带热重启的随机梯度下降。arXiv 预印本 arXiv:1608.03983，2016年。10[37] Ilya Loshchilov 和 Frank Hutter。解耦权重衰减正则化。arXiv 预印本 arXiv:1711.05101，2017年。10、11[38] Jonathon Luiten、Tobias Fischer 和 Bastian Leibe。跟踪以重建，重建以跟踪。《机器人与自动化快报》，2020年。1、2[39] Wenjie Luo、Bin Yang 和 Raquel Urtasun。风驰电掣：使用单个卷积网络实现实时端到端三维检测、跟踪与运动预测。载于 CVPR，2018 年。3 、 8[40] Nicola Marinello 、 Marc Proesmans 和 Luc Van Gool。TripletTrack：使用三元组嵌入和长短期记忆网络进行三维目标跟踪。载于 CVPR，2022年。1、6[41] Tim Meinhardt、Alexander Kirillov、Laura Leal-Taixe 和 Christoph Feichtenhofer。TrackFormer：利用 Transformer 进行多目标跟踪。载于 CVPR，2022年。2、3[42] Jiquan Ngiam、Benjamin Caine、Vijay Vasudevan、Zheng- dong Zhang、Hao-Tien Lewis Chiang、Jeffrey Ling、Rebecca Roelofs、Alex Bewley、Chenxi Liu、Ashish Venugopal 等。场景 Transformer：用于行为预测与规划的统一多任务模型。载于 ICLR，2021年。3、4、5[43] Ziqi Pang、Zhichao Li 和 Naiyan Wang。SimpleTrack：理解并重新思考三维多目标跟踪。arXiv 预印本 arXiv:2111.09621，2021年。1、2、5、7、10[44] Dennis Park、Rares Ambrus、Vitor Guizilini、Jie Li 和 Adrien Gaidon。单目三维目标检测是否需要伪 LiDAR？载于 ICCV，2021年。2[45] Chu Peng、Wang Jiang、You Quanzeng、Ling Haibin 和 Liu Zicheng。TransMOT：用于多目标跟踪的时空图 Transformer。载于 CVPR，2021年。2[46] Neehar Peri、Jonathon Luiten、Mengtian Li、Aljoˇsa Oˇsep 和 Laura Leal-Taix´e。通过未来目标检测进行激光雷达预测。载于 CVPR，2022年。3[47] John Phillips、Julieta Martinez、Ioan Andrei Bˆarsan、Sergio Casas、Abbas Sadat 和 Raquel Urtasun。用于联合定位、感知与预测的深度多任务学习。载于 CVPR，2021年。3[48] Cody Reading、Ali Harakeh、Julia Chae 和 Steven L. Waslander。用于单目三维目标检测的类别深度分布网络。CVPR，2021年。2[49] Tim Salzmann、Boris Ivanovic、Punarjay Chakravarty 和 Marco Pavone。Trajectron++：利用异构数据进行动态可行的轨迹预测。载于 ECCV，2020年。3[50] Samuel Scheidegger、Joachim Benjaminsson、Emil Rosen- berg、Amrit Krishnan 和 Karl Granstr¨om。利用深度学习检测结果和 PMBM 滤波进行单目相机三维多目标跟踪。载于 IEEE智能车辆研讨会，2018年。1、2[51] Meet Shah、Zhiling Huang、Ankit Laddha、Matthew Lang- ford、Blake Barber、Sidney Zhang、Carlos Vallespi-Gonzalez，以及 Raquel Urtasun。LiRaNet：利用时空雷达融合进行端到端轨迹预测。arXiv 预印本 arXiv:2010.00731，2020年。3[52] Shaoshuai Shi、Li Jiang、Dengxin Dai 和 Bernt Schiele。结合全局意图定位与局部运动细化的运动 Transformer。发表于 NeurIPS，2022年。3[53] Colton Stearns、Davis Rempe、Jie Li、Rares Ambrus、Sergey Zakharov、Vitor Guizilini、Yanchao Yang 和 Leonidas J Guibas。SpOT：用于 3D 目标跟踪的时空建模。发表于 ECCV，2022年。2[54] Peize Sun、Jinkun Cao、Yi Jiang、Rufeng Zhang、Enze Xie、Zehuan Yuan、Changhu Wang 和 Ping Luo。TransTrack：基于 Transformer 的多目标跟踪。arXiv 预印本 arXiv:2012.15460，2020年。2[55] Pei Sun、Henrik Kretzschmar、Xerxes Dotiwalla、Aurelien Chouard、Vijaysai Patnaik、Paul Tsui、James Guo、Yin Zhou、Yuning Chai、Benjamin Caine 等。自动驾驶感知的可扩展性：Waymo开放数据集。发表于 CVPR，2020年。3[56] Siyu Tang、Mykhaylo Andriluka、Bjoern Andres 和 Bernt Schiele。利用提升多割和人员重新识别进行多人跟踪。发表于 CVPR，2017年。2[57] Pavel Tokmakov、Jie Li、Wolfram Burgard 和 Adrien Gaidon。学习利用物体恒存性进行跟踪。发表于 ICCV，2021年。2, 6[58] Ashish Vaswani、Noam Shazeer、Niki Parmar、Jakob Uszkoreit、Llion Jones、Aidan N Gomez、Łukasz Kaiser 和 Illia Polosukhin。注意力就是你所需要的一切。NeurIPS，2017年。9[59] Qitai Wang、Yuntao Chen、Ziqi Pang、Naiyan Wang 和 Zhaoxiang Zhang。Immortal Tracker：轨迹片段永不消亡。arXiv 预印本 arXiv:2111.13672，2021年。2[60] Tai Wang、Xinge Zhu、Jiangmiao Pang 和 Dahua Lin。FCOS3D：全卷积单阶段单目三维目标检测。发表于 ICCV，2021年。2[61] Yue Wang、Vitor Campagnolo Guizilini、Tianyuan Zhang、Yilun Wang、Hang Zhao 和 Justin Solomon。DETR3D：通过 3D到2D查询从多视角图像进行三维目标检测。发表于 CoRL，2022年。1, 2, 9, 10, 11[62] Xinshuo Weng、Boris Ivanovic、Kris Kitani 和 Marco Pavone。这到底是谁的轨迹？利用基于亲和度的轨迹预测提升对跟踪误差的鲁棒性。发表于 CVPR，2022年。3[63] Xinshuo Weng、Boris Ivanovic 和 Marco Pavone。MTP：通过多假设跟踪与预测减少误差传播。发表于 IEEE智能车辆研讨会，2022年。3, 8[64] Xinshuo Weng、Jianren Wang、David Held 和 Kris Kitani。三维多目标跟踪：基线与新的评估指标。发表于 IROS，2020年。2, 6, 7, 10[65] Xinshuo Weng、Jianren Wang、Sergey Levine、Kris Kitani 和 Nicholas Rhinehart。利用 SPF2 反转姿态预测流程：用于序列姿态预测的序列点云预测。发表于 CoRL，2021年。3[66] Xinshuo Weng、Yongxin Wang、Yunze Man 和 Kris M Kitani。GNN3DMOT：结合 2D-3D多特征学习的三维多目标跟踪图神经网络。发表于 CVPR，2020年。1, 2

[67] Benjamin Wilson、William Qi、Tanmay Agarwal、John Lambert、Jagjeet Singh、Siddhesh Khandelwal、Bowen Pan、Ratnesh Kumar、Andrew Hartnett、Jhony Kaesemodel Pontes 等。Argoverse 2：面向自动驾驶感知与预测的下一代数据集。发表于 NeurIPS，2022年。1, 3[68] Nicolai Wojke、Alex Bewley 和 Dietrich Paulus。利用深度关联度量进行简单的在线实时跟踪。发表于 ICIP，2017年。2[69] Tianwei Yin、Xingyi Zhou 和 Philipp Krahenbuhl。基于中心的三维目标检测与跟踪。发表于 CVPR，2021年。2, 7, 10[70] Ye Yuan、Xinshuo Weng、Yanglan Ou 和 Kris M Kitani。AgentFormer：面向社会时间多智能体预测的智能体感知 Transformer。发表于 ICCV，2021年。3[71] Jan-Nico Zaech、Alexander Liniger、Dengxin Dai、Martin Danelljan 和 Luc Van Gool。用于三维多目标跟踪的可学习在线图表示。Robotics and Automation Letters，2022年。1, 2[72] Fangao Zeng、Bin Dong、Tiancai Wang、Xiangyu Zhang 和 Yichen Wei。MOTR：基于 Transformer 的端到端多目标跟踪。发表于 ECCV，2022年。2, 3[73] Tianyuan Zhang、Xuanyao Chen、Yue Wang、Yilun Wang 和 Hang Zhao。MUTR3D：一种通过 3D到2D查询实现的多摄像头跟踪框架。发表于 CVPRW，2022年。1, 2, 3, 6, 9[74] Yifu Zhang、Chunyu Wang、Xinggang Wang、Wenjun Zeng 和 Wenyu Liu。FairMOT：多目标跟踪中检测与重新识别的公平性。发表于《国际计算机视觉杂志》，2021年。2[75] Yunpeng Zhang、Zheng Zhu、Wenzhao Zheng、Junjie Huang、Guan Huang、Jie Zhou 和 Jiwen Lu。BEVerse：面向视觉中心型自动驾驶的鸟瞰视图统一感知与预测。arXiv 预印本 arXiv:2205 09743，2022年 3 .[76] Xingyi Zhou、Vladlen Koltun 和 Philipp Krahenbuhl。将目标作为点进行跟踪。发表于 ECCV，2020年。2, 6[77] Xizhou Zhu、Weijie Su、Lewei Lu、Bin Li、Xiaogang Wang 和 Jifeng Dai。Deformable DETR：用于端到端目标检测的可变形 Transformer。发表于 ICLR，2021年。3