

基于目标检测和场景流估计联合优化的 3D 多目标跟踪

王光明¹, 宋 亮², 沈玥伶³, 王贺升³

(1. 剑桥大学工程系, 英国 剑桥 CB2 1PZ; 2. 中国矿业大学信息与控制工程学院, 江苏 徐州 221116;

3. 上海交通大学电子信息与电气工程学院, 上海 200240)

摘要: 大多数 3D 多目标跟踪方法独立优化目标检测和帧间数据关联部分, 没有考虑单帧的特征学习和帧间关联学习的耦合性。为了实现单帧检测和帧间关联的耦合学习, 提出了一种基于目标检测和场景流估计联合优化的 3D 多目标跟踪框架 FlowDet-Track。在该框架中, 提出了一个检测引导场景流估计模块来缓解不正确的帧间关联。为了获得更准确的场景流标签, 特别是在旋转运动的情况下, 提出了一种基于框变换的场景流真值计算方法。在 KITTI MOT 数据集上的实验结果表明, 本文算法的车辆类别 HOTA 与 DetA 指标比 PointTrackNet 算法提升了 25.03% 和 30.8%, 表明本文算法的位置跟踪精度优异; 此外, 极端旋转运动条件下的对比实验进一步证明了算法的鲁棒性。

关键词: 多目标跟踪; 场景流; 目标检测

3D Multi-object Tracking Based on Simultaneous Optimization of Object Detection and Scene Flow Estimation

WANG Guangming¹, SONG Liang², SHEN Yueling³, WANG Hesheng³

(1. Department of Engineering, University of Cambridge, Cambridge CB2 1PZ, UK;

2. School of Information and Control Engineering, China University of Mining and Technology, Xuzhou 221116, China;

3. School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China)

Abstract: Most 3D multi-object tracking methods independently optimize target detection and inter-frame data association, without considering the coupling between single-frame feature learning and inter-frame association learning. To achieve the coupled learning of single-frame detection and inter-frame association, a 3D multi-object tracking framework is proposed based on the joint optimization of target detection and scene flow estimation, named FlowDet-Track. In this framework, a detection-guided scene flow estimation module is introduced to alleviate incorrect inter-frame association. To obtain more accurate scene flow labels, especially in cases of rotational motion, a box transformation-based ground truth calculation method is proposed for scene flow. The experimental results on the KITTI MOT dataset indicate that the HOTA and DetA metrics for vehicle category is improved by 25.03% and 30.8% compared with PointTrackNet algorithm, demonstrating the superior performance of the proposed method in position tracking. Moreover, comparative experiments under extreme rotational motion further validate the algorithm robustness.

Keywords: multi-object tracking; scene flow; object detection

3D 多目标跟踪是自动驾驶所需的重要感知技术, 是下游规划、导航任务的基础^[1-2]。3D 多目标跟踪技术将连续帧的 3D 点云序列作为输入, 并输出特定类别目标的轨迹。目标的轨迹由连续帧中带有跟踪标识 (ID) 的 3D 框表示。同一目标在不同帧中通过相应的唯一 ID 进行识别。

大多数方法^[1-6]都遵循先检测后跟踪的框架, 将任务分为 2 个阶段, 目标检测阶段和数据关联阶段。通过目标检测结果, 数据关联模块计算相邻帧

的每对目标之间的相似性, 并使用唯一的 ID 来识别相同的目标。由于先检测后跟踪的框架针对检测和数据关联 2 个部分进行独立优化, 没有考虑两者之间的耦合性, 因此限制了跟踪精度的提高。

传统的数据关联方法^[1-3]中用运动预测方法 (例如卡尔曼滤波器或粒子滤波器) 将相邻帧中的框关联起来。这些基于滤波器的方法在急刹车或转弯等极端运动条件下容易失效。此外, 由于不同类型的目标可能具有不同的运动模型, 因此这些方法

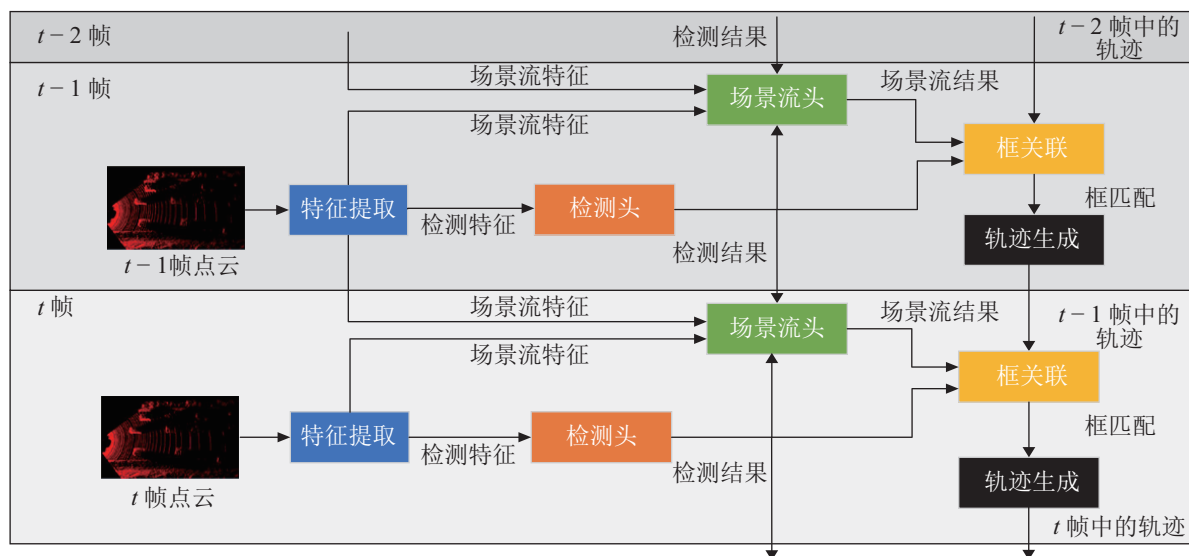


图 1 目标检测和场景流估计联合优化的 3D 多目标跟踪方法整体框架

Fig.1 Overall framework of 3D multi-object tracking method based on the joint optimization of object detection and scene flow estimation

中的固定运动参数不适用于多类别目标跟踪任务。

光流估计方法^[7-8]和场景流估计方法^[9-11]不以目标检测为基础,估计帧间每个像素或每个 3D 点的运动,而多目标跟踪任务估计的是物体的帧间运动而不是点的帧间运动。考虑到场景流估计与多目标跟踪任务之间的相关性,一些研究^[4,12-13]通过场景流估计方法获得逐点的运动估计后,用逐点的帧间运动来估计物体的帧间运动。FlowMOT 方法^[4]直接用训练好的场景流模型代替运动估计模型,提高了极端运动下的鲁棒性。然而,它仍然针对检测和场景流估计进行独立优化。PointTrackNet 方法^[13]基于场景流实现了端到端的 3D 目标检测和跟踪。该方法首先优化目标检测,然后基于场景流估计结果进行数据关联。然而,它没有考虑不同目标之间的错误关联问题。此外,在目标旋转的情况下,该方法的场景流标签也不够准确,这些都限制了多目标跟踪结果的精度。

为解决上述问题,本文提出了一种基于 3D 目标检测和场景流估计联合优化的 3D 多目标跟踪框架,以考虑 2 个任务之间的耦合关系,提高多目标跟踪的精度,并提高在极端旋转运动下的鲁棒性。

1 提出的方法(The proposed method)

1.1 总体框架

与独立优化 3D 目标检测和场景流估计不同,本文提出了一种联合优化 3D 目标检测和场景流估计的 3D 多目标跟踪框架。该框架以点云序列为输入,并输出物体的轨迹。

如图 1 所示,输入连续帧点云后,特征提取模块同时逐点提取检测特征和场景流特征。在检测结果的指导下,将相邻帧中的场景流特征输入到场景流头模块,来估计帧间各点的场景流结果。通过帧间逐点的运动估计,在框关联模块中估计帧间框的运动。帧间框的运动估计用于关联帧间的 3D 检测框,并最终生成轨迹。

通过场景流头模块获得了相邻帧之间的逐点数据关联,可以避免对运动模型参数的依赖。为了简化网络结构,在特征提取模块中同时提取目标检测特征和场景流特征。在训练阶段,目标检测和场景流的损失将被加权求和。总损失将用于优化整个网络。考虑到检测结果包含点的语义信息,将检测结果输入场景流模块作为指导,以避免属于不同目标点的数据错误地关联。为了实现多目标跟踪,根据框关联模块中的目标检测结果和场景流结果来估计帧之间的框运动,根据运动后的框之间的距离进行框匹配,轨迹生成模块将框匹配作为输入并更新轨迹。

1.2 特征提取

为了同时逐点提取点云的检测特征和场景流特征,特征提取模块采用了 PV-RCNN^[14]的骨干网络。如图 2 所示,通过体素特征提取模块从多个分辨率层面提取了体素特征。然后,通过垂直投影从特征体素的最后一层获取顶视鸟瞰的特征图。同时,通过最远点采样对点云进行下采样,并提取关键点。在关键点特征提取阶段,多种分辨率的体素特征被汇总为一小组关键点的特征嵌入。2 个并行

的多层感知机 (MLP) 以所嵌入的关键点特征为输入, 同时输出检测特征和场景流特征。

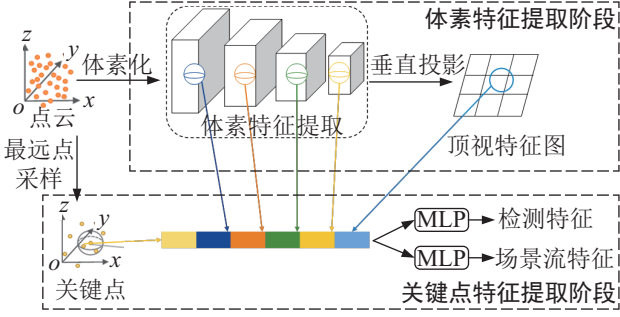


图2 特征提取框架

Fig.2 The framework of feature extraction

1.3 检测头模块

为了获得准确的检测结果, 采用了与 PV-RCNN 网络结构^[14] 相同的两阶段检测头模块。在第 1 阶段, 从顶部特征图中生成 3D 候选目标框。在候选框优化阶段, 通过内部点的检测特征对 3D 候选目标框进行优化。对于 3D 候选目标框 i , 检测头的输出包括框结果 $\mathbf{b}_i = (p_{xi}, p_{yi}, p_{zi}, w_i, l_i, h_i, \theta_i)$ 和该框属于每个类别目标的概率 s_i 。其中 (p_{xi}, p_{yi}, p_{zi}) 表示 3D 检测框 i 的中心位置坐标; w_i, l_i, h_i 表示 3D 检测框 i 的 3 维尺寸信息, 宽、长、高; θ_i 表示 3D 目标检测框 i 在 z 轴上旋转的角度。

1.4 场景流头模块

场景流头模块包括成本量体积层和场景流计算层。成本量体积层实现了 2 个点云之间的特征关联, 并输出成本量体积 (cost volume)。场景流计算层以成本量体积为输入, 为 $t-1$ 帧中的每个点输出运动到 t 帧的场景流结果。

1) 成本量体积层

成本量体积层的输入包括 2 个相邻帧的点云 $\mathbf{P}^{t-1} = \{p_i^{t-1}(x_i^{t-1}, y_i^{t-1}, z_i^{t-1})\}_{i=1}^N$ 和 $\mathbf{P}^t = \{p_i^t(x_i^t, y_i^t, z_i^t)\}_{i=1}^N$, 对应的场景流特征 $\mathbf{F}_{\text{flow}}^{t-1} = \{\mathbf{f}_i^{t-1}\}_{i=1}^N$ 和 $\mathbf{F}_{\text{flow}}^t = \{\mathbf{f}_i^t\}_{i=1}^N$, 对应的检测结果 $\mathbf{O}_{\text{det}}^{t-1} = \{\mathbf{o}_i^{t-1}\}_{i=1}^N$ 和 $\mathbf{O}_{\text{det}}^t = \{\mathbf{o}_i^t\}_{i=1}^N$, 其中 N 是每帧关键点的数量, t 是时间戳。场景流特征向量 $\mathbf{f}_i^t \in \mathbf{R}^{C_{\text{flow}}}$, C_{flow} 是场景流特征的维度。检测结果向量 $\mathbf{o}_i^t$ 通过在检测结果中拼接目标框向量 $\mathbf{b}_i^t$ 和得分 s_i^t 获得。成本量体积层输出与 $t-1$ 帧中的点 p_i^{t-1} 相对应的成本量体积 c_i^t 。

图 3 是成本量体积层的可视化结果。对于点 p_i^{t-1} , 在 t 帧中找到它的 K 个最近邻点, 表示为邻居点集合 $\Omega_i^{t-1} = \{p_{i,j}^t\}_{j=1}^K \subset \mathbf{P}^t$ 。成本量体积层通过关联中心点 p_i^{t-1} 与邻居点 Ω_i^{t-1} 的特征来学习帧间的匹配成本量体积。之前的工作^[13] 将方向向量 $\mathbf{p}_i^{t-1} - \mathbf{p}_{i,j}^t$ 、邻居点的场景流特征 $\mathbf{f}_{i,j}^t$ 以及中心点的

流特征 $\mathbf{f}_i^{t-1}$ 拼接为嵌入特征, 然后应用多层感知机和最大池化层, 计算成本量体积。然而, 这种方法在聚合过程中考虑了所有邻近点。为了获得准确的运动估计, 只应该包括属于与点 p_i^{t-1} 相同目标的邻近点。如图 3 所示, 黑色虚线是指 $t-1$ 帧中的物体轮廓, 黑色实线是指 t 帧中的物体轮廓。蓝色点指的是 $t-1$ 帧中的聚合中心点, 红色点是指 t 帧中的点。图中取 $t-1$ 帧中车上的一个蓝色点, 画了一个特征聚合的样例。从样例中可以看出, 蓝色中心点属于 $t-1$ 帧中的汽车, 尽管邻近范围内可以聚合到 t 帧的汽车上的红色点, 但是会聚合属于不同目标上的点 (树木上的点), 不同目标上的点 (树木上的点) 会干扰目标 (汽车) 的局部特征匹配, 从而影响运动的估计。因此, 本文提出了检测引导场景流模块, 来解决属于不同目标的点之间不正确的帧间关联的问题。

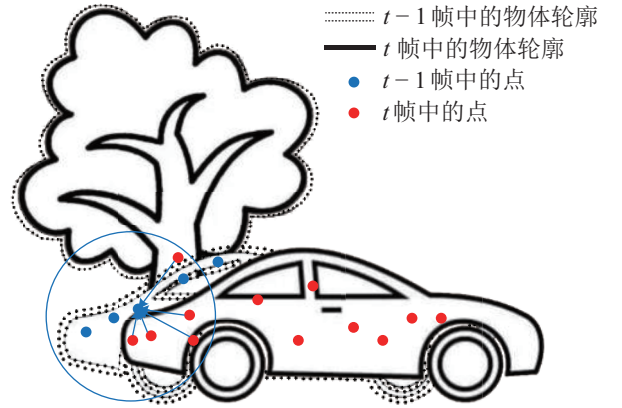


图3 成本量体积层可视化

Fig.3 The visualization of cost volume layer

具体地, 之前的场景流估计网络^[9] 以点云的场景流特征和相对坐标作为输入, 点 p_i^{t-1} 与其邻近点 $p_{i,j}^t$ 之间的成本量体积 C 按式 (1) 计算:

$$C(i, j) = h(\mathbf{f}_i^{t-1}, \mathbf{f}_{i,j}^t, \mathbf{p}_i^{t-1} - \mathbf{p}_{i,j}^t) \quad (1)$$

其中 $h(\cdot)$ 是带有可学习参数的多层感知机。接下来, 通过对点 p_i^{t-1} 的所有邻近点 $p_{i,j}^t$ ($p_{i,j}^t \in N_i^t$) 的成本量体积进行加权求和计算, 获得点 p_i^{t-1} 最终的成本量体积 c_i^{t-1} :

$$c_i^{t-1} = \sum_{p_{i,j}^t \in N_i^t} w(i, j) \times h(\mathbf{f}_i^{t-1}, \mathbf{f}_{i,j}^t, \mathbf{p}_i^{t-1} - \mathbf{p}_{i,j}^t) \quad (2)$$

其中 $w(i, j)$ 表示点 p_i^{t-1} 和点 $p_{i,j}^t$ 的匹配权重。它作为方向向量 $\mathbf{p}_i^{t-1} - \mathbf{p}_{i,j}^t$ 的连续函数, 通过多层感知机 G 进行学习。权重 $w(i, j)$ 的计算如下:

$$w(i, j) = G(\mathbf{p}_i^{t-1} - \mathbf{p}_{i,j}^t) \quad (3)$$

这些权重计算方法只考虑了方向向量作为输入。没有考虑到邻居点可能来自不同物体, 不应该参与帧间匹配的成本量体积计算。由于检测结果包含点所属的目标信息, 因此, 自然地, 可以在特征聚合中使用检测结果作为引导, 过滤离群点, 从而提升帧间匹配的精度。本文提出的权重计算方式如下:

$$w_g(i, j) = \frac{1}{\|p_i^{t-1} - p_{i,j}^t\|_2} \quad (4)$$

$$w_o(i, j) = \frac{1}{\|o_i^{t-1} - o_{i,j}^t\|_2} \quad (5)$$

$$w(i, j) = M \left(\frac{w_g(i, j)}{\sum_{j=1}^K w_g(i, j)}, \frac{w_o(i, j)}{\sum_{j=1}^K w_o(i, j)} \right) \quad (6)$$

其中 $w_g(i, j)$ 表示几何空间中距离的倒数, $w_o(i, j)$ 表示检测结果空间中距离的倒数, $M(x, y)$ 表示 x 和 y 的平均值, 权重 $w(i, j)$ 由标准化后的 $w_g(i, j)$ 和 $w_o(i, j)$ 的平均值计算而得。综上, 只有附近属于同一目标的点被赋予更大的权重。将式 (6) 代入式 (2) 中, 成本量体积 c_i^{t-1} 就可以包含点的类别信息, 通过在帧间匹配时考虑类别信息, 提升帧间匹配和多目标跟踪的精度。消融实验将证明这一点。

2) 场景流计算层

场景流计算层的输入包括成本量体积集合 $C^{t-1} = \{c_i^{t-1}\}_{i=1}^N$, 第 1 个点云的场景流特征集合 F_{flow}^{t-1} , 以及第 1 个点云 P^{t-1} 。首先, 对成本量体积和场景流特征进行逐点拼接。然后, 应用 PointConv 层^[15]在局部邻域合并拼接的特征, 随后通过 MLP 估计对应于第 1 个点云 P^{t-1} 到第 2 个点云 P^t 运动的逐点场景流 $D^{t-1,t}$, 其中每个点的估计场景流表示为 $\hat{d}_i$ 。

3) 场景流标签计算

在实际应用中往往很难获得场景流的真实标签。为了解决这个问题, 本文提出了一种基于 MOT 标注的场景流标签计算算法。MOT 真值标签提供检测结果以及每帧中目标的跟踪 ID。同一物体的跟踪 ID 在序列中保持不变。这样就可以根据跟踪 ID 匹配相邻两帧的检测结果, 从而计算出物体的帧间相对位姿变换。属于目标的点的位姿变换等同于目标的刚体位姿变换。

PointTrackNet 模型^[13]使用两帧之间边界框的 3 维移动向量作为场景流的标签。当目标在帧间旋转时, 这种基于框平移的方法不够准确。本文提出了一种基于框变换的方法来获得更准确的场景流真值标签。对于属于框 m 的点 p_i^{t-1} , 本文使用框变换 T_m 对其进行转换, 并获得点 $\tilde{p}_i^{t-1}$ 。最后, 通过框变

换后的点坐标减去框变换前的原坐标获得场景流标签 $\tilde{d}_i^{t-1}$, 如下:

$$\tilde{d}_i^{t-1} = \tilde{p}_i^{t-1} - p_i^{t-1} \quad (7)$$

1.5 框关联模块

框关联模块借助场景流结果中的帧间点的运动信息, 实现目标的帧间运动估计。首先, 利用其内部点的场景流结果来计算轨迹中每个框的帧间变换矩阵。然后, 通过计算得到的变换矩阵对前一帧轨迹框进行变换, 得到预测轨迹框。通过计算预测轨迹框与检测框之间的交并比, 计算匹配成本矩阵。最后, 利用匈牙利算法^[16]来匹配成本矩阵, 输出预测轨迹框与检测框之间的匹配关系。

框关联的关键是根据场景流中点的帧间运动结果实现帧间框的运动估计。框关联的过程是场景流真值标签计算过程的逆过程。与场景流标签计算相对应, 通过基于点云的配准解决了框的帧间转换问题。具体而言, 对于每个轨迹框 b_i^{t-1} , 提取其内部点并形成框内点云 c_i^{t-1} 。对于 c_i^{t-1} 中的每个点, 加上其估计的场景流平移向量, 得到预测的框内点云 $\tilde{c}_i^{t-1}$ 。然后, 基于奇异值分解 (SVD) 方法计算 c_i^{t-1} 与 $\tilde{c}_i^{t-1}$ 之间的变换矩阵 $T_i^{t-1,t}$ 。将变换矩阵 $T_i^{t-1,t}$ 应用于框 b_i^{t-1} , 得到预测的运动到第 2 帧的框 $\tilde{b}_i^{t-1}$ 。

1.6 轨迹生成模块

将框关联模块得到的匹配结果输入到轨迹生成模块中, 以更新轨迹。匹配结果由 3 个列表表示, 匹配列表 $I_m^{t-1,t}$ 、检测框不匹配列表 $I_{\text{um(d)}}^{t-1,t}$ 和轨迹框不匹配列表 $I_{\text{um(t)}}^{t-1,t}$ 。匹配列表 $I_m^{t-1,t}$ 由检测框和轨迹框的匹配对组成。不匹配列表 $I_{\text{um(d)}}^{t-1,t}$ 和 $I_{\text{um(t)}}^{t-1,t}$ 分别由不匹配的检测框和预测轨迹框组成。为了应对遮挡情况下产生的误检测, 轨迹生成模块根据检测框匹配和丢失的次数来添加和删除跟踪框。

1.7 网络训练的损失函数

网络训练的损失函数由 2 部分组成, 3D 目标检测损失和场景流损失, 可写为式 (8) 的形式:

$$L = L_{\text{det}} + L_{\text{scene flow}} \quad (8)$$

在 PV-RCNN 网络结构^[12]中, 检测损失可以分为区域候选损失 L_{rpn} 、关键点分割损失 L_{seg} 和候选优化损失 L_{rcnn} 。在本文的工作中, 由于固定了区域候选框 (RPN) 的参数, 因此检测损失的计算如式 (9) 所示:

$$L_{\text{det}} = L_{\text{seg}} + L_{\text{rcnn}} \quad (9)$$

对于场景流损失的计算, 本文根据 MOT 真值标签为每个点计算了其是否属于框内的掩码 M_i^{inbox} 。

场景流损失只考虑属于物体框内的点，通过均方根误差（RMSE）进行计算，如式 (10) 所示：

$$L_{\text{scene flow}} = \sqrt{\frac{\sum_{i=1}^n M_i^{\text{inbox}} \times \|d_i - \hat{d}_i\|_2^2}{N_{\text{inbox}}}} \quad (10)$$

其中， d_i 和 $\hat{d}_i$ 分别表示点 p_i^{t-1} 的场景流估计结果和场景流标签， $\|\cdot\|_2$ 代表求二范数。 N_{inbox} 是位于物体框内的点的数量。

2 实验结果与讨论（Experimental results and discussions）

2.1 数据集和评估方法

本文在 KITTI 的 MOT 数据集^[17]上训练和评估了所提出的 3D 多目标跟踪方法。KITTI 的 MOT 数据集包括 21 个训练序列和 29 个测试序列。每个序列提供激光雷达收集的 3D 点云序列和彩色单目相机收集的图像序列。将训练序列分成 2 部分：其中 Seq-0000 至 Seq-0015 用于训练，Seq-0016 至 Seq-0020 用于验证。

为了提高样本多样性并避免过拟合，引入了一种用于 3D 目标检测和场景流估计的数据增强方法。首先，截取所有真值目标检测框的标签和框内点云，使用 KITTI 目标检测数据集^[17]生成一个目标检测框数据库。然后在训练过程中，从目标检测框数据库中随机选择几个真值目标检测框，并将它们随机放置在 2 个相邻帧的场景中的相同位置，都进行随机比例缩放。通过这种方法，极大地增加了目标检测框的数量并提高了目标检测框位置的多样性。之后，对放置于第 2 帧的目标检测框进行随机角度旋转和随机平移。为了增加帧间目标运动的多样性，对每个目标独立地进行随机数据增强。

本文采用 HOTA 指标^[18]作为评估指标。HOTA 指标可以分为检测精度得分（DetA）、关联精度得分（AssA）和定位精度得分（LocA）。

2.2 网络细节

在训练过程中，固定 PV-RCNN 骨干网络的区

域候选框（RPN）参数，同步优化目标检测网络和场景流估计网络的参数。训练时使用了一个带有周期变化学习率的 Adam 优化器^[19]。学习率按照余弦函数变化，上限为 0.01。

实验使用 4 个 GeForce GTX 1080Ti GPU 进行训练，批量大小为 4。实验的深度学习环境是 CUDA 10.0，深度学习代码使用 PyTorch 实现。

2.3 定量分析

在 KITTI MOT 测试数据集上测试了训练好的模型，并将测试结果提交到 KITTI 数据集的官方网站进行精度评估。本文提出的算法被称为 FlowDet-Track，将其结果与点云输入的先检测后跟踪的在线多目标跟踪算法进行了比较。对比的方法中，AB3DMOT^[2]和 Complexer-YOLO^[20]是以点云为输入的主流算法；PointTrackNet^[13]模型是一个端到端的 3D 目标检测和跟踪网络。该方法首先优化目标检测，然后基于场景流估计结果进行数据关联。该研究在“使用场景流估计的结果进行数据关联”的这点上，和本文的思路最为接近。然而，它没有考虑不同目标之间的错误关联问题，在目标旋转的情况下，场景流标签也不够准确，这些都限制了 PointTrackNet 方法^[13]多目标跟踪结果的精度。表 1 显示了精度评估的结果（粗体字表示最佳结果）。本文方法的 HOTA 指标远远超过了其他使用点云输入的先检测后跟踪的方法，显示了本文提出的目标检测和场景流帧间关联联合学习方法的优越性。

此外，在困难数据集上测试了极端旋转运动下的鲁棒性。通过向 KITTI 的 MOT 验证集上的每个目标添加随机平移（范围为 0~0.5 m）以及绕 z 轴随机旋转（范围为 0~20°）获得了困难数据集。表 2 显示了汽车类别和行人类别的结果。与其他使用 3D 多目标先检测后跟踪的 3D 多目标跟踪方法相比，本文方法在困难数据集上仍然能够保持较高的性能。这是由于本文通过框变换的场景流标签计算方法，准确地考虑了物体的旋转和平移运动对每个

表 1 KITTI MOT 测试集上的评估结果
Tab.1 Evaluation results on KITTI MOT testsets

方法	车辆类别				行人类别			
	HOTA↑	DetA↑	AssA↑	LocA↑	HOTA↑	DetA↑	AssA↑	LocA↑
Complexer-YOLO ^[20]	49.12%	62.44%	39.34%	81.47%	14.08%	24.91%	8.15%	68.64%
AB3DMOT ^[2]	68.61%	71.06%	69.06%	86.80%	35.57%	32.99%	35.58%	71.26%
PointTrackNet ^[13]	57.20%	55.71%	59.15%	80.07%				
FlowDet-Track（本文）	71.52%	72.87%	70.89%	87.79%	39.64%	40.90%	38.72%	72.04%

注：对于表中的所有指标，带有↑的代表“该指标值越大，结果越理想”。

表 2 困难数据集上的比较评估结果
Tab.2 Comparative evaluation results on hard datasets

方法	车辆类别				行人类别			
	HOTA↑	DetA↑	AssA↑	LocA↑	HOTA↑	DetA↑	AssA↑	LocA↑
AB3DMOT ^[2]	43.34%	58.07%	32.51%	88.08%	15.14%	23.02%	10.08%	80.20%
FlowDet-Track (本文)	64.64%	67.00%	62.54%	88.59%	32.21%	46.82%	22.28%	83.88%

点的运动产生的影响,使得场景流网络训练能够准确地被真值监督,进行合理的网络优化训练,从而提升了网络训练的精度和鲁棒性。

相比于其他对比方法,本文方法的精度最高。在效率方面,本文方法相比于其他方法低一些,但仍达到了 10.6 Hz,与激光雷达常见的扫描频率 10 Hz 相比,实时性得到了有效保证。具体时间消耗如表 3 所示。

表 3 各种方法的时间耗费比较
Tab.3 Time consumption comparison of various methods

方法	推理速度 /(帧/秒)
Complexer-YOLO ^[20]	100
AB3DMOT ^[2]	207.4
Point-TrackNet ^[13]	10.1
FlowDet-Track (本文)	10.6

2.4 消融实验

设计了一项消融实验来验证所提出的数据增强模块并验证了在场景流模块中使用检测结果进行引导的有效性。结果如表 4 所示,在汽车和行人两个类别中,所提出的数据增强模块和检测结果引导的场景流模块有效提高了多目标跟踪的精度。

表 4 数据增强和检测结果引导的场景流估计消融实验结果
Tab.4 Ablation experiment results for data augmentation and detection guided scene flow estimation

方法		HOTA↑	
数据增强	检测结果引导	车辆类别	行人类别
		75.89%	42.77%
✓		77.43%	46.42%
✓	✓	78.80%	78.80%

2.5 定性分析

图 4 展示了 FlowDet-Track 算法在 MOT 数据集上的可视化结果。轨迹框被投影到图像上。具有相同跟踪 ID 的框被涂上相同的颜色。可以看到,本文方法正确地跟踪了相关类别。

图 5 展示了在场景流模块的成本层中进行的特征聚合的可视化结果。白色点是指 $t-1$ 帧的特征提取关键点,作为帧间聚合的中心点;黄色到褐色

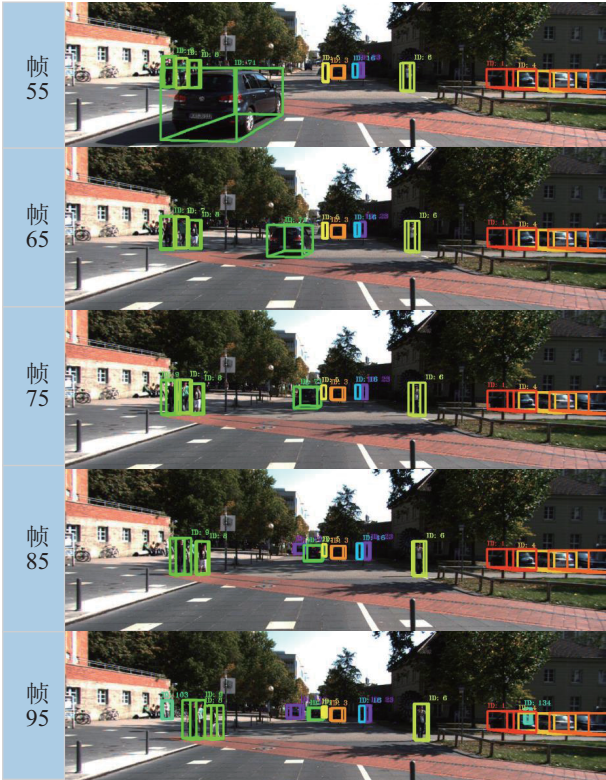


图 4 本文方法的 MOT 结果可视化
Fig.4 The visualization of MOT results using the proposed method

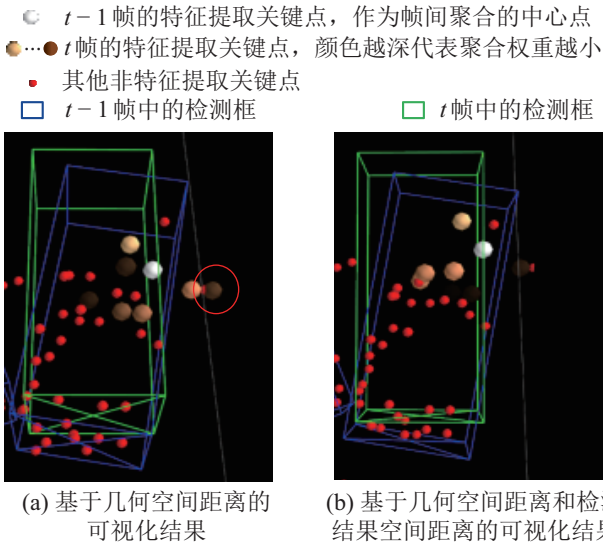
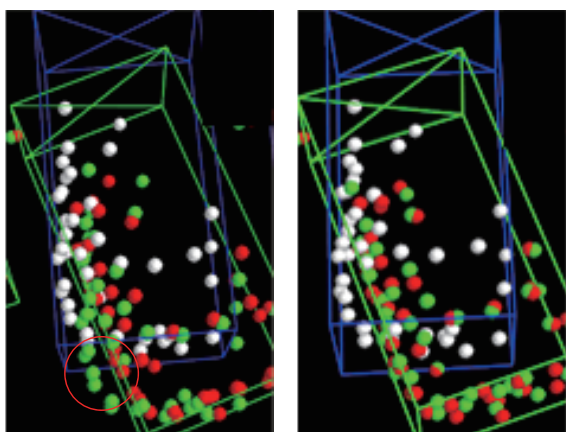


图 5 聚合权重计算的可视化结果
Fig.5 The visualization results of the aggregation weight calculation

具有渐变色的大圆点,是指 t 帧的特征提取关键点,其颜色越深代表聚合权重越小;小的红色点是非特征提取关键点,其特征已经被聚合到特征提取关键点中,这里画出仅是为了展示原始的输入点云数据。蓝色框是 $t-1$ 帧中的检测框;绿色框是 t 帧中的检测框,图中以 $t-1$ 帧的检测框中的一个白色点,画了一个特征聚合的样例。从样例中可以看出,当只考虑几何空间中的距离时,那些离聚合中心很近但并不属于目标的点将被赋予较大的聚合权重,如图5(a)中的红色圈所示。如图5(b)所示,使用所提出的检测引导特征聚合方法,这些背景点的聚合权重将小到足以被忽略。这体现了本文提出的检测引导场景流模块有效降低了属于不同目标的点对帧间关联的干扰。

图6显示了目标发生旋转运动时场景流标签计算方法的可视化结果对比。白色点是指 $t-1$ 帧的点,绿色点是 $t-1$ 帧的白色点通过场景流转换到 t 帧的点,红色点是 t 帧的点。蓝色框是 $t-1$ 帧中的检测框;绿色框是 t 帧中的检测框。理论上, $t-1$ 帧的点通过场景流转换到 t 帧后,在整体结构上应当与 t 帧的点云一致。图6(a)为使用 PointTrack-Net^[13]中基于框平移的方法的结果,红色圆圈所示的变换后的绿色点与 t 帧中的红色点有部分区域不能匹配;而图6(b)使用本文提出的基于框变换的方法,变换后的绿色点能够很好地匹配 t 帧中的红色点。这证明了所提出的基于框变换的场景流标签计算方法可以在目标旋转的情况下获得更准确的场景流真值。

● $t-1$ 帧的点 ● t 帧的点
● $t-1$ 帧的点通过场景流转换到 t 帧的点
□ $t-1$ 帧中的检测框 □ t 帧中的检测框



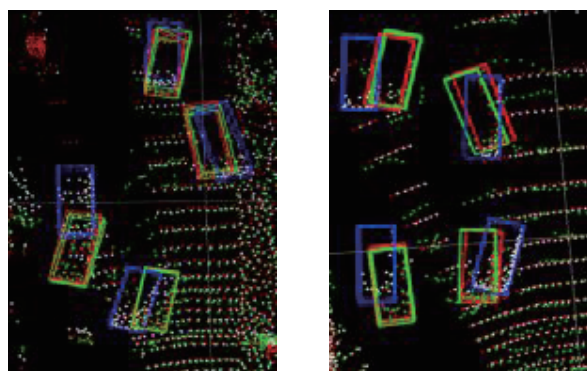
(a) 基于框平移的方法 (b) 基于框变换的方法

图6 场景流标签计算的可视化

Fig.6 The visualization results of the scene flow label calculation

图7展示了在困难数据集上测试鲁棒性时的定性结果。其中蓝色框是 $t-1$ 帧的检测框;本文通过场景流计算 $t-1$ 帧检测框的整体刚性运动,绿色框是通过该刚性运动对 $t-1$ 帧的检测框进行刚性变换得到的第 t 帧的检测框;红色框是真实的第 t 帧点云中的目标检测框。可以看出,合成的绿色框和红色框基本重合,这说明本文提出的以合成框和真实框的交并比来进行后续跟踪的方法足够有效。

● $t-1$ 帧的点 ● t 帧的点
● $t-1$ 帧的点通过场景流转换到 t 帧的点
□ $t-1$ 帧中的检测框 □ t 帧中的检测框
□ $t-1$ 帧检测框经过场景流估计得到的 t 帧的预测框



(a) 18 序列 0161、0162 帧
场景流估计与帧关联
可视化结果

(b) 20 序列 0000、0001 帧
场景流估计与帧关联
可视化结果

图7 困难数据集上场景流估计和框关联可视化结果

Fig.7 The visualization results of scene flow estimation and box association on hard datasets

3 结论 (Conclusion)

提出了一个联合优化 3D 目标检测和场景流估计的 3D 多目标跟踪框架。本文提出的检测引导场景流估计的方法探索了 3D 目标检测在促进场景流估计中的作用。提出的基于框变换的场景流标签计算方法,在目标旋转的情况下获得了更准确的场景流标签。实验结果表明,在极端旋转运动下,本文方法与其他方法相比仍然表现出具有竞争力的结果和鲁棒性。

参考文献 (References)

- [1] GUO G, ZHAO S J. 3D multi-object tracking with adaptive cubature Kalman filter for autonomous driving[J]. IEEE Transactions on Intelligent Vehicles, 2023, 8(1): 512-519.
- [2] WENG X S, WANG J R, HELD D, et al. 3D multi-object tracking: A baseline and new evaluation metrics[C]//IEEE/RSJ International Conference on Intelligent Robots and Systems. Piscataway, USA: IEEE, 2020: 10359-10366.
- [3] SHENOI A, PATEL M, GWAK J Y, et al. JRMOT: A real-time 3D multi-object tracker and a new large-scale dataset[C]//IEEE/RSJ International Conference on Intelligent Robots and Systems. Piscataway, USA: IEEE, 2020: 10335-10342.

- [4] ZHAI G Y, KONG X, CUI J H, et al. FlowMOT: 3D multi-object tracking by scene flow association[DB/OL]. (2021-03-05) [2023-01-05]. <https://arxiv.org/abs/2012.07541>.
- [5] WENG X S, WANG Y X, MAN Y Z, et al. GNN3DMOT: Graph neural network for 3D multi-object tracking with 2D-3D multi-feature learning[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, USA: IEEE, 2020: 6498-6507.
- [6] WANG G M, PENG C S, GU Y Y, et al. Interactive multi-scale fusion of 2D and 3D features for multi-object vehicle tracking [J]. IEEE Transactions on Intelligent Transportation Systems, 2023, 24(10): 10618-10627.
- [7] DOSOVITSKIY A, FISCHER P, ILG E, et al. FlowNet: Learning optical flow with convolutional networks[C]//IEEE International Conference on Computer Vision. Piscataway, USA: IEEE, 2015: 2758-2766.
- [8] SUN D Q, YANG X D, LIU M Y, et al. PWC-NET: CNNs for optical flow using pyramid, warping, and cost volume [C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, USA: IEEE, 2018: 8934-8943.
- [9] WU W X, WANG Z Y, LI Z W, et al. PointPWC-NET: A coarse-to-fine network for supervised and self-supervised scene flow estimation on 3D point clouds[DB/OL]. (2020-07-31) [2023-09-03]. <https://arxiv.org/abs/1911.12408>.
- [10] LIU X Y, QI C R, GUIBAS L J. FlowNet3D: Learning scene flow in 3D point clouds[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, USA: IEEE, 2019: 529-537.
- [11] WANG G M, WU X R, LIU Z, et al. Hierarchical attention learning of scene flow in 3D point clouds[J]. IEEE Transactions on Image Processing, 2021, 30: 5168-5181.
- [12] ZHANG J M Y, ZHOU S P, CHANG X, et al. Multiple object tracking by flowing and fusing[DB/OL]. (2020-01-30) [2023-01-15]. <https://arxiv.org/abs/2001.11180>.
- [13] WANG S K, SUN Y X, LIU C J, et al. PointTrackNet: An end-to-end network for 3-D object detection and tracking from point clouds[J]. IEEE Robotics and Automation Letters, 2020, 5(2): 3206-3212.
- [14] SHI S S, GUO C X, JIANG L, et al. PV-RCNN: Point-voxel feature set abstraction for 3D object detection[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, USA: IEEE, 2020: 10526-10535.
- [15] WU W X, QI Z A, LI F X. PointConv: Deep convolutional networks on 3D point clouds[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, USA: IEEE, 2019: 9613-9622.
- [16] KUHN H W. The Hungarian method for the assignment problem[J]. Naval Research Logistics Quarterly, 1955, 2(1/2): 83-97.
- [17] GEIGER A, LENZ P, URTASUN R. Are we ready for autonomous driving? The KITTI vision benchmark suite[C]//IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, USA: IEEE, 2012: 3354-3361.
- [18] LUITEN J, OSEP A, DENDORFER P, et al. HOTA: A higher order metric for evaluating multi-object tracking[J]. International Journal of Computer Vision, 2021, 129: 548-578.
- [19] SMITH L N, TOPIN N. Super-convergence: Very fast training of neural networks using large learning rates[C]//Artificial Intelligence and Machine Learning for Multi-domain Operations Applications. Bellingham, USA: SPIE, 2019: 369-386.
- [20] SIMON M, AMENDE K, KRAUS A, et al. Complexer-YOLO: Real-time 3D object detection and tracking on semantic point clouds[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Piscataway, USA: IEEE, 2019: 1190-1199.

作者简介:

王光明 (1996—), 男, 博士, 博士后。研究领域: 机器人, 计算机视觉, 自动驾驶。

王贺升 (1980—), 男, 博士, 教授。研究领域: 视觉伺服, 机器人, 计算机视觉。