控制与决策
Control and Decision

图像与点云三维体信息交互的3D多目标跟踪网络

杨浩冉, 李辉, 艾晓雪, 赵国伟, 郭颖

引用本文:
杨浩冉, 李辉, 艾晓雪, 等. 图像与点云三维体信息交互的3D多目标跟踪网络[J]. 控制与决策, 2024, 39(12): 4127-4135.

在线阅读 View online: https://doi.org/10.13195/j.kzyjc.2023.1660

图像与点云三维体信息交互的3D多目标跟踪网络

杨浩冉, 李 辉†, 艾晓雪, 赵国伟, 郭 颖

(青岛科技大学 数据科学学院，山东 青岛 266061)

摘 要: 多目标跟踪是自动驾驶领域中的一个关键问题. 然而,仅依赖单一图像信息或点云信息难以克服复杂场景下的跟踪挑战,目前,多模态融合的跟踪方法在融合性能、数据关联、轨迹管理等方面仍然存在许多问题. 为此,提出图像与点云三维体信息交互的3D多目标跟踪网络. 首先,设计三维体特征交互模块来获取目标的三维体形态信息,得到更有判别性的特征,提升复杂场景下的定位精度;然后,设计基于三维综合运动估计的数据关联,利用卡尔曼滤波以及目标在点云中的运动信息,获取目标在下一帧中的位置预测,从而提升目标在帧间的一致性;最后,为进一步增强轨迹关联的鲁棒性,设计一种基于三维体特征的轨迹管理模块,以更好地克服目标消失-重现的关联问题. 在KITTI数据集上的实验结果表明,与其他方法相比,所提出跟踪方法具有更好的跟踪性能.

关键词: 点云；3D多目标跟踪；三维体特征；信息交互；运动估计

中图分类号: TP391.4 **文献标志码:** A

DOI: 10.13195/j.kzyjc.2023.1660

引用格式: 杨浩冉, 李辉, 艾晓雪, 等. 图像与点云三维体信息交互的3D多目标跟踪网络[J]. 控制与决策, 2024, 39(12): 4127-4135.

3D multi-object tracking network based on 3D volume information interaction between image and point cloud

YANG Hao-ran, LI Hui†, AI Xiao-xue, ZHAO Guo-wei, GUO Ying

(School of Data Science，Qingdao University of Science and Technology，Qingdao 266061，China)

Abstract: Multi-object tracking is a crucial area in autonomous driving. However, it is difficult to overcome the challenges of complex scenes by only relying on a single image or point cloud information, and the current multi-modal fusion tracking methods still have many problems in fusion strategy, data association and trajectory management. Therefore, we propose a 3D multi-object tracking network based on 3D volume information interaction between image and point cloud. Firstly, we design the 3D volume feature interaction module to obtain the 3D volume morphology information of the object, so as to get more discriminative features and improve the localisation accuracy in complex scenes. Secondly, we design the data association based on 3D integrated motion estimation, using Kalman filtering and the object's motion information in the point cloud to obtain the prediction of the object's position in the next frame, so as to enhance the consistency of the object between frames. Finally, to further enhance the robustness of trajectory association, we design a 3D volume feature-based trajectory management module to better overcome the object disappearance-reappearance association problem. Experimental results on the KITTI dataset show that the proposed method has better tracking performance compared with other methods.

Keywords: point cloud；3D multi-object tracking；3D volume feature；information interaction；motion estimation

0 引 言

3D多目标跟踪是自动驾驶的一项基础任务. 当前基于深度学习的跟踪方法[1]取得了不错的性能. 但是由于复杂场景存在目标相似度高、目标重叠等问题[2],现有方法在复杂场景下跟踪性能较低. 由于图像中丰富的纹理等外观信息,基于图像的目标跟踪能够更易对目标建模,并取得良好的跟踪精度[3-4]. 但是,图像的二维局限性导致单目图像缺少深度信息,在三维空间定位不准确. 随着激光雷达广泛应用于自动驾驶领域,其获取的三维点云数据很好地弥补了图像的缺陷,具有精确的空间位置信息,保留了物体的三维形态[5]. 但是,激光雷达获取的点云缺乏色彩

收稿日期: 2023-11-30；**录用日期:** 2024-03-26.

基金项目: 国家重点研发计划项目(2023YFF0612102)；青岛市关键技术攻关及产业化示范类项目(23-7-2-qljh-4-gx, 24-1-2-qljh-19-gx).

†通讯作者. E-mail: lihui@qust.edu.cn.

纹理信息,且有稀疏、近密远疏等缺陷,对小目标、较远目标难以准确地跟踪.

图像与点云特征相融合能够有效克服单模态跟踪方法的局限[6-7],进一步提升目标跟踪的性能,为自动驾驶提供更准确的环境感知信息. 但是由于两种模态的数据类型不同,多模态的3D多目标跟踪仍然存在以下问题: 1) 现有方法[8]专注于提取更精确的3D检测框,忽略了目标的三维形态信息,导致目标三维特征判别性较差; 2) 特征融合仅利用点云位置信息增强图像特征[9]或利用图像增强点云的语义特征[10],无法真正达到特征互补的效果; 3) 由于跟踪场景的复杂性,目标遮挡、消失-重现等情况导致ID频繁切换,对跟踪性能造成不利影响.

针对以上问题,本文提出图像与点云三维体信息交互的3D多目标跟踪网络,主要内容如下: 1) 提出三维体特征交互融合模块. 该模块提取目标三维体特征,并利用局部特征交互和全局特征融合得到精准的外观特征. 2) 提出基于三维综合运动估计的数据关联. 综合三维中心点运动估计和两帧运动估计,构建运动亲和度矩阵,并联合外观特征亲和度完成数据关联,减少目标间的错误匹配. 3) 设计基于三维体特征的消失目标图. 利用三维体特征保留消失目标的轨迹特点,增加目标的匹配机会,有效解决ID频繁转换问题.

1 相关工作

1) 基于图像的3D多目标跟踪.

基于图像的3D多目标跟踪分为前置摄像机完成的单目3D多目标跟踪和多摄像机完成的多视图3D多目标跟踪. 两者均使用二维图像估计目标的三维信息. Li等[11]组建了时空信息流模块聚合目标的几何特征和外观特征,再将其聚合到特定的帧中,计算时间域中所有目标的亲和度,以此完成跟踪. 由于基于图像的方法跟踪性能受深度估计精度的限制,与基于其他模态的3D多目标跟踪相比性能较差. 为了弥补单目图像没有深度信息的缺点,研究者使用多视图图像来构建3D空间. Li等[12]使用一系列图像构建伪点云,引入了相机参数作为深度估计的先验,利用点云数据来监督预测的深度,解决了基于图像的方法深度估计不足的问题. 其可对多视图图像的数据集进行跟踪,但是在训练时需要使用点云数据来保证预测深度的准确性.

2) 基于点云的3D多目标跟踪.

点云具有不规则、无序等特点,因此传统卷积网络无法直接应用于点云. 为了解决这一问题,研究者对点云进行体素化处理或映射为BEV图等伪图像,规范点云,使其能够使用卷积网络. 也有研究者直接使用点云数据构建深度学习模型. PointNet是其中的开创性工作,但是PointNet只考虑了全局特征,丢失了点云的局部信息. 在其基础上改进的PointNet ++采用分层抽取特征的思想,不仅可以提取点云的局部特征,还解决了点云密度不均问题. Hussain等[13]使用Pillar或Voxel网络进行体素化,通过混合时间中心图辅助跟踪,消除了传统的启发式匹配,较好地应对遮挡问题. AB3DMOT[1]使用卡尔曼滤波加匈牙利匹配算法的方式进行跟踪,仅使用3DIOU进行目标关联,验证了运动估计的有效性; CenterPoint[14]将目标检测问题转化为无监督的聚类和回归问题,生成3D检测框并用其进行跟踪; SimpleTrack[15]在CenterPoint的基础上使用Generalized IoU (GIoU),改变检测框的匹配方式,解决了IOU过小导致的轨迹消失问题. 但是这些方法忽略了点云三维体形态信息的判别性,导致基于点云的3D多目标跟踪在复杂场景下有较多的误检和错误匹配.

3) 基于点云和图像多模态的3D多目标跟踪.

为了充分利用图像和点云的优势,获得更具有判别性的特征,研究者融合两个模态的特征进行跟踪[6-8]. 其中: GNN3DMOT[16]使用图神经网络进行目标间的特征交互,使得相似的特征更加相似,能够很好地跟踪快速移动的物体,但是在特征聚合上还需改进. Kim等[8]使用现有的目标检测算法生成2D检测框和3D检测框,关联属于同一目标的2D检测结果和3D检测结果,再利用多阶段关联结果完成跟踪,忽略了目标的三维形态对跟踪的影响. JMODT[7]对点云和图像进行逐点特征融合,完成了跨模态特征交互. 但是,JMODT缺少全局间的跨模态特征融合,目标消失-重现等情况下会频繁切换ID. 为了应对这个问题, JoDT[17]设计了目标与目标关联的图神经全连接网络,构建了时空特征聚合网络SFANet,利用空间和时间的上下文信息减少了ID切换次数. 基于点云和图像多模态的3D多目标跟踪方法可实现更好的跟踪性能,但是仍然存在不能充分利用两种模态优势、ID频繁切换等问题.

2 网络框架及其实现

2.1 网络框架

所提出方法的网络结构如图1所示,主要包括3个部分: 多模态特征交互融合、综合运动和外观的数据关联、基于三维体特征的轨迹管理. 首先,输入为连续两帧的点云和图像,图像采用双分支并行注意力增强语义特征,点云在与图像完成局部的逐点特征融

合后,提取并融合三维体特征;然后,利用多尺度融合网络,完成多模态全局语义融合. 在数据关联方面,除预测两帧的运动外,使用三维点运动估计计算两帧间三维体中心点的速度,预测上一帧目标在当前帧的位置;最后,经过消失目标图完成轨迹管理,判断出新生目标、消亡目标和跟踪轨迹,完成跟踪.

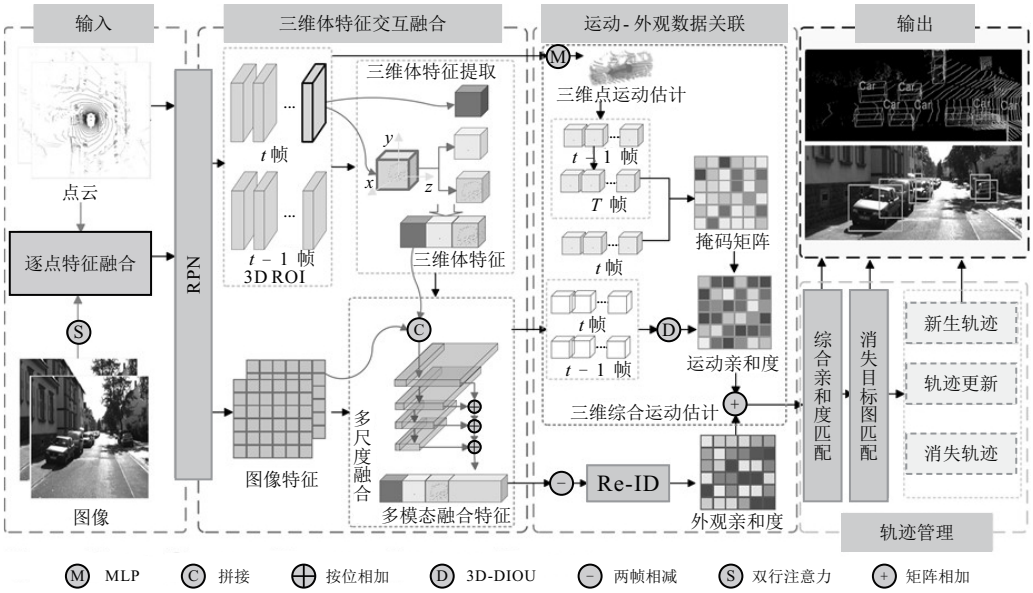


图 1　本文方法的网络结构

2.2　图像与点云三维体信息交互融合

2.2.1　三维体特征提取与交互

图像可为点云提供先验信息,但是在拍摄过程中受光源等因素影响,会出现清晰度不足等问题,直接使用图像特征难以适应复杂场景的跟踪任务. 因此,在图像与点云逐点融合前,使用双分支并行注意力增强图像语义特征,提升特征的表示能力. 双分支并行注意力包括通道注意力和位置注意力:通道注意力每层对应一种特征,可学习不同特征的相关性和重要性,增强网络对关键特征的关注;位置注意力根据上下文信息构建语义特征,提升特征的表示能力. 与串行计算方式相比,并行方式减小了分支间的影响. 与其他注意力[18]相比,双分支并行注意力降低了参数量,可构建语义间的依赖关系. 完成图像特征增强后,通过卷积得到与原始图像相同分辨率的特征图,根据已知传感器校准矩阵,建立特征图二维位置与三维点云间的点对应关系,然后利用双线性插值在连续坐标下得到相应的图像特征 F'_{2d} 交互,补充点云的语义信息,完成点云特征与图像特征的局部融合,如图2所示. 点云中目标的三维形态信息,不仅能够加强轮廓位置信息,回归出更精确的检测框,也可作为外观特征,判别目标. 为了强调目标的三维信息,利用点云提取三维体特征,再与点云特征和图像特征融合,获得更具有判别力的特征. 如图2中下半部分所示:点云经过3D RPN网络得到 N 个3D ROI,每个ROI内有 M 个点,将其进行一次MLP操作,得到三维体的中心点坐标,第 i 个ROI的中心点坐标为 (x_i, y_i, z_i). 对该ROI内的第 j 个点求取相对坐标,其

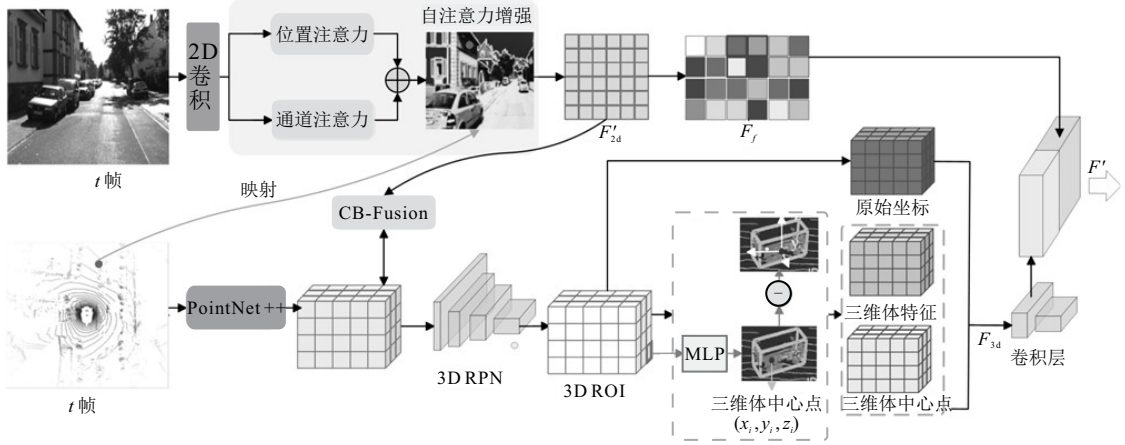


图 2　三维体特征交互融合网络

操作可描述为

$$(x'_{ij}, y'_{ij}, z'_{ij}) = (x_{ij}, y_{ij}, z_{ij}) - (x_i, y_i, z_i). \quad (1)$$

将原始点(x_{ij}, y_{ij}, z_{ij})、相对点$(x'_{ij}, y'_{ij}, z'_{ij})$、三维体中心点$(x_i, y_i, z_i)$进行拼接后,再经过两层卷积,得到三维体特征. 将图像特征F_f与三维体特征$F_{3\mathrm{d}}$拼接得到特征F',输入至多尺度融合网络.

2.2.2 多尺度融合网络

上一阶段完成了图像与点云局部信息的交互,但是仍然难以准确判别出不同尺度的同一目标,且将图像特征图像特征F_f与三维体特征$F_{3\mathrm{d}}$简单拼接的融合方法过于粗糙,不能很好地将图像特征与三维体特征融合在一起. 因此,本文设计了图像与点云分级融合的多尺度融合网络. 逐层融合多模态特征,分级交互目标的多模态信息并学习目标的多尺度信息. 紧密融合图像特征和三维体特征的同时,加强了两个模态的相关性和鲁棒性,减小了目标尺度变化对跟踪的影响. 图3为多尺度融合网络. 如图3所示: 多尺度融合网络共分4级,使用步长为1、2、4、8的计算方式实现. $F_{3\mathrm{d}}$经过4个不同步长的Block块,得到不同分辨率的三维体特征. 每个Block由下采样卷积层、ReLu激活层、跳跃层组成. F_f同样要经下采样得到不同分辨率的图像特征. 每级使用FL模块将三维体特征作为先验信息叠加在图像特征中,得到三维体和图像的融合特征. 然后将不同分辨率的融合特征上采样到统一维度,分别输出特征F'、F'_1、F'_2、F'_3. 将4个特征的元素相加,得到的图像-三维体融合特征F为$F = F' + F'_1 + F'_2 + F'_3$.

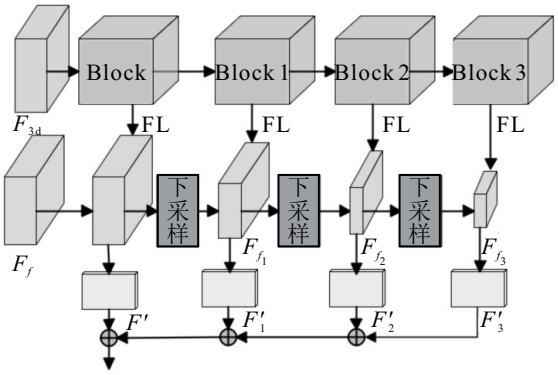


图3 多尺度融合网络

2.3 基于三维综合运动估计的数据关联

现有的数据关联方法中,运动特征通常基于卡尔曼滤波估计来构建关联矩阵,对于长时间遮挡的目标估计误差较大. 因此,本文设计三维综合运动估计模块,在卡尔曼滤波的基础上构建更鲁棒的运动特征关联矩阵. 图4为关联匹配过程. 如图4所示: 首先,对两帧3D ROI中点云经过MLP得到速度v,作为三维体中心的估计速度,再将上一帧的三维体中心点转移至当前帧的坐标下,提升目标在帧间的相似性; 然后,计算三维体中心点的相关性作为目标掩码; 接着,通过两帧融合特征F_{t-1}与F_t得到两帧的3D检测结果D_t和D_{t-1}的3D-DIOU; 最后,输出的运动亲和度矩阵为$A^{\mathrm{iou}} = \{\alpha^{\mathrm{diou}}_{d_t, d_{t-1}}, d_t \in D_t, d_{t-1} \in D_{t-1}\}$.

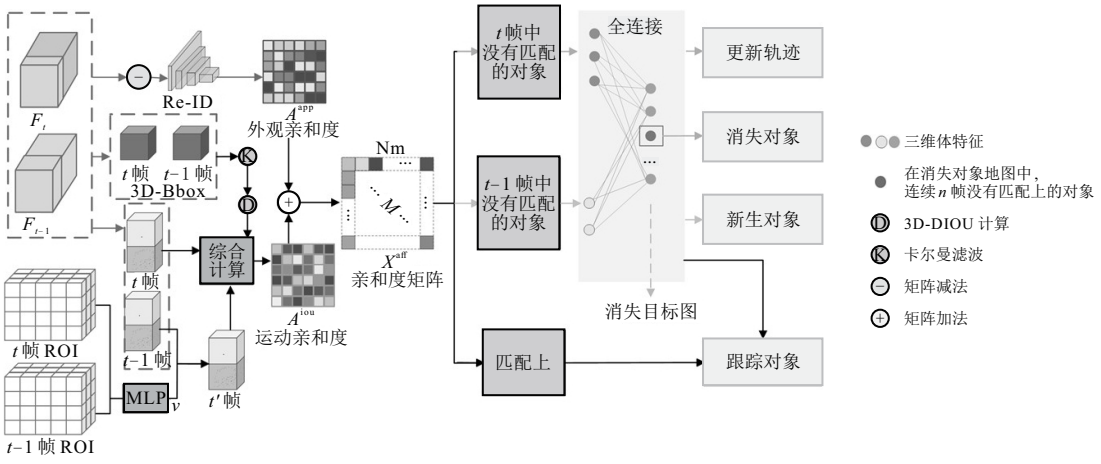


图4 关联匹配过程

3D-DIOU计算公式为

$$a^{\mathrm{diou}}_{t,t-1} = \left(1 - \frac{\rho(b_t, b_{t-1})}{l}\right) + \frac{B_t \bigcap B_{t-1}}{B_t \bigcup B_{t-1}}. \quad (2)$$

其中: ρ为欧氏距离; l为能够完全覆盖两个3D检测框的最小框对角线长度; B_t和B_{t-1}分别为目标i在t帧和$t-1$帧的3D检测框; b_t为B_t的中心; 同理, b_{t-1}为B_{t-1}的中心; $B_t \bigcap B_{t-1}$为目标i的两帧3D检测框的交集; $B_t \bigcup B_{t-1}$为目标i的两帧3D检测框的并集.

外观特征的相似性是判别目标的重要依据,可以帮助网络更加快速、准确地判别目标. 因此,本文利用相邻两帧目标的外观相似性特征计算外观亲和度. 如图4所示: 两帧的相似性特征$F_{t-1,t} = |F_t - F_{t-1}|$,

将其输入ReID网络得到外观亲和度矩阵A^{app}. 综合亲和度矩阵X^{aff}计算如下所示:

$$X^{\text{aff}} = \alpha A^{\text{app}} + \beta A^{\text{iou}}, \tag{3}$$

其中$\alpha + \beta = 1$.

2.4　基于三维体特征的消失目标图匹配

上一阶段的数据关联只匹配相邻两帧间的目标, 保证相邻两帧中同一目标ID不变, 并不能保证整个跟踪过程中目标维持同一ID. 当有遮挡等情况时, 仅利用上一阶段进行跟踪会导致ID频繁切换. 为了克服这一问题, 设计基于三维体特征的消失目标图匹配, 构建消失目标图, 存储消失目标的轨迹, 判断目标的轨迹状态(新生、跟踪更新、消失). 三维体特征与三维体中心点相关, 无需将特征转移至当前帧坐标下, 可直接完成较为准确的跨帧匹配, 与使用整个多模态融合特征作为存储特征相比, 单独使用三维体特征简化了匹配方法、减少了内存的使用. 如图4所示: 对未匹配成功的目标$d \in D_t$、D_{t-1}和消失目标图中的目标$k \in K$进行三维体特征的全连接匹配, 若匹配成功, 则d继承目标k的ID. 若t帧中目标匹配失败, 则判定为新生目标并分配新的ID. 若$t-1$中的目标匹配失败, 则判定为预消失目标加入消失目标图K中, 目标k若连续帧与d匹配失败, 则判定为消失目标. 若设置Y为正在被跟踪的目标, y^{aff}为相邻两帧间匹配的目标, y^{dis}为与消失地图匹配的目标, X^{dis}为相邻两帧匹配失败的目标与消失地图中目标的亲和度矩阵, 则

$$Y = y^{\text{aff}} + y^{\text{dis}}, \tag{4}$$

$$y^{\text{aff}} = X^{\text{aff}} |D_t - D_{t-1}|, \tag{5}$$

$$y^{\text{dis}} = X^{\text{dis}}(D_t, K) + X^{\text{dis}}(D_{t-1}, K). \tag{6}$$

消失地图中的目标总数为

$$K = K + X^{\text{dis}}(D_{t-1}, K). \tag{7}$$

2.5　Loss计算

总损失包括RPN和RCNN的分类回归损失, 即

$$L_{\text{total}} = L_{\text{rpn}} + L_{\text{rcnn}}. \tag{8}$$

对于RPN的分类损失, 本文使用如下焦点损失:

$$\text{FL}(p_t) = -\alpha_t (1 - p_t)^{\gamma} \log(p_t). \tag{9}$$

其中: p_t为预测类别的概率, 即Sigmoid的输出概率; 以α和γ来平衡正负样本. 使用L1损失对真实框进行优化. 对于回归出的3D检测框$(x, y, z, l, h, w, \theta)$使用L1损失进行优化后, 采用基于框的回归损失.

所提出方法的RPN损失为

$$L_{\text{rpn}} = L_{\text{cls}} + L_{\text{reg}}, \tag{10}$$

$$L_{\text{cls}} = -\alpha (1 - C_p)^{\gamma} \log C_p, \tag{11}$$

$$L_{\text{reg}} = \sum_{u \in x,z,\theta} E(b_u, \hat{b}_u) + \sum_{u \in x,y,z,h,w,l,\theta} S(r_u, \hat{r}_u). \tag{12}$$

其中: E为交叉熵损失, S为L1损失, $\hat{b}_u$和$\hat{r}_u$分别为3D真实框和残差偏移量. RCNN损失为分类损失、回归损失、关联模块损失的和, 分类损失和回归损失与RPN损失计算相同, 关联模块损失使用L1损失.

3　实验结果和分析

3.1　数据集和评价指标

为了评估所提出跟踪方法的性能, 在KITTI数据集上进行实验. KITTI包括3种输入数据: 前向摄像机的图像数据、64E激光雷达采集的点云数据和传感器校准数据. 所提出方法在GPU为RTX 3090的服务器上训练和测试. 实验中: JMODT为所提出方法的基线网络模型, 预训练模型为EPNet++[19]在GitHub上提供的检测模型. 训练轮次为50, 批处理大小为2, 学习率为2e-4. 亲和度计算中, 超参数经交叉验证, 最终设置如下: $\beta = 10\alpha$, $\theta_{\text{cls}} = 0.85$, $\alpha = 0.25$, $\gamma = 2.0$.

3.2　消融实验

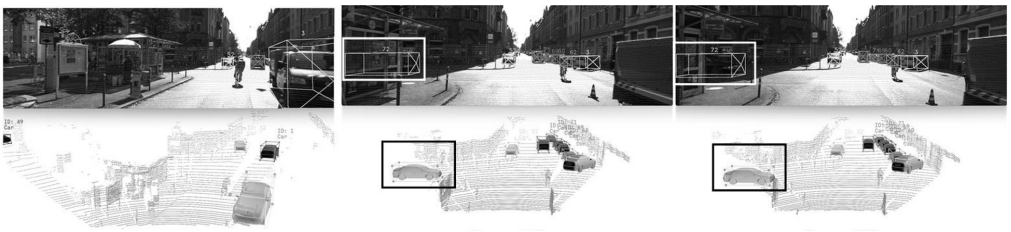
3.2.1　三维体特征交互融合的消融实验

为了验证三维体特征交互融合模块(VMFusion)的有效性, 在基线网络模型采用CB-Fusion的基础上, 逐次增加三维体特征交互模块(3D Vm)、多尺度融合模块(Mul)进行消融实验, 实验结果如表1所示. 由表1可见: 与基线方法相比, 在添加3D Vm后, MOTA上升了1.44%, MOTP上升了0.8%; 只添加Mul后, MOTA上升了0.15%, MOTP上升了0.12%; 两者同时使用时, MOTA达到86.38%, 上升了2.01%, MOTP达到88.38%, 上升了1.14%, 取得了最好的跟踪效果. 为了判断融合部分对检测的影响, 在这次消融实验中增加了指标MODP来判断模型的检测精度. 在对比中发现, 增加3D Vm和Mul后, MODP上升了0.63%. 因此, 所提出方法的VMFusion有助于提高三维检测精度. 另外, 所提出方法选用传统多尺度融合网络(Mul_T)与Mul进行对比实验. 对比表1第4行与第5行可以看出: 在其他条件相同时, 使用Mul的网络MOTA、MOTP、FN等指标均高于Mul_T. 对于跟踪任务而言, 3D Vm使得点云特征增加了丰富的语义信息, 同时, 图像特征增加了目标的3D空间信息, 提高了网络判别目标的正确率. Mul对交互后的特征进行分级多尺度融合, 使得图像特征与三维体特征可在不同尺度上紧密融合, 也使得融合特征增加了目标的多尺度信息. VMFusion能够更加细致、有效地融合图像特征和点云特征, 使得网络适应目标尺度的变化, 提高了网络的鲁棒性.

表 1 多模态特征融合消融实验

CB-Fusion	3D Vm	Mul	Mul_T	MOTA/%↑	MOTP/%↑	FN↓	FP↓	MT/%↑	ML↓	IDS↓	MODP/%↑
✓				84.37	87.24	1 226	506	80.09	2.31	4	90.23
✓	✓			85.81	88.10	1 000	546	83.10	**1.39**	**2**	90.56
✓		✓		84.52	87.36	1 215	**498**	81.02	2.32	7	90.08
✓	✓		✓	84.41	88.26	1 225	457	80.13	2.31	7	90.34
✓	✓	✓		**86.38**	**88.38**	**973**	569	**86.65**	2.31	4	**90.86**

图5为增加三维体特征交互融合模块的实验结果. 如图5所示: 被粗框标出的ID为72车辆在第112帧图片中被严重遮挡, 在第113帧图片中出现, 但是也存在遮挡情况. 所提出方法完成了对该车辆的跟踪, 表明VMFusion提升了复杂场景下判别目标的性能. 在第96帧中, 所提出方法对远处的小目标也有很好的跟踪结果, 表明所提出方法能够适应目标尺度的变化.



frame 96 frame 112 frame 113

图 5 增加三维体特征交互融合模块的实验结果

3.2.2 三维综合运动估计消融实验

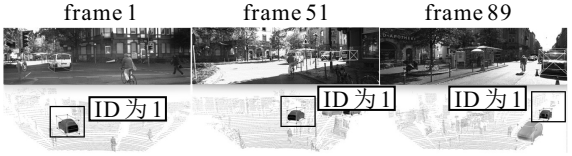
为了验证三维点运动估计的有效性, 进行如下消融实验. 如表2所示: 在基线网络上增加三维点运功估计后, MOTA上升了1.9%. 同时增加VMFusion和三维点运动估计模块后, MOTA上升了3.43%. 不使用VMFusion, 仅增加三维点运动估计, MOTA也有提升. 这是因为三维点运动估计在卡尔曼滤波运动估计前, 得到目标三维体中心点在两帧的相对运动估计作为掩码, 能够更加精准地判断两帧间目标的运动关系, 达到更好的跟踪性能.

表 2 三维点运动估计消融实验

VMFusion	三维点运动估计	消失目标图	MOTA/%↑	MOTP/%↑	FN↓	FP↓	MT/%↑	ML↓	IDS↓
			84.37	87.24	1 226	506	80.09	2.31	4
	✓		86.27	85.41	978	571	85.21	2.32	**1**
✓			86.38	88.38	973	569	86.65	2.31	4
✓	✓		87.80	88.39	**956**	596	87.05	1.39	2
✓	✓	✓	**88.51**	**89.94**	961	**542**	**87.14**	**1.37**	**1**

3.2.3 消失目标图消融实验

如表2所示: 在使用VMFusion和三维点运动估计的基础上, 增加消失目标图后模型的IDS更低, MOTA上升了0.71%, MOTP上升了1.55%, FP降低至542. 表明使用消失目标图能够增加跟踪准确度, 改善ID频繁转换的问题. 消失目标图可在两帧匹配的基础上, 建立轨迹, 保存曾出现目标, 有利于长时间跟踪目标. 图6为消失目标图消融实验结果对比. 如图6所示: 在使用消失目标图时, 被粗框标出的ID为1的车辆因拐弯消失再出现, 且在图像中被遮挡, 该车辆ID仍然为1; 在不使用消失目标图时, 该车辆拐弯后再出现, ID转换为35.



(a) 使用Disappear-object map的实验结果

(b) 不含Disappear-object map的实验结果

图 6 消失目标图消融实验结果对比

3.3 跟踪性能对比分析

为了更加公平地与其他方法进行对比, 采用KITTI官方指定的评估标准和指标, 分别将汽车(Car)和行人(Pedestrian)类别的跟踪结果与其他先进多模态3D多目标跟踪方法进行对比评估. 表3为Car类别

对比结果,表4为Pedestrian类别的对比结果. 由于部分多模态3D多目标跟踪方法并没有在Pedestrian类别上进行实验,本文Car类别的对比方法与Pedestrian类别的对比方法稍有不同.

表 3　KITTI基准跟踪数据集上Car的跟踪性能对比

方法	MOTA/%↑	MOTP/%↑	HOTA/%↑	FN↓	FP↓	DetPr/%↑	MT/%↑	ML↓	IDS↓	LocA/%↑
JMODT [7]	85.35	85.37	70.73	1 249	3 438	84.02	77.39	2.92	350	86.95
EagerMOT [8]	87.82	85.69	74.39	**454**	3 497	86.42	76.15	2.46	239	87.17
JRMOT [20]	85.10	85.28	69.61	787	4 067	85.07	70.92	4.62	271	86.72
mmMOT [21]	83.23	85.03	62.05	752	4 284	84.89	72.92	2.92	733	86.58
DeepFusion-MOT [22]	84.63	85.02	75.46	601	4 601	85.25	68.61	9.08	84	86.70
MSA-MOT [23]	**88.01**	85.45	**78.52**	2 060	1 974	82.21	**86.77**	**1.23**	91	87.00
YONTD-MOT [24]	85.09	86.98	78.08	1 188	3 899	85.71	67.54	7.08	**42**	88.23
FNC 2 [25]	84.21	85.86	73.19	2 472	2 763	81.67	75.85	6.00	195	87.31
FANTrack [26]	75.84	82.46	60.85	1 305	6 262	80.82	62.77	8.77	743	84.72
ours	87.23	**88.72**	75.87	2 931	**1 493**	**86.57**	81.02	1.31	81	**89.66**

表 4　KITTI基准跟踪数据集上Pedestrian的跟踪性能对比

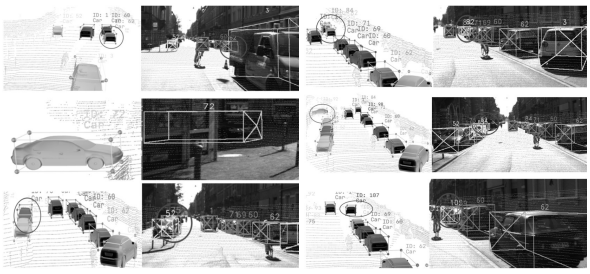
方法	MOTA/%↑	MOTP/%↑	HOTA/%↑	FN↓	FP↓	DetPr/%↑	MT/%↑	ML↓	IDS↓	LocA/%↑
EagerMOT [8]	49.82	64.42	39.38	2 161	**8 959**	61.49	24.49	24.05	496	71.25
JRMOT [20]	45.31	72.22	34.24	10 207	1 822	66.64	24.40	29.55	631	76.64
EAFFMOT [27]	42.01	64.57	40.20	2 431	10 793	60.03	21.99	35.40	201	71.25
MMTrack [28]	56.19	75.34	**49.28**	**886**	9 081	72.98	30.93	32.65	**175**	79.26
MSA-MOT [23]	47.86	64.35	44.73	4 101	7 761	55.94	**33.68**	**16.15**	209	71.21
YONTD-MOT [24]	26.19	65.66	25.89	2 882	13 137	54.99	11.00	31.96	1 068	72.10
Be-Track [29]	50.85	72.45	43.36	1 225	9 953	69.03	22.34	32.65	199	76.78
ours	**60.27**	**77.58**	47.76	1 011	10 321	**74.52**	28.68	23.14	181	**80.51**

3.3.1　Car类别跟踪性能对比分析

对比结果如表3所示: 所提出方法跟踪精度MOTP达到88.72%,在表3中排名第1,比排名第2的方法高1.74%. 误报率FP、检测精度DetPr和定位精度LocA也高于其他方法,分别为1 493、 86.57%、89.66%. MT、ML和IDS虽然不是最好,但是也仅低于MSA-MOT与YONTD-MOT. MSA-MOT采用了分层匹配和定制轨道管理的多级关联的方法,缓解了单阶段关联鲁棒性差的问题,MOTA和HOTA最好. 但是MSA-MOT并没有对多模态融合提出较大改进,且跟踪时忽略了目标的三维形态,因此其MOTP远低于所提出方法. YONTD-MOT主要针对ID频繁转化问题,探讨了历史轨迹回归置信度对跟踪的影响,因此其有着最好的IDS,但其LocA低于所提出方法1.43%. EagerMOT使用多阶关联的方法,适用范围广,可结合不同检测算法使用,在表3方法中有着最好的漏检FN,同时MOTA和DetPr位居第2. 但是,EagerMOT忽略了外观特征的重要性,相较于所提出方法,其他指标均较差. DeepFusion-MOT同样利用检测结果进行多阶段关联,但其仅有FN较好. 这反应出仅使用检测结果并不能很好地判别目标. 原因为检测结果仅含有检测框,失去了外观等有利于区分目标的信息. 所提出方法可以更好地区分不同目标,同时也能更好地识别消失-重现的目标,表明三维体特征对于检测和跟踪均是有利的.

由表3可知,所提出方法的MOTA、HOTA和FN有待提升. 这是因为所提出方法注重跟踪精度和检测准确度,会出现一定的目标丢失或误判,但是MOTA、HOTA仍然排名第3,优于大多数方法.

对于Car的跟踪结果,主要分析复杂场景下的遮挡和远处小目标的检测跟踪,如图7所示. 由图7可见: 图像中由圆圈标出的小目标和没有圈出的大目标均被严重遮挡,但是均被检测出,表明所提出方法对图片中被遮挡的目标和小目标也有比较好的检测结果.



(a) 被遮挡目标　　(b) 被遮挡的小目标

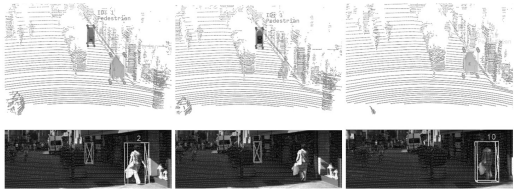
图 7　复杂场景下对Car的跟踪结果

3.3.2　Pedestrian类别跟踪性能对比分析

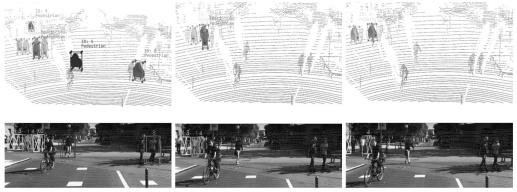
如表4所示,所提出方法在Pedestrian类别的评估结果中表现出最佳性能. 其中: MOTA、 MOTP、DetPr、 LocA在所有对比方法中排名第1,分别为60.27%、77.58%、74.52%、80.51%. 其他指标中所提出方法的HOTA、FN、ML、IDS在所有方法中排名第

2, FP和MT还有改进空间. EagerMOT并没有针对Pedestrian类别进行单独的训练, 而是使用检测结果利用欧氏距离进行关联跟踪, 因此跟踪效果稍差. 与EagerMOT相比, 所提出方法MOTA提高了10.45 %. YONTD-MOT将目标检测和多目标跟踪集成至同一个模型中, 主要是针对Car的类别进行改进, 对于小目标跟踪效果较差, 因此在Pedestrian类别中表现较差. MMTrack只针对Pedestrian类别训练, 通过特征解耦的方法缓解了目标检测与目标匹配间的竞争, 有着最好的HOTA和FP. 但是MMTrack的MOTA和MOTP与所提出方法相比, 分别低4.08 %和2.24 %. 所提出方法在Car类别和Pedestrian类别均具有良好的跟踪性能, 也验证了所提出方法具有泛化性.

对于Pedestrian类别的跟踪结果, 主要从ID转换问题和漏检问题进行分析. 图8为基线方法对Pedestrian的跟踪结果, 图9为所提出方法对Pedestrian的跟踪结果. 图8的场景1中: 第1帧检测出的两个行人ID 1和ID 2在第3帧中出现漏检, 且在第8帧, 行人ID 2再次被检测时ID变为10. 场景2的连续3帧中, 第2帧和第3帧均出现了严重漏检. 图9的场景1中没有漏检, 且两个行人的ID保持不变. 场景2连续帧中出现的目标均已检测出. 表明所提出方法可改善针对小目标的漏检和ID频繁转换的问题, 提高了对于小目标的检测和跟踪性能.

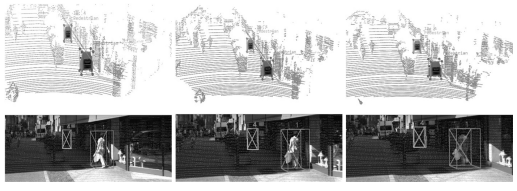


(a) 场景1

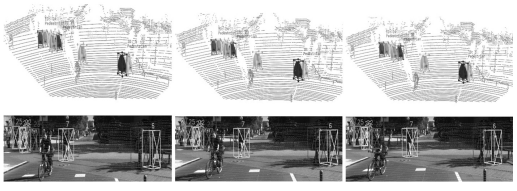


(b) 场景2

图 8 基线方法对Pedestrian的跟踪结果



(a) 场景1



(b) 场景2

图 9 本文方法对Pedestrian的跟踪结果

4 结 论

本文提出了一种图像与点云三维体信息交互的3D多目标跟踪网络, 该方法通过点云和图像特征交互, 增强了特征的融合效果. 通过引入表示物体三维形态的三维体特征, 获得了更好的三维跟踪效果, 并在实验中得到了验证. 由实验结果可以看出, 消失目标图能够有效地改善ID频繁转换的问题, 提高了长时间跟踪的精度. 未来计划通过跨帧分享特征的方式进一步改善外观亲和度计算方法, 优化网络模型, 降低目标的丢失率, 以提高检测和跟踪的准确度.

参考文献(References)

[1] Weng X S, Wang J R, Held D, et al. 3D multi-object tracking: A baseline and new evaluation metrics[C]. Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems. Las Vegas, 2020: 10359-10366.

[2] Chen Y K, Liu J H, Zhang X Y, et al. VoxelNeXt: Fully sparse VoxelNet for 3D object detection and tracking[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, 2023: 21674-21683.

[3] 仲训杲, 范东嘉, 仲训昱, 等. 融合多模板注意力深度网的自适应目标框跟踪算法[J]. 控制与决策, 2024, 39(4): 1123-1132.
(Zhong X G, Fan D J, Zhong X Y, et al. Adaptive target box tracking algorithm by integrating multi-template attention deep network[J]. Control and Decision, 2024, 39(4): 1123-1132.)

[4] 朱姝姝, 王欢, 严慧. 基于帧内关系建模和自注意力融合的多目标跟踪方法[J]. 控制与决策, 2023, 38(2): 335-344.
(Zhu S S, Wang H, Yan H. Multi-object tracking based on intra-frame relationship modeling and self-attention fusion mechanism[J]. Control and Decision, 2023, 38(2): 335-344.)

[5] Sadjadpour T, Li J, Ambrus R, et al. ShaSTA: Modeling shape and spatio-temporal affinities for 3D multi-object tracking[J]. IEEE Robotics and Automation Letters, 2024, 9(5): 4273-4280.

[6] Wang L, Zhang X Y, Qin W Y, et al. CAMO-MOT: Combined appearance-motion optimization for 3D multi-object tracking with camera-LiDAR fusion[J].

IEEE Transactions on Intelligent Transportation Systems, 2023, 24(11): 11981-11996.

[7] Huang K M, Hao Q. Joint multi-object detection and tracking with camera-LiDAR fusion for autonomous driving[C]. IEEE/RSJ International Conference on Intelligent Robots and Systems. Prague, 2021: 6983-6989.

[8] Kim A, Ošep A, Leal-Taixé L. EagerMOT: 3D multi-object tracking via sensor fusion[C]. IEEE International Conference on Robotics and Automation. Xi'an, 2021: 11315-11321.

[9] Wang Y, Guizilini V, Zhang T Y, et al. DETR3D: 3D object detection from multi-view images via 3D-to-2D queries[C]. Proceedings of the 5th Conference on Robot Learning. London, 2021: 180-191.

[10] Yan J J, Liu Y F, Sun J J, et al. Cross modal transformer: Towards fast and robust 3D object detection[C]. IEEE/CVF International Conference on Computer Vision. Paris, 2023: 18222-18232.

[11] Li P X, Jin J Y. Time3D: End-to-end joint monocular 3D object detection and tracking for autonomous driving[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, 2022: 3875-3884.

[12] Li Y H, Ge Z, Yu G Y, et al. BEVDepth: Acquisition of reliable depth for multi-view 3D object detection[J]. AAAI Technical Track on Computer Vision II, 2023, 37(2): 1-9.

[13] Hussain M I, Azam S, Munir F, et al. Multiple objects tracking using radar for autonomous driving[C]. IEEE International IOT, Electronics and Mechatronics Conference. Vancouver, 2020: 1-4.

[14] Yin T W, Zhou X Y, Krähenbühl P. Center-based 3D object detection and tracking[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, 2021: 11784-11793.

[15] Pang Z Q, Li Z C, Wang N Y. SimpleTrack: Understanding and rethinking 3D multi-object tracking[C]. Computer Vision — ECCV 2022 Workshops: Tel Aviv, Israel. New York, 2022: 680-696.

[16] 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. Seattle, 2020: 6498-6507.

[17] Koh J, Kim J, Yoo J H, et al. Joint 3D object detection and tracking using spatio-temporal representation of camera image and LiDAR point clouds[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2022, 36(1): 1210-1218.

[18] Woo S, Park J, Lee J Y, et al. CBAM: Convolutional block attention module[C]. European Conference on Computer Vision. Munich, 2018: 3-19.

[19] Liu Z, Huang T T, Li B L, et al. EPNet++: Cascade bi-directional fusion for multi-modal 3D object detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(7): 8324-8341.

[20] Shenoi A, Patel M, Gwak J, 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. Las Vegas, 2020: 10335-10342.

[21] Zhang W W, Zhou H, Sun S Y, et al. Robust multi-modality multi-object tracking[C]. IEEE/CVF International Conference on Computer Vision. Seoul, 2019: 2365-2374.

[22] Wang X Y, Fu C Y, Li Z K, et al. DeepFusionMOT: A 3D multi-object tracking framework based on camera-LiDAR fusion with deep association[J]. IEEE Robotics and Automation Letters, 2022, 7(3): 8260-8267.

[23] Zhu Z M, Nie J H, Wu H, et al. MSA-MOT: Multi-stage association for 3D multimodality multi-object tracking[J]. Sensors, 2022, 22(22): 8650.

[24] Wang X Y, Fu C Y, He J W, et al. You only need two detectors to achieve multi-modal 3D multi-object tracking[J/OL]. 2023, arXiv: 2304.08709.

[25] Jiang C, Wang Z L, Liang H W, et al. A fast and high-performance object proposal method for vision sensors: Application to object detection[J]. IEEE Sensors Journal, 2022, 22(10): 9543-9557.

[26] Baser E, Balasubramanian V, Bhattacharyya P, et al. FANTrack: 3D multi-object tracking with feature association network[C]. IEEE Intelligent Vehicles Symposium. Paris, 2019: 1426-1433.

[27] Jin J Y, Zhang J D, Zhang K P, et al. 3D multi-object tracking with boosting data association and improved trajectory management mechanism[J]. Signal Processing, 2024, 218: 109367.

[28] Xu L B, Huang Y P. Rethinking joint detection and embedding for multiobject tracking in multiscenario[J]. IEEE Transactions on Industrial Informatics, 2024, PP(99): 1-10.

[29] Dimitrievski M, Veelaert P, Philips W. Behavioral pedestrian tracking using a camera and LiDAR sensors on a moving vehicle[J]. Sensors, 2019, 19(2): 391.

作者简介

杨浩冉 (1997–), 女, 硕士生, 主要研究方向为计算机视觉、3D多目标跟踪, E-mail: hr18303800515@163.com;

李辉 (1984–), 男, 副教授, 博士, 主要研究方向为计算机视觉、行为识别, E-mail: lihui@qust.edu.cn;

艾晓雪 (2000–), 女, 硕士生, 主要研究方向为计算机视觉、3D目标跟踪, E-mail: axx_edu@163.com;

赵国伟 (1998–), 男, 硕士生, 主要研究方向为计算机视觉、3D多目标跟踪, E-mail: 2035963788@qq.com;

郭颖 (1999–), 女, 硕士生, 主要研究方向为计算机视觉、多目标跟踪, E-mail: guoying_official@163.com.