3D稀疏卷积结构下融合空间点与体素关系建模的LiDAR点云跟踪方法

田胜景[1],韩一男[2],赵宪通[3],刘秀平[2,3],张 明[1]
(1. 中国矿业大学经济管理学院,江苏徐州 221116;2. 大连理工大学白俄罗斯国立大学联合学院,辽宁大连 116024;
3. 大连理工大学数学科学学院,辽宁大连 116024)

摘 要: 稀疏卷积在处理激光雷达点云单目标跟踪时的潜力尚未得到充分发掘. 目前,绝大多数点云跟踪算法使用基于球邻域的骨干网络,其显存计算资源占用大并且目标感知的关系建模不充分. 针对此问题,本文提出一种基于稀疏卷积结构的LiDAR(Lightlaser Detection And Ranging)点云跟踪算法,并创新性地融合了空间点与体素双通道的关系建模模块,以高效适应稀疏框架下目标判别信息的嵌入. 首先,本文采用3D稀疏卷积残差网络来分别提取模板和搜索区域的特征,并利用反卷积来获取逐点特征来保证跟踪任务中对空间位置特性的要求. 其次,关系建模模块进一步在模板与搜索区域特征之间计算相似度语义查询表. 为了捕捉到模板与搜索区域间细粒度的关联性,该模块一方面在空间点通道中利用近邻算法找出每个搜索区域点的模板近邻点,并根据语义查询表提取对应特征;另一方面,在体素通道中以每个搜索区域点为中心构建局部多尺度体素,并根据落入体素单元的模板点索引计算语义查询表中值的累计和. 最后,将双通道的特征融合并送入基于鸟瞰图的候选包围盒生成模块来回归目标包围盒. 为了验证所提出方法的优越性,本文在KITTI和NuScenes数据集进行了测试,对比其他使用稀疏卷积的算法,本文方法平均成功率和精确率分别提升了11.0%和12.0%. 本文方法在继承了稀疏卷积高效特点的同时还实现了跟踪精度的提高.

关键词: 点云理解;目标跟踪;机器视觉;稀疏卷积;特征融合
基金项目: 国家自然科学基金(No.62301562);中国博士后科学基金(No.2023M733756);中央高校基本科研业务费专项资金资助(No.2023QN1055)
中图分类号: TP391.4 **文献标识码:** A **文章编号:** 0372-2112(2024)10-3527-14
电子学报URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20231009

LiDAR Point Cloud Tracking Method Using Point-Voxel Relationship Modeling Under 3D Sparse Convolutional Framework

TIAN Sheng-jing[1], HAN Yi-nan[2], ZHAO Xian-tong[3], LIU Xiu-ping[2,3], ZHANG Ming[1]
(1. *School of Economics and Management*, *China University of Mining and Technology*, *Xuzhou*, *Jiangsu* 221116, *China*;
2. *DUT-BSU Joint Institute*, *Dalian University of Technology*, *Dalian*, *Liaoning* 116024, *China*;
3. *School of Mathematical Sciences*, *Dalian University of Technology*, *Dalian*, *Liaoning* 116024, *China*)

Abstract: The potential of sparse convolution in the field of single target tracking from LiDAR (Lightlaser Detection And Ranging) point cloud has not been fully explored. The vast majority of point cloud tracking algorithms use point-based backbone networks which require higher computation costs and the target-aware relationship modeling is insufficient. To address this problem, this paper proposes a 3D target tracking algorithm based on a sparse convolutional framework, and incorporates it with a point-voxel dual channel relationship modeling module to facilitate the embedding of target discrimination information in the such sparse framework. Firstly, this work uses a 3D convolutional residual network to extract the features of the template and search area separately, then uses deconvolution to obtain pointwise features for the spatial position in tracking tasks. Secondly, the relationship modeling module further calculates a semantic similarity query table based on the above features of the template and the search area. In order to capture the fine-grained correlation, on the one hand, the module utilizes the nearest neighbor algorithm in the spatial point channel to find the template points for each search area point, and extracts corresponding features based on the query table; on the other hand, local multi-scale voxels are constructed

收稿日期:2023-10-27;修回日期:2024-05-19;责任编辑:朱梅玉

with each search area point as the center in the voxel channel, and the accumulated similarity of templates falling into voxel units is used as clues to extract features. Finally, the dual channel feature fusion is sent into the candidate bounding box generation module based on bird's-eye view to estimate the target bounding box. To verify the superiority of the proposed method, we evaluated it on the KITTI and NuScenes datasets, and compared with the baseline algorithm adopting sparse convolution, the mean success and precision rates achieved a considerable improvement of 11.0% and 12.0%. The proposed method not only inherits the efficient characteristics of sparse convolution but also improves tracking accuracy.

Key words: point cloud understanding; object tracking; machine vision; sparse convolution; feature fusion

Foundation Item(s): National Natural Science Foundation of China (No.62301562); China Postdoctoral Science Foundation (No.2023M733756); Fundamental Research Funds for the Central Universities (No.2023QN1055)

1 引言

目标跟踪作为计算机视觉领域中经典的任务之一[1~5],其相关技术被广泛应用于各种领域,如自动驾驶汽车[6]、移动机器人[7]、增强现实[8]等. 随着激光雷达(Light Detection And Ranging,LiDAR)设备的逐渐普及,点云这一数据模态逐渐进入人们的视野. 由于三维数据拥有二维数据不具备的深度信息,且深度信息能有效提高跟踪任务精度[9],因此目标跟踪算法也逐渐从处理传统图像过渡到处理点云[10,11]. 点云目标跟踪又可以根据同时跟踪的目标个数分成单目标跟踪和多目标跟踪,二者有着各自不同的应用场景. 本文聚焦于点云单目标跟踪,旨在计算资源消耗和跟踪性能之间进行有效平衡.

具体来说,点云单目标跟踪任务根据第一帧中指定的任意物体作为目标模板,期望算法能在该模板的引导下从后续帧中估计出三维包围盒来定位目标. 目前主流的点云单目标跟踪算法有两种不同的范式. 一种是以SC3D[11]为代表的基于形状匹配的跟踪方法. 该类方法先在搜索区域中提取出目标的候选点云,再将其对应的候选点云特征与模板特征进行相似度计算,从而估计出目标的状态. 然而这类方法在候选点云生成时依赖概率采样,并且无法进行端到端训练. 另外一种是基于孪生网络的方法[12~15],其克服上述形状匹配中存在的缺陷并在跟踪性能上获得了显著提升[13]. 该类方法首先通过孪生网络计算模板与搜索区域的特征,然后借助特征融合模块将目标信息嵌入到搜索区域当中,最后将融合特征送入候选框预测网络回归出三维包围盒的参数. 当前基于孪生网络的跟踪器在点云目标跟踪领域备受青睐. 这些方法从特征增强[13]、几何尺寸先验[12]、鸟瞰图[14]以及跨模态图像信息[15]等出发,获得了优异的跟踪性能. 然而,上述方法在跟踪过程中却忽略了目标在稀疏和稠密场景中的变化. 随后,研究者们[16,17]受点云Transformer启发,利用模板点和待搜索点之间的交叉注意力来应对不同场景中的变化. 上述孪生网络的框架大都基于PointNet++[18]. 尽管点云跟踪算法基于此框架取得了显著进展,但是其对每个空间点进行球查询并提取特征的过程会占用大量计算资源,在大规模场景高效训练也存在挑战. 此外更重要的是,当前方法在模板和搜索区域特征之间的关系建模有待提升,逐点间的相似匹配仍缺乏对长距关联的考虑.

针对上述所提出的训练资源占用大的问题,本文采用稀疏卷积作为骨干网络来提取特征. 由于稀疏卷积只对体素网格中非空单元进行处理,因此能够避免冗余计算从而大大降低网络计算资源消耗. 稀疏卷积已经在目标检测[19]、分割[20]、识别[21]等领域取得了显著成功,但就目前所知,基于稀疏卷积的点云目标跟踪方法暂没有受到广泛关注. LTTR跟踪器[22]虽然基于P2B框架[13]进行了初步探索,其重心在于帧间注意力的变化,并且仍依赖于空间点的直接投票和聚类的方式来生成候选包围盒. 这种方式状态估计空间大且计算资源昂贵. 鉴于此,为了充分挖掘稀疏卷积在目标跟踪领域的潜力,本文采用端到端的卷积响应图估计来进行点云单目标跟踪. 一方面,其采用3D稀疏残差网络来提取模板和搜索区域的特征,并利用反卷积计算逐点特征以保证跟踪任务中对空间位置特性的要求;另一方面,其从鸟瞰图视角压缩3D卷积特征并把每一个特征点作为锚点(anchor)来预测前景响应图和对应偏移量,这种基于锚点的方式相比于上述的直接投票聚类降低了状态估计难度.

针对模板和搜索区域特征关系建模不充分的问题,本文提出了一种融合空间点和体素双通道的关系建模模块来适配稀疏卷积框架. 具体来说,基于空间点通道的关联方式能够获取到模板和搜索区域之间局部细节的关系,而基于体素通道的关联方式则更擅长够捕捉两者间长距的关联和多尺度特征[23,24]. 考虑到点云跟踪误差累积导致漂移,本文设计的双通道融合方式既能提取局部细节又能捕捉目标间的长距信息. 在KITTI[25]和nuScenes[26]数据集上的测试结果证明,本文所提方法显著超越了当前基准方法,有效节约了计算资源,并提升了跟踪的效率.

总结来说,本文的主要贡献如下:

(1)提出了一种基于3D稀疏卷积的响应图估计方法来进行点云单目标跟踪,有效降低了训练资源消耗.

不同于以往仅在三维空间中利用散点投票和聚类的候选生成方式,其完成了从特征提取和候选生成过程按照卷积响应图的方式进行候选目标回归,避免使用基于球邻域的大量资源占用.

(2)提出了融合点与体素双通道的关系建模模块.该模块能较好地与稀疏卷积框架相耦合,体素通道能够感知长距信息,空间点通道能够捕捉局部结构特征,二者结合能有效提高跟踪精度.

(3)在两个主流数据集上验证了所提方法的优越性.对比同样使用稀疏卷积的基准算法,其平均成功率和精确率分别高出11.0%和12.0%,获得显著提升.

2　相关研究工作

2.1　稀疏卷积

卷积神经网络已经被证明能够较好的处理二维图像,但是对于三维数据来说,维度的增加会使卷积的计算量几何倍地增加.此外,与二维图像不同,三维空间中会存在大量的空白区域,这使三维的体素数据往往是稀疏的,使用传统卷积方法进行处理会造成大量无意义的冗余计算,造成计算资源的浪费.针对这一问题,Facebook公司的Graham等人[20]率先提出了子流形稀疏卷积,用数据结构保存非空体素,只按索引对其进行非空卷积.英伟达公司Choy等人[21]提出MinkowskiEngine来处理时空四维稀疏数据.这些方法已被应用于多个点云场景任务[27](如配准、分割、特征匹配),是进行端到端稀疏卷积的开端.由于稀疏卷积在处理三维数据时拥有资源占用低、速度快的优势,目前被广泛应用于各类室外LiDAR感知任务中.

在特征匹配领域,基于稀疏卷积的网络已经在速度和精度显著超越诸如PPFNet、Spin Image、PointNet等方法[27].在目标检测领域,基于PointNet++框架的方法[28,29]的表现也已渐渐被稀疏卷积超越[19,23,30].其中最具代表性的稀疏卷积工作包括SECOND[19]和PV-RCNN[23].前者使用子流形稀疏卷积处理体素化的场景并在特征图上直接使用锚点来回归包围盒,后者结合稀疏卷积和原始点来辅助提取多尺度特征.这种思路的额外优势是其可以通过鸟瞰图转换来借鉴二维图像检测任务的经验策略.在该任务中,稀疏卷积的引入使得它们无论是计算资源消耗还是性能表现都大幅领先其他方法,并且已成为评价点云目标检测的基准.尽管基于稀疏卷积的方法已在多个领域崭露头角,但在点云跟踪领域未被充分研究.当前,基于稀疏卷积的目标跟踪算法仅有LTTR[22].该方法借助深度霍夫投票进行空间点聚类,以此完成候选框的生成,其没有充分发挥基于卷积特征响应图的包围盒优势,且缺乏对模板和搜索区域进行有效的关系建模.因此,相较于前沿跟踪方法,稀疏卷积的跟踪框架仍有巨大潜力.这一发现引起我们的重新思考,并以此为切入点给出了适应稀疏卷积的目标跟踪方法.

2.2　点云目标跟踪

SC3D[11]是这一领域的开创性工作,该方法先在搜索区域中以人工经验的方式选取一些候选区域与模板点云一起送入孪生编码器提取深度特征,然后通过比较候补区域特征与模板特征的相似性来确定目标.Fang等人[31]受二维视频跟踪启发提出了一种3DSiam-RPN跟踪器,但其存在模板和搜索区域特征对齐问题.P2B[13]是继SC3D之后又一里程碑式的工作,也是基于特征融合的目标跟踪范式的开创者,它使用孪生的PointNet++[18]对模板和搜索区域进行编码和降采样,然后使用目标特征增强模块将模板信息嵌入到搜索区域中,最后使用基于深度霍夫投票[32]的检测头回归最合适的包围盒.P2B解决了SC3D无法端到端训练以及候补目标的自适应选取问题,但该方法仍然存在以下缺点.首先,它对搜索区域的下采样点进行操作,这可能会忽略目标对象的细粒度特征.其次,所生成的包围盒容易受到球内聚合点的扰动影响,进而导致跟踪结果的波动.之后研究者对P2B进行了许多优化.BAT[12]将包围盒的几何尺寸信息与模板信息一同嵌入到搜索区域中.MLVSNet[33]对特征进行多尺度处理以适应点云的稀疏性.DSDM[34]则提出基于多种子的监督下降方法来优化包围盒.SA-P2B[35]引入结构信息来辅助目标对象的细粒度特征提取.PTTR[16]则提出了一种目标感知的下采样方式来提升目标跟踪精度.这些方法虽然都在P2B的基础上获得了显著提升,但是在目标判别性信息的嵌入方面还需要进一步的改进和研究.

随着研究逐渐深入,研究者们进一步探索了结合Transformer[17,22]、图网络[36]以及BEV鸟瞰图[14]的思想来提高性能.具体地,PTT[17]方法在特征融合后使用注意力机制进行关系建模.GPT[36]则通过构建模板点云和搜索区域点云的连接关系,并利用图网络进行前景与背景之间的信息传递.V2B[14]发现用搜索区域的鸟瞰图进行候选包围盒的预测比基于霍夫投票的任务头更精准.该方法通过特征体素化和最大池化将融合特征投影到鸟瞰图上,然后将鸟瞰图送入二维卷积层回归目标包围盒.LTTR[22]则基于稀疏卷积先将模板和搜索区域特征分别投影到鸟瞰图上,然后设计Transformer结构让二者进行完成特征的交互融合.然而,上述跟踪方法单纯地使用模板和搜索区域的几何外观特征,忽略了目标时间上的运动信息.为了继续提高跟踪精度,一些方法则利用帧间的运动增量进行建模.最近,MMTrack[37]将两帧点云进行拼接并在特征维度引入时间信息,最终通过预测相邻帧间的相对运动来辅

助包围盒的生成. 该类方法会预估目标的运动轨迹并以此为辅助帮助定位目标,取得了更高的跟踪精度. 然而,该方法跟踪结果的好坏很大程度依赖于特定类别(车或人)运动线索估计的精度,若测试与训练时的运动轨迹或模式出现较大偏差,那么跟踪任务也会失败.

上述绝大部分方法都利用PointNet++网络提取特征,并针对性地设计了模板和搜索区域的关系建模方式. 受点云目标检测启发,我们发现基于稀疏卷积的方法除了资源占用低外,还可以更直接借助以往二维跟踪的先验来提高点云跟踪的性能. 尽管当前稀疏卷积的单目标跟踪研究仍有巨大的探索空间,但是现有稀疏卷积框架应用至该领域的主要障碍是缺少与之相适配的关系建模. 鉴于此,本文研究了以稀疏卷积为骨干网络的目标跟踪框架,并给出了融合空间点和体素关系建模的方法.

3 本文算法框架

点云单目标跟踪任务是指在目标模板引导下从搜索区域中定位感兴趣的目标. 具体来说,该任务是以模板点云 $\boldsymbol{P}_{\mathrm{t}} \in \mathbb{R}^{N_{\mathrm{t}} \times 3}$ 和搜索区域点云 $\boldsymbol{P}_{\mathrm{s}} \in \mathbb{R}^{N_{\mathrm{s}} \times 3}$ 为输入(其行向量代表空间点的三维坐标,此处分别记作 $\boldsymbol{p}_{\mathrm{t}}^{i}$ 和 $\boldsymbol{p}_{\mathrm{s}}^{i}$),最终输出待跟踪目标的包围盒信息 $\boldsymbol{B}=(x,y,z,l,w,h,\theta)$ 的过程. 其中,(x,y,z) 是包围盒中心点坐标,(l,w,h) 是包围盒的长宽高,θ 是包围盒在平面上的偏航角度.

为了提高跟踪性能并且降低点云训练的计算资源占用,本文基于MinkowskiEngine[21]给出了以稀疏卷积为骨干网络的目标跟踪算法,总体框架参见图1,具体分解为三个部分:卷积残差网络的特征提取、空间点与体素双通道的关系建模以及候选包围盒生成模块. 首先,对模板和搜索区域点云进行体素化,将其送入孪生的稀疏卷积和反卷积层提取特征,并将该稀疏卷积特征赋予对应点得到逐点特征. 然后,将模板特征和搜索区域特征送入双通道关系建模模块,通过模板特征嵌入的方式来获取目标感知的判别性特征. 最后,将融合特征利用鸟瞰图视角压缩送入候选包围盒生成模块来回归包围盒信息.

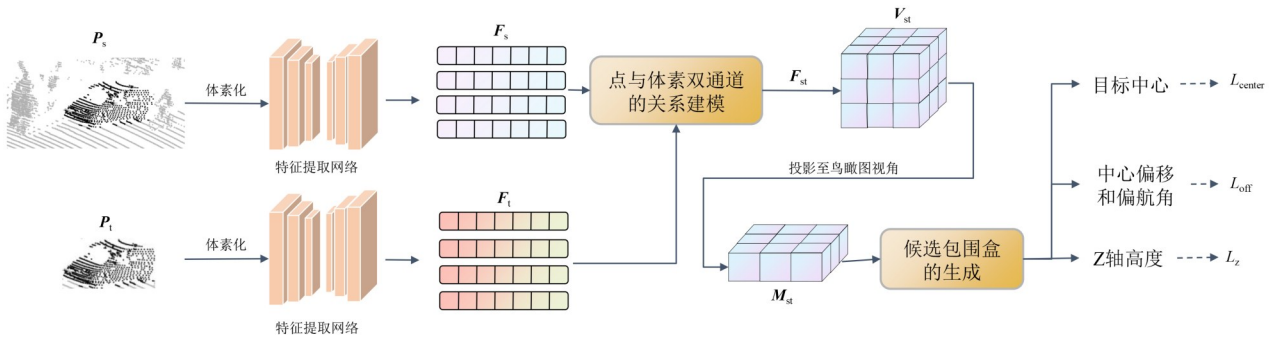

图1 方法整体流程图.

3.1 特征提取模块

稀疏卷积凭借其优秀的特征编码能力和高效的计算方式被广泛应用于各类任务中,如目标检测、特征匹配、语义分割等. 在目标跟踪领域,基于原始点的特征编码方式是当前的主要趋势. 但相比之下,稀疏卷积可以有效编码多尺度特征并有助于产生高质量的锚点[23]. 为了深入挖掘稀疏卷积在目标跟踪领域的潜力,本文采用3D稀疏卷积残差网络来提取模板和搜索区域的特征,并利用反卷积来获取逐点特征,以保证跟踪任务中对空间位置特性的要求. 具体结构如图2所示,其中,括号内依次为卷积核大小、步长以及特征输出维度.

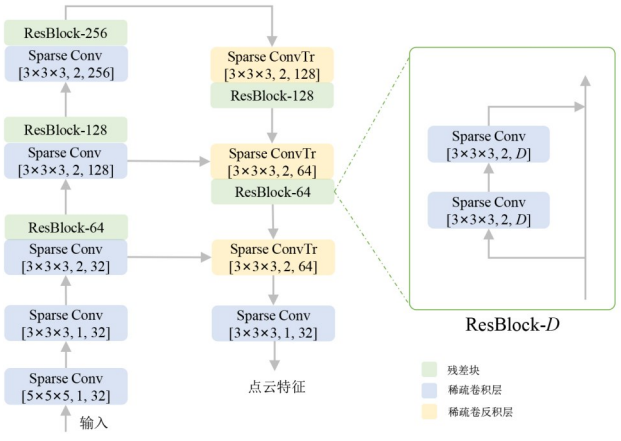

图2 稀疏卷积网络结构示意图

首先,本文对模板点云 $\boldsymbol{P}_{\mathrm{t}}$ 和搜索区域点云 $\boldsymbol{P}_{\mathrm{s}}$ 按照一定步长进行体素化,划分成为分辨率为 $W_{\mathrm{t}} \times L_{\mathrm{t}} \times H_{\mathrm{t}}$ 和 $W_{\mathrm{s}} \times L_{\mathrm{s}} \times H_{\mathrm{s}}$ 的体素,并记录各点所属的体素单元索引. 其次,将模板和搜索区域的体素送入孪生稀疏卷积结构(图2)抽取特征,同时将特征图分辨率降至 $\frac{W_{\mathrm{t}}}{8} \times \frac{L_{\mathrm{t}}}{8} \times \frac{H_{\mathrm{t}}}{8}$ 和 $\frac{W_{\mathrm{s}}}{8} \times \frac{L_{\mathrm{s}}}{8} \times \frac{H_{\mathrm{s}}}{8}$. 借助UNet结构再将抽取的

特征通过反卷积将分辨率还原. 最后,由于原始空间点与体素单元之间的对应关系未变,本文将体素单元的特征重新赋予对应点作为该点的特征. 在这一过程中,若同一体素单元中有多个点则多个点的特征相同. 当分辨率足够大时(即体素单元个数设置较多时),该过程可以使每个非空体素单元中仅有一个LiDAR点,从而降低几何信息的损失. 对于传统卷积方法来说分辨率越大计算负担越大,但稀疏卷积只将非空值单元纳入计算的特性可以避免该问题. 将上述编码过程记作φ_θ,则可以通过该编码器获得模板特征和搜索区域特征:

$$\boldsymbol{F}_{\mathrm{t}}=\varphi_\theta\left(\boldsymbol{P}_{\mathrm{t}}\right)\in\mathbb{R}^{N_{\mathrm{t}}\times D} \tag{1}$$

$$\boldsymbol{F}_{\mathrm{s}}=\varphi_\theta\left(\boldsymbol{P}_{\mathrm{s}}\right)\in\mathbb{R}^{N_{\mathrm{s}}\times D} \tag{2}$$

其中,D是特征维度.

3.2　融合空间点与体素双通道的关系建模

稀疏卷积虽然具有较好的计算特性,但是仍需面向特定的任务设计合适的关系建模模块. 针对点云跟踪问题,为了更好地建模目标模板与搜索区域之间的相似性关系,本文提出了一个新颖的空间点与体素双通道的关系建模模块. 该模块以模板点云和搜索区域点云的三维坐标和特征向量作为输入,其作用在于将目标的判别信息嵌入到搜索区域之中,如图3所示.

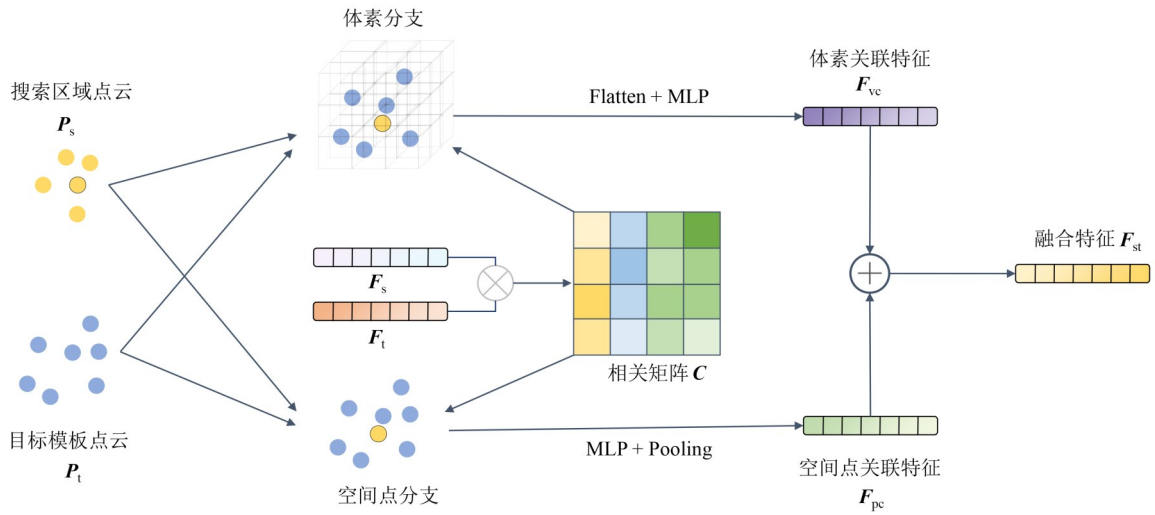

图3　空间点与体素双通道的关系建模示意图.

首先,根据模板与搜索区域的稀疏卷积特征$\boldsymbol{F}_{\mathrm{t}}$和$\boldsymbol{F}_{\mathrm{s}}$计算二者之间的相关矩阵$\boldsymbol{C}$,将其作为后续关联操作的特征查询表. 然后,基于相关矩阵$\boldsymbol{C}$,空间点通道和体素通道利用各自设计的规则分别得到关联特征$\boldsymbol{F}_{\mathrm{pc}}$和$\boldsymbol{F}_{\mathrm{vc}}$. 其中,空间点通道使用K近邻算法寻找临近点并捕捉局部特征,体素通道使用中心体素化捕捉长距关联. 最后,通过MLP将双通道特征维度对齐,二者特征相加后即得到了蕴含模板信息的融合特征$\boldsymbol{F}_{\mathrm{st}}$. 融合后的特征既包含了空间散点捕捉到的局部几何信息,同时又包含了多尺度体素对长距的感知,将会进一步送入到下游候选包围盒生成模块.

在后续章节单独介绍每个通道的工作原理之前,此处先计算模板特征和搜索区域特征的相关矩阵:

$$\boldsymbol{C}=\boldsymbol{F}_{\mathrm{s}}\boldsymbol{F}_{\mathrm{t}}^{\mathrm{T}}=\left[c_{ij}\right]\in\mathbb{R}^{N_{\mathrm{s}}\times N_{\mathrm{t}}} \tag{3}$$

其中,c_{ij}为相关矩阵$\boldsymbol{C}$第i行第j列的元素,表示搜索区域第i个点的特征与模板第j个点的特征之间的相似度. 与之前方法不同,计算得到的相关矩阵$\boldsymbol{C}$不会直接与搜索区域特征发生交互而是作为一个特征查询表参与后续环节,具体做法见第3.2.1节和第3.2.2节.

3.2.1　基于空间点的通道

基于空间点的关联方式能够获取模板和搜索区域之间局部特征的相似关系. 在这一分支中,首先需要确定模板点与搜索区域点之间的位置关系. 为此,本文先将经过中心变换的模板点云与搜索区域点云放置到同一坐标系下:

$$\boldsymbol{P}=\mathrm{concat}\left(\boldsymbol{P}_{\mathrm{t}},\boldsymbol{P}_{\mathrm{s}}\right)=\begin{bmatrix}\boldsymbol{P}_{\mathrm{t}}\\ \boldsymbol{P}_{\mathrm{s}}\end{bmatrix}\in\mathbb{R}^{\left(N_{\mathrm{t}}+N_{\mathrm{s}}\right)\times 3} \tag{4}$$

其中,$\mathrm{concat}(\cdot,\cdot)$表示矩阵的按行拼接.

获得融合点云后,为了进一步确定搜索区域中的点与模板中哪些点关系密切,本文采用K近邻策略去为搜索区域$\boldsymbol{P}_{\mathrm{s}}$中的每个点检索其在模板点$\boldsymbol{P}_{\mathrm{t}}$中的近邻,并将$\boldsymbol{p}_{\mathrm{s}}^{i}$的K近邻点集合记为$N_K^i=N\left(\boldsymbol{p}_{\mathrm{s}}^{i}\right)\in\mathbb{R}^{3\times K}$. 通过近邻点集合中对应点的索引$d\left(N_K^i\right)$,该通道可以在相关矩阵$\boldsymbol{C}$中提取出近邻点集$N_K^i$与搜索区域点集之间的矩阵元素值,记作$\boldsymbol{C}_K^i=\boldsymbol{C}\left(d\left(N_K^i\right)\right)\in\mathbb{R}^{1\times K}$. 然后通过下述处理可得到空间点$\boldsymbol{p}_{\mathrm{s}}^{i}$的关联特征:

$$\boldsymbol{f}_{\mathrm{pc}}^{i}=\mathrm{pooling}\left(\mathrm{MLP}\left(\mathrm{concat}\left(\boldsymbol{C}_{K}^{i},N_{K}^{i}-\boldsymbol{p}_{\mathrm{s}}^{i}\right)\right)\right)\in\mathbb{R}^{D\times 1}\tag{5}$$

其中，pooling(·)表示最大池化，MLP(·)表示多层感知机. 最后对搜索区域的所有点按式(5)计算对应关联特征，即可得到空间点关联特征矩阵：

$$\boldsymbol{F}_{\mathrm{pc}}=\left[\boldsymbol{f}_{\mathrm{pc}}^{1}\quad\boldsymbol{f}_{\mathrm{pc}}^{2}\quad\cdots\quad\boldsymbol{f}_{\mathrm{pc}}^{N_{\mathrm{s}}}\right]^{\mathrm{T}}\in\mathbb{R}^{N_{\mathrm{s}}\times D}\tag{6}$$

基于空间点的通道能提取到模板与搜索区域之间细粒度的空间分布相似信息. 虽然这些局部信息的利用能够获取模板与搜索区域之间的局部相关性，但其固定了近邻数量，容易纳入干扰点，缺少对场景不同距离尺度的控制. 为了解决上述问题，本文在点通道的基础上额外增加了体素通道，旨在同时获得模板与搜索区域间的长距与局部相似性匹配.

3.2.2 基于体素的通道

与以往的直接将模板和搜索区域整体进行体素化处理的方式不同，本文方法则以搜索区域中的点为中心，在式(4)中的融合点云$\boldsymbol{P}$上进行局部体素化. 具体来说，在计算搜索区域点$\boldsymbol{p}_{\mathrm{s}}^{i}$与模板间的关联特征时，首先以$\boldsymbol{p}_{\mathrm{s}}^{i}$为中心构建体素边长为$r$、分辨率为$a\times a\times a$的局部体素块：

$$V_{r,a}^{i}=\left\{V_{(\zeta)}^{i}\middle|\,\zeta\in\mathbb{Z}^{3}\right\}\tag{7}$$

$$V_{(\zeta)}^{i}=\left\{\boldsymbol{p}_{\mathrm{s}}^{i}+\zeta\times r+\triangle r\middle|\,\|\triangle r\|_{1}\leqslant\frac{r}{2}\right\}\tag{8}$$

其中，$\zeta=[l,h,w]$，$\left\lceil-\frac{a}{2}\right\rceil\leqslant l,h,w\leqslant\left\lfloor\frac{a}{2}\right\rfloor\in\mathbb{Z}$表示局部体素单元的索引；$V_{(\zeta)}^{i}$是对应索引下的单元所占据的坐标连续空间.

随后，可以确定场景内落在体素单元$V_{(\zeta)}^{i}$中的模板点集，记为$\boldsymbol{P}_{(\zeta)}^{i}=\left\{\boldsymbol{p}_{\mathrm{t}}^{j_1},\boldsymbol{p}_{\mathrm{t}}^{j_2},\boldsymbol{p}_{\mathrm{t}}^{j_3},\cdots,\boldsymbol{p}_{\mathrm{t}}^{j_{n_\zeta}}\right\}$，其中，$n_\zeta$表示落在体素单元$N_{(\zeta)}^{i}$中的模板点的数量. 进一步地，根据相关矩阵$\boldsymbol{C}$和模板点索引$j_{n_\zeta}$，此模块可以查询出当前体素中心$\boldsymbol{p}_{\mathrm{s}}^{i}$内的所有相关值，将其记作$C_{(\zeta)}^{i}=\left\{c_{i,j_1},c_{i,j_2},c_{i,j_3},\cdots,c_{i,j_{n_\zeta}}\right\}$. 本文将此$n_\zeta$个相关值的累加和作为当前体素单元的特征. 若单元中没有模板点(即$n_\zeta=0$)，则规定该体素单元值为0. 最后，将整个体素块依序展开得到长度为a^3的特征向量，经过MLP特征变换，得到搜索区域点$\boldsymbol{p}_{\mathrm{s}}^{i}$在体素通道对应的关联特征：

$$\boldsymbol{f}_{\mathrm{vc}}^{i}=\mathrm{MLP}\left(\mathrm{flatten}\left(\left\{\sum_{j=j_1}^{j_{n_\zeta}}c_{i,j}\middle|\,\zeta\in\mathbb{Z}^{3}\right\}\right)\right)\in\mathbb{R}^{D\times 1}\tag{9}$$

其中，flatten(·)操作表示将体素块按顺序展开成向量. 与点通道处理方式相同，对搜索区域中的所有点计算关联特征即可得到体素通道的关联特征矩阵：

$$\boldsymbol{F}_{\mathrm{vc}}=\left[\boldsymbol{f}_{\mathrm{vc}}^{1}\quad\boldsymbol{f}_{\mathrm{vc}}^{2}\quad\cdots\quad\boldsymbol{f}_{\mathrm{vc}}^{N_{\mathrm{s}}}\right]^{\mathrm{T}}\in\mathbb{R}^{N_{\mathrm{s}}\times D}\tag{10}$$

因为当r和a取值一定时体素单元占据固定空间，即在固定位置寻找模板点，较大的值设置能够覆盖较广的区域，所以基于体素的分支能够帮助捕捉到长距的关联特征，并与点分支形成互补. 此外为了获得感知不同尺度目标的能力，受体素特征常用的金字塔结构启发，本文通过设置多个体素边长r得到不同尺度下的联合特征，再将不同尺度的特征拼接送入多层感知机调整维度获得最终的多尺度的体素关联特征.

在体素通道和空间点通道都完成了上述模板信息嵌入以后，二者特征相加后即得到最终的融合特征，记作$\boldsymbol{F}_{\mathrm{st}}=\boldsymbol{F}_{\mathrm{vc}}+\boldsymbol{F}_{\mathrm{pc}}$.

3.3 候选包围盒估计

通过上述关系建模得到融合特征之后，本文继而考虑在稠密的特征响应图上进行包围盒估计. 具体而言，先将融合特征$\boldsymbol{F}_{\mathrm{st}}\in\mathbb{R}^{N_{\mathrm{s}}\times D}$转化为规则体素$\boldsymbol{V}_{\mathrm{st}}\in\mathbb{R}^{L\times W\times H\times D}$，然后通过最大池化操作将其转换成鸟瞰特征图$\boldsymbol{M}_{\mathrm{st}}\in\mathbb{R}^{L\times W\times H\times D_M}$. 此处$D_{\boldsymbol{M}}$是鸟瞰图的特征通道数，$L=\left\lfloor\frac{x_{\max}-x_{\min}}{v}\right\rfloor+1$和$W=\left\lfloor\frac{y_{\max}-y_{\min}}{v}\right\rfloor+1$是鸟瞰图的长和宽，$(x_{\max},x_{\min})$和$(y_{\max},y_{\min})$是搜索区域在$x$轴方向和$y$轴方向的坐标最大与最小值，$v$是边长. 另外，为了后续损失函数的计算，本文将目标中心在$\boldsymbol{M}_{\mathrm{st}}$中的相对位置表示为$\boldsymbol{c}=(x_c,y_c)$，像素索引表示为$\tilde{\boldsymbol{c}}=\left(\lfloor x_c\rfloor,\lfloor y_c\rfloor\right)$，其中，$x_c=\frac{x-x_{\min}}{v}$，$y_c=\frac{y-y_{\min}}{v}$.

进一步，本文将鸟瞰特征图$\boldsymbol{M}_{\mathrm{st}}$送入二维卷积层中，在保持分辨率不变的情况下实现目标特征聚合. 该过程首先通过连续的卷积操作实现特征重分配，以便能够将鸟瞰图中丰富的信息聚焦到要跟踪的目标上. 最后，本文的损失函数通过约束全卷积回归出的包围盒中心、高度以及偏航角来训练网络. 具体地，损失函数的计算其由L_{center}、L_{off}和L_{z}三个方面组成. 下面分别对其进行阐述.

L_{center}表示目标中心损失. 本文用特征响应图的峰值来表示目标中心位置，并借助改进的交叉熵损失Focal Loss[13]来进行约束：

$$\begin{aligned}L_{\mathrm{center}}=-\sum_{i,j}&\left(\sigma\left(\boldsymbol{\Lambda}_{ij}\right)\cdot\left(1-\boldsymbol{\Lambda}_{ij}\right)^{\alpha}\log\left(\tilde{\boldsymbol{\Lambda}}_{ij}\right)\right.\\&\left.+\left[1-\sigma\left(\boldsymbol{\Lambda}_{ij}\right)\right]\cdot\left(1-\boldsymbol{\Lambda}_{ij}\right)^{\beta}\boldsymbol{\Lambda}_{ij}^{\alpha}\log\left(1-\tilde{\boldsymbol{\Lambda}}_{ij}\right)\right)\end{aligned}\tag{11}$$

其中，α和β分别为Focal Loss中困难样本和负样本的权重调节因子，$\boldsymbol{\Lambda}\in\mathbb{R}^{H\times W\times 1}$是目标所在位置的真值响应图，$\tilde{\boldsymbol{\Lambda}}\in\mathbb{R}^{H\times W\times 1}$是神经网络预测出的目标中心响应图，$\sigma\left(\boldsymbol{\Lambda}_{ij}\right)$为示性函数：

$$\sigma\left(\boldsymbol{\Lambda}_{ij}\right)=\begin{cases}1,&\boldsymbol{\Lambda}_{ij}=1\\0,&\boldsymbol{\Lambda}_{ij}\neq 1\end{cases}\tag{12}$$

L_{off}是中心偏移和旋转损失:

$$L_{\text{off}}=\sum_{i,j}\sigma(i,j)\cdot\left\|\tilde{\boldsymbol{\Gamma}}_{ij}-\boldsymbol{\Gamma}_{ij}\right\|_1 \tag{13}$$

其中,$\boldsymbol{\Gamma}\in\mathbb{R}^{H\times W\times 3}$是中心偏移量(即$\boldsymbol{c}-\tilde{\boldsymbol{c}}$)与偏航角的真值,$\tilde{\boldsymbol{\Gamma}}\in\mathbb{R}^{H\times W\times 3}$则表示中心偏移量与偏航角的估计值.$\sigma(i,j)$为示性函数:

$$\sigma(i,j)=\begin{cases}1, & \left\|(i,j)-\tilde{\boldsymbol{c}}\right\|_2<\delta\\ 0, & \text{otherwise}\end{cases} \tag{14}$$

表示只在目标中心$\tilde{\boldsymbol{c}}$的δ邻域内进行损失约束.

L_z是目标在z轴高度上的损失:

$$L_z=\sum_{i,j}\sigma(i,j)\cdot\left\|\tilde{\boldsymbol{Z}}_{ij}-\boldsymbol{Z}_{ij}\right\|_1 \tag{15}$$

其中,$\tilde{\boldsymbol{Z}}\in\mathbb{R}^{H\times W\times 3}$是目标在$z$轴上的高度预测值,$\boldsymbol{Z}$是目标的真实高度.

最终总损失函数为

$$L_{\text{total}}=\lambda_1 L_{\text{center}}+\lambda_2 L_{\text{off}}+\lambda_3 L_z \tag{16}$$

其中,λ_1,λ_2和λ_3是权重参数,用于平衡损失函数数值大小.

4　实验

本文实验运行环境为Ubuntu 20.04,CUDA 11.7和PyTorch 1.8.1,显卡设备为2块GeForce RTX 3090. 为了全面地评估所提出方法的有效性,本文在当前主流的点云跟踪数据集上进行实验,并分别给出了与其他方法的对比结果、自身的消融实验结果以及跟踪可视化结果.

4.1　实验设置

4.1.1　数据集

为了公平比较,本文与以往方法保持一致,沿用点云跟踪领域主流的大规模数据集:KITTI Tracking[25]和NuScenes[26].

KITTI Tracking是目前广为应用的三维激光雷达点云数据集之一. 它由64线激光雷达扫描而来,包含了21个室外扫描场景的808个目标实例. 前16个场景用于训练,其他用于验证和测试.

NuScenes拥有更大的数据规模,由32线激光雷达扫面得到,包含横跨23个对象类的1 000个驾驶场景,场景稀疏更具挑战性. 其中,700个场景是训练集,验证集和测试集各150个场景. 为了进行目标跟踪,使用训练集中的"train_track"子集进行训练.

4.1.2　性能评价指标

本文采用基于中心位置误差和包围盒重叠率的评价指标[38]. 具体地,中心位置误差是指预测的3D包围盒$\boldsymbol{B}=(x,y,z,l,w,h,\theta)$与对应的真值包围盒$\boldsymbol{B}^*$之间的中心距离:

$$e_{\text{center}}=\sqrt{\left(x-x^*\right)^2+\left(y-y^*\right)^2+\left(y-y^*\right)^2}$$

包围盒重叠率指的是预测的3D包围盒与对应的真值包围盒交集的体积交并比:

$$e_{\text{iou}}=\frac{\text{Volume}\left(\boldsymbol{B}\cap\boldsymbol{B}^*\right)}{\text{Volume}\left(\boldsymbol{B}\cup\boldsymbol{B}^*\right)}$$

基于这两个指标,评估算法的做法为:给定某一阈值,去统计测试帧中超过这一阈值的所占比例. 具体来说,以阈值为横轴,超过包围盒重叠率阈值的帧数占比为纵轴,所围成的曲线面积作为成功率指标σ_{succ}. 同理,以阈值为横轴,超过中心位置误差阈值的帧数占比作为纵轴,所围成的曲线面积作为精确度指标σ_{prec}.

4.2　实验结果对比

按照特征提取方式不同,当前点云跟踪方法可以分为基于稀疏卷积(sparse convolution)的方法和基于原生点云(PointNet++)的方法. 前者目前仅有LTTR[22],后者代表包括P2B[13]、BAT[12]和V2B[14]等. 受稀疏卷积在其他任务(如3Ddetection)的优异性能推动,本文在稀疏卷积框架下提出了一种点与体素融合的关系建模方法(记为PVT),来发挥其在跟踪任务上的潜在性能. 为了说明所提方法的有效性,本节首先将PVT与其基准方法LTTR对比,然后再与其他前沿方法进行对比.

4.2.1　与LTTR方法的跟踪结果对比

LTTR方法虽然在稀疏卷积框架下提取特征,但在后续关系建模时其将特征投影至2D鸟瞰图视角,并利用自注意力机制在目标模板与搜索区域之间进行信息传递. 这种投影的关系建模方式无疑会造成三维空间信息的丢失. 不同于此,本文基于点与体素双分支直接在空间中进行关系建模,充分考虑空间和语义信息. 二者的对比结果如表1所示.

由表1实验结果可知,根据成功率σ_{succ}和精确率σ_{prec},本文方法PVT全面超越了当前基准方法LTTR,分

表1　与当前稀疏卷积方法在成功率和精确率的比较结果　　单位:%

特征提取方式	方法	数据形式	Car (6 424)	Pedestrian (6 088)	Van (1 248)	Cyclist (308)	Mean (14 068)
稀疏卷积	LTTR[22]	LiDAR	65.0 / 77.1	33.2 / 56.8	35.8 / 45.6	66.2 / 89.9	48.7 / 65.8
	PVT(本文)	LiDAR	68.4 / 81.8	50.7 / 77.2	56.1 / 66.9	69.0 / 92.8	59.7 / 78.7
			↑ 3.4 / 4.7	↑ 17.5 /20.4	↑ 20.3 / 21.3	↑ 2.8 / 2.9	↑ 11.0 / 12.9

注:"/"前部表示成功率,后部表示精确率.下划线标注了每列最优结果.

别平均指标实现了11.0%和12.9%的显著提升. 特别地,在场景中常见物体行人(Pedestrian)类别上,本文方法在成功率和准确率指标上比LTTR大幅提升了17.5%和20.4%. 其主要原因是LTTR关系建模时的鸟瞰图投影对于这种尺寸较小的目标容易造成空间信息丢失,而本文方法同时从原生点查找K近邻和局部坐标系构建体素的设计能够很大程度缓解该问题.

4.2.2 与其他方法的跟踪结果对比

除了与稀疏卷积的方法对比外,本文还对比了点云跟踪领域的其他12种方法,其中包括发表于CVPR、AAAI、ICCV和NeurIPS会议上的P2B[13]、BAT[12]、GPT[36]、PTT[17]、PTTR[16]和V2B[14]等前沿方法.

KITTI数据集的结果如表2所示. 总体来说,在上述所有的方法中,本文方法在所有类别的平均成功率和精确率上获得了最好的结果,分别达到了59.7%和78.7%,具有竞争力. 具体来说,在汽车的精确率指标上,PTT和本文方法PVT获得了最佳性能(81.8%);而在汽车的成功率指标上,V2B和PVT分别为70.5%和68.4%,位于前两名. 在行人类别上,PTTR和本文方法则在成功率指标上旗鼓相当,分别为50.9%和50.7%;而其在精确率指标上,PTTR比PVT高了4.4%. 分析背后原因可能在于,PTTR中基于语义相似性的种子点采样保留了较小目标的前景信息,有利于后续状态回归. 未来基于稀疏卷积的方法也可考虑引入前景掩码的策略来进一步提升跟踪性能.

表2 KITTI上与其他前沿方法在成功率和精确率的比较结果 单位:%

特征提取方式	方法	数据形式	Car (6 424)	Pedestrian (6 088)	Van (1 248)	Cyclist (308)	Mean (14 068)
原生点云	SC3D[11]	LiDAR	41.3 / 57.9	18.2 / 37.8	40.4 / 47.0	41.5 / 70.4	31.2 / 48.5
	P2B[13]	LiDAR	56.2 / 72.8	28.7 / 49.6	40.8 / 48.4	32.1 / 44.7	42.4 / 60.0
	SA-P2B[35]	LiDAR	58.0 / 75.1	34.6 / 63.3	51.2 / 63.1	32.0 / 43.6	46.7 / 68.2
	3DSiamRPN[31]	LiDAR	58.2 / 76.2	35.2 / 56.2	45.7 / 52.9	36.2 / 49.0	46.6 / 64.9
	PTT[17]	LiDAR	67.8 / 81.8	44.9 / 72.0	43.6 / 52.5	37.2 / 47.3	55.1 / 74.2
	V2B[14]	LiDAR	70.5 / 81.3	48.3 / 73.5	50.1 / 58.0	40.8 /49.7	58.4 / 75.2
	DSDM[34]	LiDAR	65.1 / 75.8	32.5 / 51.4	58.9 / 67.9	37.4 / 46.5	49.8 / 63.9
	PTTR[16]	LiDAR	65.2 / 77.4	50.9 / 81.6	52.5 / 61.8	65.1 / 90.5	57.9 / 78.1
	GPT[36]	LiDAR	59.1 / 75.6	35.2 / 63/6	49.6 / 60.6	34.3 / 46.3	47.4 / 68.4
	MLVSNet[33]	LiDAR	56.0 / 74.0	34.1 / 61.1	52.0 / 61.4	34.3 / 44.5	45.7 / 66.6
	BAT[12]	LiDAR	60.5 / 77.7	42.1 / 70.1	52.4 / 67.0	33.7 / 45.4	51.2 / 72.8
	F-Siamese[15]	LiDAR+RGB	37.1 / 50.6	16.2 / 32.2	—	47.0 / 77.2	—
稀疏卷积	PVT(本文)	LiDAR	68.4 / 81.8	50.7 / 77.2	56.1 / 66.9	69.0 / 92.8	59.7 / 78.7

注:"/"前部表示成功率,后部表示精确率.下划线标注了每列最优结果."—"表示此单元格结果为空.

基于NuScenes数据集的点云跟踪评价库是BAT方法中首次提出的评估基准. 由于其复杂性,当前只有少数跟踪方法在该数据集上进行评估测试. 本文作为点云单目标跟踪领域中基于稀疏卷积的尝试工作,也对比了基于原生点云特征提取的先驱SC3D[11]和里程碑方法P2B[13]和BAT[12]. 表3给出了该数据集上不同方法的对比结果.

整体来说,本文提出的方法PVT在11万多个目标帧中的跟踪结果表现十分具有竞争力,在各类的平均成功率和精确率指标上取得了37.9%和47.2%的结果. 与基于原始点云特征提取的方法相比,本文方法PVT在行人(Pedestrian)、汽车(Car)和公交车(Bus)这三个类别的跟踪上高于SC3D、P2B和BAT,将稀疏卷积框架下的跟踪向前推进了一步. 具体地,本文的PVT在各类跟踪性能显著高于先驱方法SC3D,平均指标高出17.6%/27.2%. 这得益于基于模板和搜索区域之间有效

表3 NuScenes上与其他前沿方法在成功率和精确率的比较结果 单位:%

特征提取方式	方法	数据形式	Car (64 159)	Pedestrian (33 227)	Truck (13 587)	Bus (2 953)	Mean (113 926)
原生点云	SC3D[11]	LiDAR	22.3 / 21.9	11.3 / 12.7	30.7 / 27.7	29.4 / 24.1	20.3 / 20.0
	P2B[13]	LiDAR	38.8 / 43.2	28.4 / 52.2	42.9 / 41.6	32.9 / 27.4	36.1 / 45.2
	BAT[12]	LiDAR	40.7 / 43.3	28.8 / 53.3	45.3 / 42.6	35.4 / 28.0	37.6 / 45.7
稀疏卷积	PVT(本文)	LiDAR	40.0 / 44.1	29.9 / 56.0	47.8 / 44.6	37.3 / 29.1	37.9 / 47.2

注:"/"前部表示成功率,后部表示精确率.下划线标注了每列最优结果.

关系建模以及端到端候选包围盒的生成,而SC3D仅利用PointNet提取特征、缺少特征间关系建模,并且借助卡尔曼滤波生成候选框. 此外,P2B引入PointNet++[18]孪生框架以后性能获得提升,但仍低于本文方法. 这说明基于稀疏卷积的孪生框架的有效性. 最后,虽然BAT通过引入几何先验进一步提升P2B的性能,但是本文在汽车上与其结果相当,在其他类(如行人、汽车)也高于该方法. 这得益于稀疏框架以响应图回归的内在优势,我们认为,在以后的工作中,以本文的稀疏框架为基准,进一步引入背景信息、时序记忆机制以及几何先验(如部件点到中心点距离、尺寸信息)等额外机制可进一步提升跟踪性能.

4.2.3　不同方法的计算性能对比

此外,为了说明跟踪方法在性能指标和计算资源消耗之间的平衡,图4给出了本文方法与当前代表性方法在KITTI上的对比,包括P2B、BAT、V2B、STNet以及LTTR. 图4中横坐标为每个方法训练的计算资源消耗,纵坐标为衡量跟踪结果的成功率指标,圆形的大小表示该方法占用内存资源的多少. 在此图中,越靠近左上角表示该方法较为高效. 观察图4可知,在相同的训练配置下(批大小batch size为64),相对于其他方法本文PVT在跟踪成功率和资源消耗取得了较好的平衡. 具体来说,与基于PointNet++框架的方法(如P2B、V2B)相比,PVT从最高~24 G降低为~5 G. 此外,与稀疏框架的LTTR相比,PVT在内存消耗和实时帧率(FPS)都有一定改善. 分析其中的原因在于,LTTR中在关系建模时,其注意力机制占据大量计算内存,并且使得跟踪帧率(23 FPS)低于PVT(50 FPS).

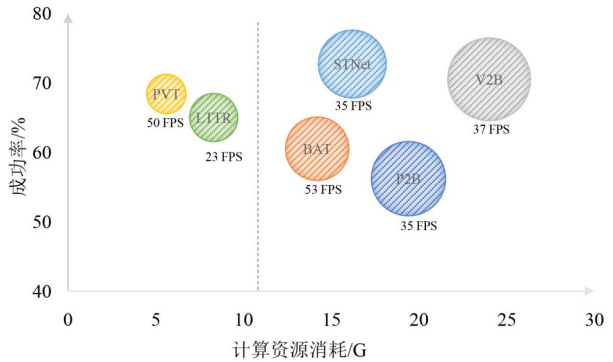

图4　跟踪性能与计算资源消耗之间平衡对比图

4.3　可视化结果对比

除了上述定量比较,本文给出了在KITTI数据集上的可视化结果. 图5给出了与其他方法P2B和BAT的对比结果. 图5展示了一个稠密汽车目标实例、一个稀疏汽车跟踪两实例以及一个行人跟踪实例,其中,绿色和红色的三维包围盒分别为真值和本文的跟踪结果. 从行人实例可以看出,本文PVT更加贴合目标真值,并且P2B随着目标移动以及出现干扰时发生了后续的偏移. BAT虽然略好于P2B但最后也偏离了原本目标. 从汽车实例上,其他两个方法在高度和水平发生了偏差,而PVT则能较好的贴合目标真实包围盒.

此外,图6专门给出了与基准方法LTTR的可视化对比结果. 同样地,本文用绿色和红色表示真值和本文结果. 为了说明本文所提方法在考虑长距和局部信息整合的有效性,本文给出了行人移动目标和汽车的快速移动目标的结果展示,其中行人移动较慢但局部干扰多,考验关系建模的局部感知能力;而车辆速度快考验关系建模时的长距信息关联能力. 从图6可以看出,本文提出的PVT方法均能较为准确地跟踪目标,而LTTR发生明显偏移. 特别地,在行人实例中,局部干扰使LTTR后续产生完全漂移,而本文方法全程捕捉到了目标. 在汽车实例中,LTTR估计的包围盒在多帧间发生了滞后于目标前进的现象,而本文方法则较好地避免该现象. 以上主要是基于体素和空间点双通道融合的关系建模在长距关联和局部信息判别发挥了一定作用.

4.4　消融实验与参数分析

本节针对本文提出的方法,从关系建模以及稀疏卷积框架两方面进行分析,包括关系建模中单个点分支、单个体素分支、点与体素分支融合方式、稀疏卷积框架中体素分辨率以及嵌入该框架的其他关系建模方式.

4.4.1　关系建模中单个空间点通道的结果

在第3节中,本文介绍了空间点与体素双通道的关系建模模块. 为了验证双通道有效性,本节首先对比了单个空间点通道的跟踪结果. 只保留依赖KNN近邻选取的点分支,并进一步探索近邻点的个数N_{knn}在进行关系建模时的影响. 该对比实验运行在KITTI数据集上. 如图7所示,横坐标表示近邻点个数,纵坐标为成功率指标σ_{succ}和精确率指标σ_{prec}. 由图7可知,近邻点的个数对结果具有一定的影响,当N_{knn}取12时,σ_{succ}和σ_{prec}取得了最优值. 其背后原因可解释为,邻居点太少,目标信息刻画不充分,而太多则会引入其他背景点的干扰.

4.4.2　关系建模中单个体素通道的结果

第4.4.1节对比了只使用空间点通道带来的影响,本节进一步给出只使用基于相似性累加和的体素通道的结果. 同时,探索以每个搜索区域点为中心时的局部体素尺寸r对跟踪结果的影响. 如图8所示,横坐标表示每个正方体体素的尺寸,纵坐标为成功率σ_{succ}和精确率σ_{prec}. 由图8可知,体素尺寸对该分支具有细微的影响,当体素大小为0.1时,所提的方法取得了最佳的性能.

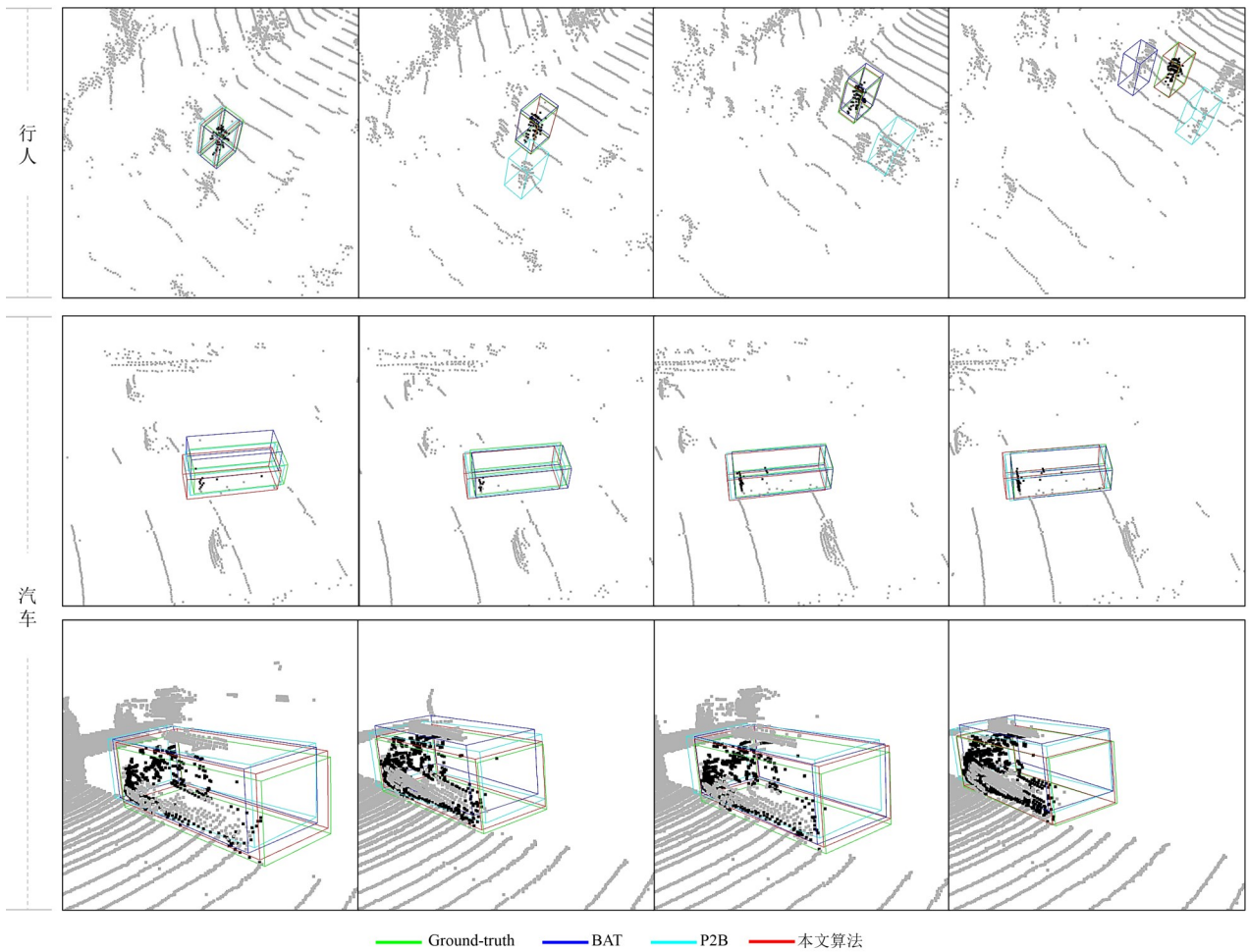

图5 本文方法与BAT、P2B方法的可视化对比结果

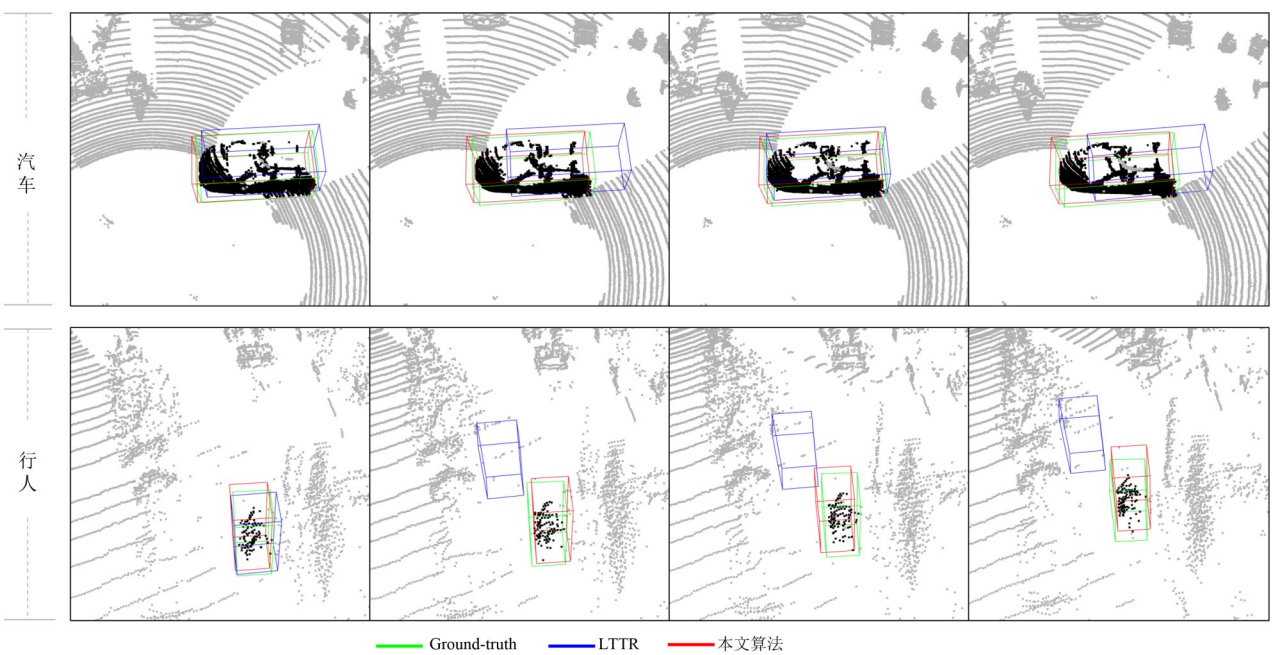

图6 本文方法与LTTR方法的可视化对比结果

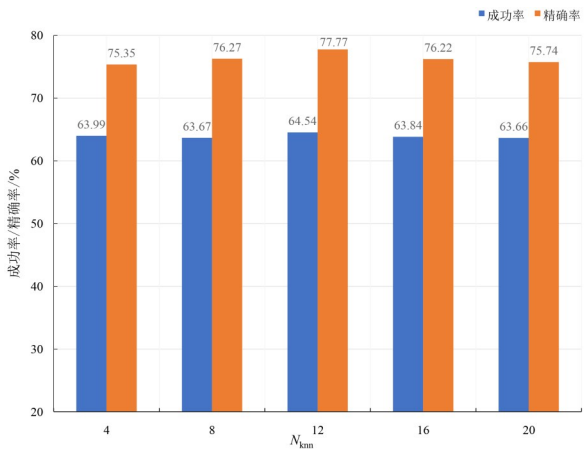

图7　单个空间点通道中不同近邻数量的表现

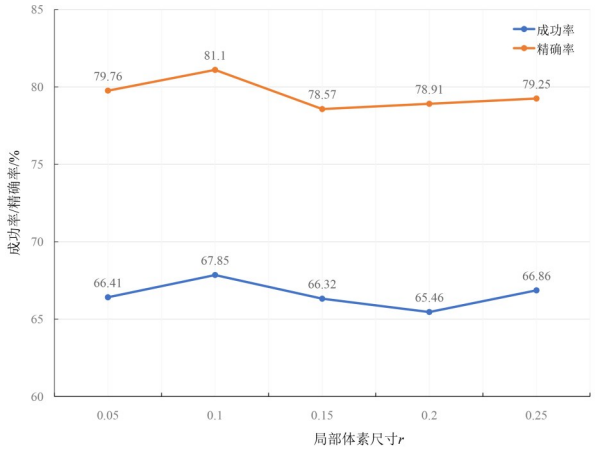

图8　单体素通道不同体素尺寸的表现

4.4.3　关系建模中点与体素分支融合方式对比

本小节探究了如何将本文提出的点与体素分支进行有效融合. 此处给出了两种不同的融合方式:特征对应通道相加以及特征通道拼接. 实验结果如表4所示,通道拼接的成功率和精确率分别为68.38%和81.80%;通道相加的方式则分别为68.18%和79.68%. 在整体表现上,通道拼接略高于通道相加的方式. 其主要原因是,前者将两种空间语义信息以独立通道反应能够增加目标判别性,而后者可能出现语义特征通道不一致相加的情况.

表4　关系建模中两种不同的融合方式　　单位:%

融合方式	成功率	精确率
对应通道相加	68.18	79.68
对应通道拼接	68.38	81.80

4.4.4　稀疏卷积框架中体素分辨率对比

为了全面评估方法本身的有效性,本节考虑稀疏卷积框架中体素分辨率大小带来的影响. 具体地,我们考虑了将整个跟踪场景帧按照0.02 m、0.03 m、0.04 m、0.05 m和0.06 m的体素边长来训练整个网络. 图9给出了不同参数设置下,本文方法训练的实验结果. 由图9可知,体素分辨率对整体跟踪性能有轻微影响,其现象为体素越小跟踪性能则较高. 较小的体素会有较高的分辨率,也会保留原始三维场景较多的空间信息,因此其跟踪性能表现相对较高.

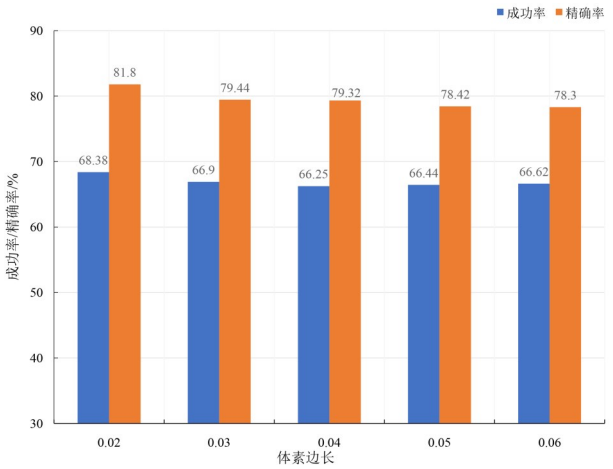

图9　稀疏卷积特征提取框架中不同体素大小的表现

4.4.5　稀疏卷积框架中与其他关系建模方式对比

为了说明本文提出的关系建模方式与稀疏卷积框架的适用性,我们将基于PointNet++框架的P2B和BAT方法中模板与搜索区域的关系建模方式放入本文稀疏框架中来进行对比验证,其分别为基于特征相似和基于包围盒几何尺寸先验的模块. 为了方便表述,将其分别记为XCorrP2B和XCorrBAT.

表5给出了二者与本文提出的关系建模方式在KITTI数据集上的对比. 从成功率和精确率整体指标来看,本文方法均高于XCorr_P2B和XCorr_BAT. 此外,值得注意的是,XCorr_P2B比原始的P2B提升了5.3%和3.4%,说明了稀疏框架的有效性. 同样,BAT的实验结果具有相同的结论. 另一方面,本文提出的双通道关系建模以6.82%和5.55%的指标优势高出XCorr_P2B,考虑了空间近邻和体素两方面的目标信息,更能适配稀疏卷积框架.

表5　与其他关系建模方式的结果对比　　单位:%

特征提取方式	关系建模方式	成功率	精确率
稀疏卷积	Xcorr_P2B	61.56	76.25
原始点云	P2B	56.2	72.8
稀疏卷积	XCorr_BAT	65.47	77.88
原始点云	BAT	60.5	77.7
稀疏卷积	本文方法	68.38	81.80

5　结语

本文提出了融合空间点与体素关系建模的3D稀疏卷积跟踪方法,从特征提取以及模板和搜索区域之

间的关系建模两个方面出发来提升LiDAR点云跟踪的性能. 针对特征提取,本文从PointNet++框架向稀疏卷积框架转变,有效节省了训练资源,推动了稀疏卷积框架在目标跟踪领域的表现;针对关系建模,本文从空间点K近邻和体素内相似性累计两个通道来描述目标在搜索区域中的实例判别特性. 在KITTI和NuScenes数据集上的定量和可视化实验表明,本文提出的PVT比基准方法获得显著提升,并且与其他前沿代表性方法相比,取得了具有竞争力的表现.

在将来的工作中,我们主要从三个方面进行研究. 首先,在特征提取上,将三维体素特征与图像模态特征相融合,从特征跨模态蒸馏来进一步改进;其次,在目标模板和搜索区域的关系建模上,考虑加入背景信息、时序记忆机制以及几何先验;最后,本文当前聚焦于前面两个方面,在候选包围盒生成上仍有空间,接下来考虑适应空间不同物体规模的候选生成方式.

参考文献

[1] 张伟俊, 钟胜, 徐文辉, 等. 融合显著性与运动信息的相关滤波跟踪算法[J]. 自动化学报, 2021, 47(7): 1572-1588.
ZHANG W J, ZHONG S, XU W H, et al. Correlation filter based visual tracking integrating saliency and motion cues[J]. Acta Automatica Sinica, 2021, 47(7): 1572-1588. (in Chinese)

[2] 陈丹, 姚伯羽. 运动模型引导的自适应核相关目标跟踪方法[J]. 电子学报, 2021, 49(3): 550-558.
CHEN D, YAO B Y. Adaptive response kernel correlation target tracking method guided by motion model[J]. Acta Electronica Sinica, 2021, 49(3): 550-558. (in Chinese)

[3] 林彬, 王华通, 封全喜. 基于双模型竞争机制的目标跟踪算法[J]. 电子学报, 2023, 51(5): 1381-1387.
LIN B, WANG H T, FENG Q X. Object tracking algorithm based on dual-model competition mechanism[J]. Acta Electronica Sinica, 2023, 51(5): 1381-1387. (in Chinese)

[4] 黄鹤, 李文龙, 吴琨, 等. 动态自适应特征融合的MFOPA跟踪器[J]. 电子学报, 2023, 51(5): 1350-1358.
HUANG H, LI W L, WU K, et al. MFOPA tracker with dynamic adaptive feature fusion[J]. Acta Electronica Sinica, 2023, 51(5): 1350-1358. (in Chinese)

[5] MARVASTI-ZADEH S M, CHENG L, GHANEI-YAKHDAN H, et al. Deep learning for visual tracking: A comprehensive survey[J]. IEEE Transactions on Intelligent Transportation Systems, 2022, 23(5): 3943-3968.

[6] LI P L, QIN T, SHEN S J. Stereo vision-based semantic 3D object and ego-motion tracking for autonomous driving[C]//European Conference on Computer Vision. Cham: Springer, 2018: 664-679.

[7] SINGH P, AGRAWAL P, KARKI H, et al. Vision-based guidance and switching-based sliding mode controller for a mobile robot in the cyber physical framework[J]. IEEE Transactions on Industrial Informatics, 2019, 15(4): 1985-1997.

[8] WANG B, WU V, WU B C, et al. LATTE: Accelerating LiDAR point cloud annotation via sensor fusion, one-click annotation, and tracking[C]//2019 IEEE Intelligent Transportation Systems Conference (ITSC). Piscataway: IEEE, 2019: 265-272.

[9] ASVADI A, GIRÃO P, PEIXOTO P, et al. 3D object tracking using RGB and LIDAR data[C]//2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC). Piscataway: IEEE, 2016: 1255-1260.

[10] PANG Z Q, LI Z C, WANG N Y. Model-free vehicle tracking and state estimation in point cloud sequences[C]//Proceedings of the International Conference on Intelligent Robots and Systems (IROS). Piscataway: IEEE, 2021: 8075-8082.

[11] GIANCOLA S, ZARZAR J, GHANEM B. Leveraging shape completion for 3D Siamese tracking[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2019: 1359-1368.

[12] ZHENG C D, YAN X, GAO J T, et al. Box-aware feature enhancement for single object tracking on point clouds[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2021: 13199-13208.

[13] QI H Z, FENG C, CAO Z G, et al. P2B: Point-to-box network for 3D object tracking in point clouds[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2020: 6329-6338.

[14] HUI L, WANG L P, CHENG M M, et al. 3D Siamese voxel-to-bev tracker for sparse point clouds[C]//Proceedings of Advances in Neural Information Processing Systems (NeruIPS). San Diego: MIT Press, 2021: 28714-28727.

[15] ZOU H, CUI J H, KONG X, et al. F-siamese tracker: A frustum-based double Siamese network for 3D single object tracking[C]//2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Piscataway: IEEE, 2020: 8133-8139.

[16] ZHOU C Q, LUO Z P, LUO Y R, et al. PTTR: relational

3D point cloud object tracking with Transformer[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2022: 8521-8530.

[17] SHAN J Y, ZHOU S F, CUI Y B, et al. Real-time 3D single object tracking with Transformer[J]. IEEE Transactions on Multimedia, 2022, 25: 2339-2353.

[18] QI C R, YI L, SU H, et al. PointNet++: Deep hierarchical feature learning on point sets in a metric space[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc., 2017: 5105-5114.

[19] YAN Y, MAO Y X, LI B. SECOND: Sparsely embedded convolutional detection[J]. Sensors, 2018, 18(10): 3337.

[20] GRAHAM B, ENGELCKE M, VAN DER MAATEN L. 3D semantic segmentation with submanifold sparse convolutional networks[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 9224-9232.

[21] CHOY C, GWAK J AND SAVARESE S. 4D spatio-temporal convnets: Minkowski convolutional neural networks[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2019: 3070-3079.

[22] CUI Y B, FANG Z, SHAN J Y, et al. 3D object tracking with Transformer[C]//Proceedings of British Machine Vision Conference (BMVC). Durham: British Machine Vision Association, 2021: 1-13.

[23] SHI S S, GUO C X, JIANG L, et al. PV-RCNN: Point-voxel feature set abstraction for 3D object detection[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2020: 10526-10535.

[24] WEI Y, WANG Z Y, RAO Y M, et al. PV-RAFT: Point-voxel correlation fields for scene flow estimation of point clouds[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2021: 6950-6959.

[25] GEIGER A, LENZ P, STILLER C, et al. Vision meets robotics: The KITTI dataset[J]. International Journal of Robotics Research, 2013, 32(11): 1231-1237.

[26] CAESAR H, BANKITI V, LANG A H, et al. NuScenes: A multimodal dataset for autonomous driving[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2020: 11618-11628.

[27] CHOY C, PARK J, KOLTUN V. Fully convolutional geometric features[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2019: 8957-8965.

[28] 李宗民, 姚纯纯, 刘玉杰, 等. 点云场景下基于结构感知的车辆检测[J]. 计算机辅助设计与图形学学报, 2021, 33(3): 405-412.
LI Z M, YAO C C, LIU Y J, et al. Vehicle detection based on structure perception in point cloud[J]. Journal of Computer-Aided Design & Computer Graphics, 2021, 33(3): 405-412. (in Chinese)

[29] SHI S S, WANG X G, LI H S. PointRCNN: 3D object proposal generation and detection from point cloud[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2019: 770-779.

[30] 周锋, 陶重犇, 张祖峰, 等. 体素点云融合的三维动态目标检测算法[J]. 计算机辅助设计与图形学学报, 2022, 34(6): 901-912.
ZHOU F, TAO C B, ZHANG Z F, et al. 3D dynamic target detection algorithm based on voxel point cloud fusion[J]. Journal of Computer-Aided Design & Computer Graphics, 2022, 34(6): 901-912. (in Chinese)

[31] FANG Z, ZHOU S F, CUI Y B, et al. 3D-SiamRPN: An end-to-end learning method for real-time 3D single object tracking using raw point cloud[J]. IEEE Sensors Journal, 2021, 21(4): 4995-5011.

[32] QI C R, LITANY O, HE K M, et al. Deep Hough voting for 3D object detection in point clouds[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2019: 9276-9285.

[33] WANG Z T, XIE Q, LAI Y K, et al. MLVSNet: Multi-level voting siamese network for 3D visual tracking[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2021: 3101-3110.

[34] TIAN S J, LIU B, TAN H C, et al. Deep supervised descent method with multiple seeds generation for 3D tracking in point cloud[J]. IEEE Transactions on Industrial Informatics, 2022, 18: 5077-5086.

[35] ZHOU X Y, WANG L, YUAN Z A, et al. Structure aware 3D single object tracking of point cloud[J]. Journal of Electronic Imaging, 2021, 30(4): 043010.

[36] PARK M, SEONG H, JANG W, et al. Graph-based point tracker for 3D object tracking in point clouds[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2022: 2053-2061.

[37] ZHENG C D, YAN X, ZHANG H M, et al. Beyond 3D Siamese tracking: A motion-centric paradigm for 3D single object tracking in point clouds[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2022: 8101-8110.

[38] WU Y, LIM J, YANG M H. Object tracking benchmark[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(9): 1834-1848.

作者简介

田胜景 男,1994年出生,山东济宁人. 于大连理工大学获得博士学位,现入职中国矿业大学师资博士后. 主要研究方向为点云理解、3D视觉.

E-mail: tye.dut@gmail.com

韩一男 男,1994年出生,辽宁沈阳人. 于东北大学获得学士学位,现为大连理工大学博士研究生. 主要研究方向为点云理解、跨模态表示学习.

赵宪通 男,1999年出生,山东泰安人. 于中国矿业大学获得学士学位,现为大连理工大学博士研究生. 主要研究方向为点云跟踪、多模态学习.

刘秀平 女,1964年出生,辽宁鞍山人. 于吉林大学获得博士学位,现为大连理工大学教授. 主要研究方向为计算机视觉、计算机图形学.

张 明 男,1980年出生,山东博兴人. 于大连理工大学获得博士学位,现为中国矿业大学教授. 主要研究方向为大数据管理与应用、复杂网络.