文章编号:1671-4598(2024)10-0077-09 DOI:10.16526/j.cnki.11-4762/tp.2024.10.011 中图分类号:TP212.9;TP242 文献标识码:A

基于 BEV 视角的多传感融合 3D 目标检测

张 津[1], 朱冯慧[1], 王秀丽[1,2], 朱 威[1,2]
(1. 浙江工业大学 信息工程学院, 杭州 310023; 2. 浙江省嵌入式系统联合重点实验室, 杭州 310023)

摘要: 3D 目标检测是自动驾驶在道路环境感知任务中的重要环节, 现有主流框架通过搭载多种感知设备获取多模态的数据信息来实现多传感融合检测; 传统相机与激光雷达的传感器融合过程中存在几何失真, 以及信息优先级不对等, 导致传感融合的 3D 目标检测性能不足; 对此, 提出了一种基于鸟瞰视角 (BEV) 的多传感融合 3D 目标检测算法; 利用提升-展开-投射 (LSS) 方式, 获取图像的潜在深度分布建立图像在 BEV 空间下的特征; 采用 PV-RCNN 的集合抽象法建立点云在 BEV 空间下的特征; 该算法在统一的 BEV 共享空间中设计了低复杂度的特征编码网络融合多模态特征实现 3D 目标检测; 实验结果表明, 所提出的算法在检测精度上相较于纯激光方法提升 4.8%, 相较于传统的融合方案减少了 47%的参数, 并保持了相近的精度, 较好地满足了自动驾驶系统道路环境感知任务的检测要求。

关键词: 3D 目标检测; 鸟瞰图视角; 多传感融合; 自动驾驶; 道路环境感知

3D Object Detection for Multi-sensor Fusion Based on BEV Perspective

ZHANG Jin[1], ZHU Fenghui[1], WANG Xiuli[1,2], ZHU Wei[1,2]
(1. College of Information Engineering, Zhejiang University of Technology, Hangzhou 310023, China;
2. Joint Key Laboratory of Embedded Systems of Zhejiang Province, Hangzhou 310023, China)

Abstract: 3D object detection is an important part in the road environment perception of autonomous driving. Existing mainstream framework is to obtain multi-modal data by using multiple sensing devices, which achieves multi-sensor fusion. There are the shortages of geometric distortions and unequal information priorities in the fusion process of traditional cameras and LiDARs, resulting in insufficient 3D object detection performance of sensor fusion. To address this issue, a multi-sensor fusion 3D object detection algorithm based on bird's-eye view (BEV) is proposed. The lift-splat-shot (LSS) method is used to obtain the potential depth distribution of the image, and establish the feature map of the image in the BEV space. The set abstraction method of point-voxel region convolutional neural networks (PV-RCNN) is used to establish the feature map of the point cloud in the BEV space. A low-complexity feature encoding network is designed for fusing multi-modal features in a unified BEV space in the proposed method to achieve 3D object detection. Experimental results show that the proposed method improves the accuracy by 4.8% compared to the LiDAR method, reduces the parameters by 47% compared to the traditional fusion methods, and maintains similar accuracy. The proposed method meets the detection requirements of the road environment perception of autonomous driving system.

Keywords: 3D object detection; BEV view; multi-sensor fusion; autonomous driving; road environment perception

0 引言

随着高速领航辅助驾驶 (高速 NOA) 功能的推广, 智能驾驶系统已经进入到了大众的工作和生活。在不断完善高速辅助驾驶功能的同时, 城市内封闭道路领航辅助驾驶 (城区 NOA) 功能也正逐步研发实现。不同于高速公路环境, 车辆在城市中的周围环境信息会变得极为复杂, 因此精准、高效的 3D 目标检测成为自动驾驶系统不可或缺的一环[1]。多传感融合的 3D 目标检测算法能够通过利用搭载多种传感器设备提升对道路目标 (如行人、自行车、车辆) 的感知能力[2], 然而由于设备间的检测视场差异, 以及传感器融合中的几何失真问题, 导致多传感器融合的方法有时甚至不如单传感器的方法。因此本文首先对多传感器进行联合标定展开研究, 将各个传感器转换到相同的坐标系中, 为后续融合奠定基础。

根据融合阶段划分, 传感融合的目标检测的融合策略分为 3 种类型: 原始数据端融合 (前期融合), 特征端融合 (中期融合) 和决策端融合 (后期融合)。数据端的前期融合方法包括: PointPainting[3]、PointAugmenting[4], 需要构建复杂的映射关系, 把点云数据映射到图像数据, 或是把图像映射到点云数据。传统特征端的中期融合方法包括: MV3D[5]、AVOD[6], 使用 CNN 提取点云与 RGB 图像的特征, 输入 RPN 后进行融合。决策端的后期融合方法包括: CLOCs[7]、Fast CLOCs[8], 使用低复杂度的多模态融合框架将独立的点云检测与图像检测候选的一致性关系, 输入

收稿日期:2024-04-30; **修回日期**:2024-05-13。
基金项目:国家自然科学基金青年项目(62303414);浙江省自然科学基金探索青年项目(LQ23F030016)。
作者简介:张 津(1998-),男,硕士研究生。
通讯作者:朱 威(1982-),男,博士,副教授。
引用格式:张 津,朱冯慧,王秀丽,等. 基于 BEV 视角的多传感融合 3D 目标检测[J]. 计算机测量与控制,2024,32(10):77-85.

稀疏卷积计算得到最终的融合结果。

基于算法处理的数据结构，将 3D 目标检测算法主要分为三类：第一类是基于原始点云的 3D 目标检测[9]，第二类是基于网格的 3D 目标检测，第三类是基于鸟瞰图的 3D 目标检测。基于原始点云的 3D 目标检测方法包括：Pointnet[10]、Pointnet++[11]、Point-RCNN[12]，这类方法直接使用原始点云的方法无需将点云转换为其他网格表示的结构，能够最大限度地保留原始的几何细节。基于网格的 3D 目标检测方法包括：SECOND[13]、PointPillars[14]，不同于直接的原始点云方法，为了克服点云的大量级、无序性、不均一的难点，这些方法将点云在 X，Y，Z 三个坐标轴上进行划分，转换为其他网格形式的编码表示。基于鸟瞰图视角的 3D 目标检测方法包括：BEV Det[15]、Bevpool[16]，这些方法通过将特征从图像视图转换为 BEV 的视图，进一步编码使用 BEV 中目标预测头实现目标检测，在鸟瞰图视角能够表示更加全面的完整场景。

以上方法虽然实现方式各不相同，但目的都是在寻找点云中得到的精准特征与计算量相平衡的最优方案。原始点云方法虽然能最大限度地保留原始的几何信息，但是其计算开销也是最大的；基于网格方法虽然减少了复杂度，拥有更加高效的速度，但是存在部分信息的缺失。

为了有效解决上述问题，本文在多传感器联合标定的基础上，提出了一种基于 BEV 视角的多传感融合 3D 目标检测算法。首先，采用提升－展开－投射（LSS）方式[17]，获取图像的潜在深度分布得到图像在鸟瞰图（BEV）空间下的特征；随后，使用 PV-RCNN[18]的集合抽象法得到点云在 BEV 空间下的特征；最后，在统一的 BEV 共享空间中融合多模态特征，用于 3D 目标检测。本文主要创新点如下：

1）通过设计相机与激光雷达的标定方案，设计了一种自动标定方法，得到精确的传感器位置关系，提升传感早期的数据端融合精度。

2）对特征融合方法进行改进，采用提升－展开－投射（LSS）方式，并增加 PV-RCNN 的体素集合抽象维度，获取图像与点云在鸟瞰图（BEV）空间下相匹配的特征，使多模态特征拥有统一的编码表示。

3）设计了低复杂度的特征编码网络，在 BEV 共享空间中融合图像与点云的多模态特征，抑制了多模态信息优先级不对等问题，提升 3D 目标检测精度。

1 数据级融合标定算法

不同的传感器在系统中处于不同的位置，角度以及坐标系中，这些差异称之为传感器的空间差异性，因此需要通过对传感器进行联合标定，从而将传感器转换到相同的坐标系中。本节介绍相机与激光雷达联合标定的具体原理与操作步骤，以得到校准后的统一坐标系。

由于相机坐标系与激光雷达坐标系之间存在空间差异性，需要确定相机和激光雷达之间的相对位置和姿态关系，即两个坐标系之间的刚性转换。能否得到正确的刚性转换矩阵，建立两者之间的准确映射关系是后续的相机与激光雷达在同一坐标系下进行有效融合的基础。整体标定流程如图 1 所示。

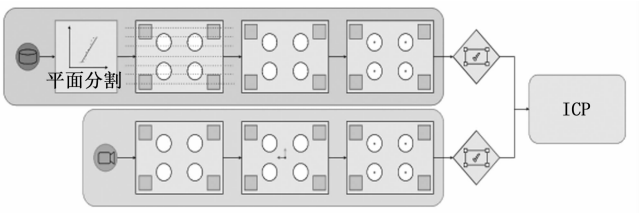
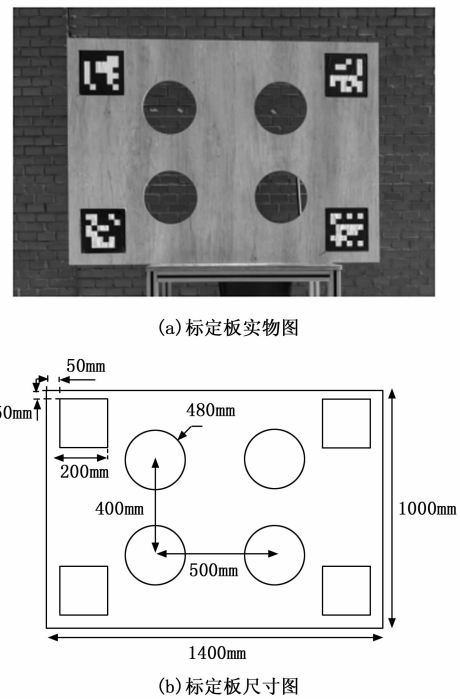


图 1 标定流程图

本文采用了专门设计的标定板，如图 2 所示。该标定板中包含 4 个 ArUco 标记码，以及 4 个空心圆孔结构，分别用于 4 个特征点的三维空间坐标提取。整个校准标定过程分为两个不同的阶段：第一个阶段涉及校准目标的分割和参考点在各自传感器坐标系中的定位；第二个阶段计算两个传感器坐标之间的刚性变换参数，使变换参数能够配准参考点。



(a)标定板实物图

(b)标定板尺寸图

图 2 velo2cam 标定板

1.1 单目相机特征点提取

启动单目相机作为一号传感器，通过输入的图像识别位于校准目标 4 个角上的 ArUco，检测到 ArUco 的 4 个角点 p_{corner}。通过 p_{corner}、ArUco 的现实尺寸 x、相机内参 I_k，解决 n 点透视问题（PnP）得到 ArUco 的位姿。如图 3，通过其中的三个角点作为参考点，构建相似三角形，使用余弦定理得到式（1）：

$$\begin{aligned} d_1^2+d_2^2-2d_1d_2\cos\theta_{12}&=d_{12}^2\\ d_2^2+d_3^2-2d_2d_3\cos\theta_{23}&=d_{23}^2\\ d_1^2+d_3^2-2d_1d_3\cos\theta_{13}&=d_{13}^2 \end{aligned} \tag{1}$$

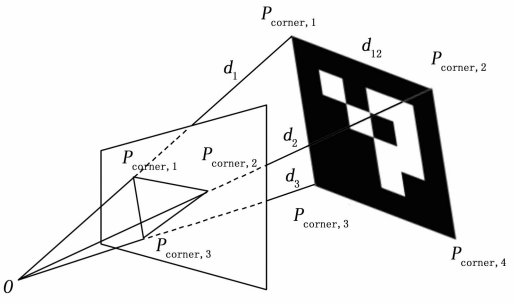


图 3　P3P 几何原理结构图

其中：$d_i = ||\overrightarrow{OP}||_{corner,i}$，$d_{i,j} = ||\overrightarrow{OP}||_{corner,ij}$。不妨设式（1）中的 $x = \frac{d_2}{d_1}$，$y = \frac{d_3}{d_1}$，$v = \frac{d_{12}^2}{d_1^2}$，$u = \frac{d_{23}^2}{d_{12}^2}$，$w = \frac{d_{13}^2}{d_{12}^2}$。代入消除 v，可以得到式（2）：

$$\begin{aligned}(1-u)y^2 - ux^2 - \cos\theta_{23}y + 2uxy\cos\theta_{12} + 1 = 0\\(1-w)x^2 - uy^2 - \cos\theta_{13}x + 2wxy\cos\theta_{12} + 1 = 0\end{aligned} \tag{2}$$

通过内参 I_k 以及像素坐标系中的 p_{corner}，可以求得 $\cos\theta$，通过 ArUco 的现实尺寸可以求得 u，w，因此式（2）为二元二次方程。由于 4 个角点不共面，该方程最多可能有 4 个解，最后使用验证点 $p_{corner,4}$ 来计算确切的最终解，得到 p_{corner} 在相机坐标系下的三维坐标，使用 ICP 计算相机的位姿得到旋转向量 $\boldsymbol{R}_{vec}$，平移向量 $\boldsymbol{T}_{vec}$。将 4 组 ArUco 位姿的均值作为标定板的初始位姿猜测 $\boldsymbol{R}_{vec}$，平移向量 $\boldsymbol{T}_{vec}$，对角点计算重投影误差后，使用 LM 优化算法，得到相机坐标系下精准的标定板中心位姿。通过圆孔与标定板在实际中的位置关系，计算圆孔在相机坐标系中的 3D 位置 p_p^M，如图 4 所示。

1.2　激光雷达特征点提取

启动激光雷达作为二号传感器，使用直通滤波器，将激光雷达点云中的额外点滤除，保留标定目标以及背景墙，得到 p_1^L。使用 RANSAC 拟合标定板平面，得到平面内点 $p_{inliers}^L$，以及标定板平面的参数 $coeff_{plane}$。通过式（3）计算内点梯度，$p_{i,r}$ 是 $p_{inliers}^L$ 中点 p_i 与激光雷达的距离值，p_{i+1} 和 p_{i+1} 是同一行扫描中与 p_i 相邻的点。

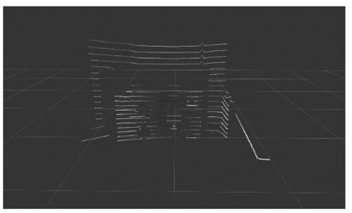
$$\Delta p_{i,r} = max(p_{i-1,r} - p_{i,r}, p_{i+1,r} - p_{i,r}, 0) \tag{3}$$

将 $\Delta p_{i,r} < 0.1$ 与距离标定板平面超过 0.1 的点滤除，便得到了标定板上的边缘点。

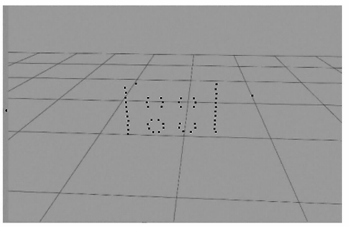
已知保留的点共面，为了便于计算通过旋转变换，将当前得到的边缘点统一到 xy 平面上转换为二维得到 p_2^L，简化后续求解。循环使用 RANSAC 拟合圆孔，使其半径 $r = 1.2 \pm 0.1$ 直至剩余的点数少于 3。得到多组圆心 p_{conter}^L。p_{conter}^L 通过排列组合方式得到 4 个一组的候选。

建立四边形，并通过计算所得到的四边形与目标矩阵的几何一致性（周长、对角线、长、宽误差小于 0.06），确认最优的候选点组，将其从 xy 平面重新变换回激光雷达的坐标中，最终得到 4 个圆孔在相机坐标系中的 3D 位置 p_p^L，如图 4 所示。

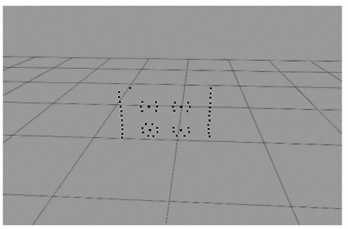
(a) 标定板图像关键点



(b) 标定板点云过滤图



(c) 标定板点云边缘点



(d) 标定板点云关键点

图 4　特征点检测

1.3　单目相机与激光雷达坐标变换求解

由于将校准场景保持静态一段时间通常是可行的，因此在 30 个数据帧上累积组成的点 p_p 以生成，然后 p'_p 对执行欧几里得聚类。如果发现超过 4 个集群，则认为数据不可靠，排除该组数据；否则，对得到的 4 个集群分别计算类内点云的质心 p_c，用作中心位置的合并估计。

通过考虑 4 个以上的参考点来提高估计的准确性。如果针对校准目标相对于传感器的 N 个不同姿态重复分割过程，则利用每个姿态获得的 p_c 被累积在 p'_c 中，其中 $4 \times N$ 个参考点可用于执行配准阶段。对于每个姿势的分割，假设传感器和目标都是静态的。

每个传感器提供了一组点云 p'_c。p'_c 中包含了传感器输入的某一帧经过计算后得到的中心 3D 位置估计。配准步骤的目标是找到最佳参数 $\theta = (t_x, t_y, t_z, r_x, r_y, r_z)$，使得当应用所得到的最佳参数所组成的变换矩阵时，能够得到从两个传感器获得的参考点之间的最佳对准。因此该问题可以被视为具有 $4 \times N$ 个目标函数的多目标优化。如式（4）：

$$\text{argmin} = \frac{1}{4 \cdot N} \| p_i^M - T_{ML} p_i^L \|^F \tag{4}$$

使用奇异值分解（SVD）来最小化目标函数，最终可以计算得到矩阵 $\boldsymbol{T}_{ML}$，最终得到 6 个参数 $\theta=(t_x, t_y, t_z, r_x, r_y, r_z)$。

2 特征融合目标检测算法

不同的传感器拥有不同的数据类型、数据密度以及环境适应性，这些差异称之为传感器的数据差异性，因此需要通过对传感器的融合方式进行设计，从而将有效的传感器融合应用于 3D 多目标检测算法。传统方法将一种模态映射到另一种模态空间下，以便于使用单一模态的成熟方案实现感知任务。然而，这种模态间映射的投影过程引入了严重的几何失真，同时将一种模态映射到另一种模态空间会导致信息优先级不对等，这使得它在面向后续感知任务中的效果较差。为此本节提出轻量化的 BEV 视场融合方案，以在 BEV 空间中统一多模态特征，用于学习保持了几何结构和语义信息，保证精度与效率。

2.1 单目相机 2D 特征提取

传统的计算机视觉任务中，通常在相机图像坐标中，直接使用卷积或是编解码方式，提取特征用于最终的预测。但是自动驾驶的感知任务对环境进行感知表示的空间往往需要在 BEV 视图下完成，以便于后续任务流程的实现。为此本节基于 LSS 方式，获取图像的潜在深度分布。对图像特征进行升维表示，并使用上一节中的标定结果指导特征从相机图像坐标映射至 BEV 视角的空间中，在 BEV 坐标系中找到场景的栅格化的占有表示，完成对 2D 图像特征的提取，以准备后续在 BEV 视角下将 2D 特征用于融合工作。

2.1.1 基于 LSS 提取潜在深度特征

外参矩阵 $\boldsymbol{E}_k$ 和内参矩阵 $\boldsymbol{I}_k$ 一起定义了相机中从世界坐标到局部像素坐标的映射。而单目相机与激光雷达的融合挑战在于，需要深度来转换局部像素坐标为三维点云参考系中的坐标。但每个像素相关的深度在本质上的模糊性，导致了单目相机与激光雷达的融合往往存在几何的不一致性。本小节基于 LSS 为每个像素生成所有可能的深度表示，用于将图像像素转为以相机为中心的视锥结构。

设 $x_k \in R^{3\times H\times W}$ 是一个具有外参 $\boldsymbol{E}_k$ 和内参 $\boldsymbol{I}_k$ 的相机得到的像素空间图像，设 p 是图像中坐标为 (h, w) 的一个像素。将图像中的每个像素与 $|D|$ 个点相关联得到三维点 $\{(h, w, d) \in R^2 \mid d \in D\}$，其中 D 是一组离散深度，定义的深度如 $\{d_0+\Delta, \cdots, d_0+|D|+\Delta\}$。通过这个方式，构建了视锥结构的点云空间，类似于多平面图像。至此完成了将图像像素转为视锥结构的空间定义。

点云中每个点的初始上下文向量都包含了两组信息，以匹配图像特征和离散深度推理的概念。在像素处，网络预测上下文向量为 $\boldsymbol{c} \in R^C$，随深度的预测的概率分布表示为 $\alpha \in D$。对于不同预测深度的点 p_d，其最终的关联特征由图像特征 $\boldsymbol{c}_d \in R^C$ 与深度概率放缩后得到 $\boldsymbol{c}_d = \alpha_d \boldsymbol{c}$。当 α 定义为一个独热编码时，则点 p_d 处的上下文特征向量将仅针对某一处唯一的深度为非零。如果网络预测 α 为均匀分布，则网络将预测平均地分配给像素 p 的每个点，深度信息则被忽视。因此，理论上，该网络能够将图像中的上下文特征分布在从鸟瞰图的特定位置至整个空间射线之间。

2.1.2 相机特征转 BEV 特征

为了便于后续的融合，现在得到深度信息放缩后的特征处于视锥空间中，本小节实现如何将视锥空间的特征统一至体素（Voxel）从而得到栅格化后的占有形式。体素是三维空间中的体积元素，将连续的三维空间划分为规则的、离散的小立方体，每个立方体就是一个体素，通过将稠密的点云简化为体素网格的形式，有效降低数据量，提高处理效率。为了后续融合的统一性，创建体素空间感受野 $RF=[0, -50, -10, 50, 50, 1]$，体素尺寸为 $[0.5, 0.5, 0.2]$。利用内外参，对相机的视锥进行坐标变换，输出视锥点云在自车周围物理空间的位置索引。将每个点分配给其所属的体素，并执行求和池化以在每一个 z 轴上创建一个张量，从而将体素最终投影至 BEV 视角。

虽然图像像素给出了密集的点云分布，但由于体素本身的几何性质，许多区域在投影过程中没有被点云所覆盖。导致在这些区域内的体素没有包含任何点云数据，从而产生很多空的体素，影响求和池化的效率。因此使用累积和技巧实现池化的计算过程，此操作可以有效计算的分析梯度，减少重复计算，提高池化操作的效率。

给出区间索引 $Id=[idex_1, idex_2, \cdots, idex_n]$ 以及特征值 $A=[a_1, a_2, \cdots, a_n]$，区间索引提供了每个点进行排序后所属于的体素序列索引，求和池化对每个体素内的点的特征进行求和计算，需要使用累积和技巧，首先计算区间索引的梯度得到 $d \in [d_1, d_2, \cdots, d_n]$，其中梯度满足如下式（5）要求计算得到：

$$d_i = \begin{cases} 0 & idex_j = idex_{j+1} \\ 1 & idex_j \neq idex_{j+1} \end{cases} \tag{5}$$

随后特征值的累积和 $SUM_A = \left[a_1, a_1+a_2, \cdots, \sum_{i=1}^{n} a\right]$；梯度 d_i 提供了体素与体素之间对于点的划分间隔。求和池化的求和操作即从位置 i 到位置 j 的元素总和，只需转化为对应位置的和的差值，即可得到最终的求和结果 S_{ij} 如下式：

$$S_{ij} = d_j \cdot a_j - d_i \cdot a_i \tag{6}$$

至此完成了将 2D 图像特征进行潜在深度估计后，转为视锥结构的 3D 点，并通过池化操作，将点云特征转为 BEV 视角的 2D 特征。

2.1.3 2D 特征网络结构

整个网络流程如图 5 所示。对于图像存在以下超参，输入图像 $H\times W$ 的大小为 128×224，并根据内外参对图像畸变进行修复。并通过 16 倍下采样后，事先构建离散深度 $D=[4, 45]$，间隔为 1 m 的视锥结构，用于后续的特征提取，得到 $\left[41\times\frac{H}{16}\times\frac{W}{16}\right]$ 的视锥结构。

对于图像进行上下文特征提取操作的网络，本文采用

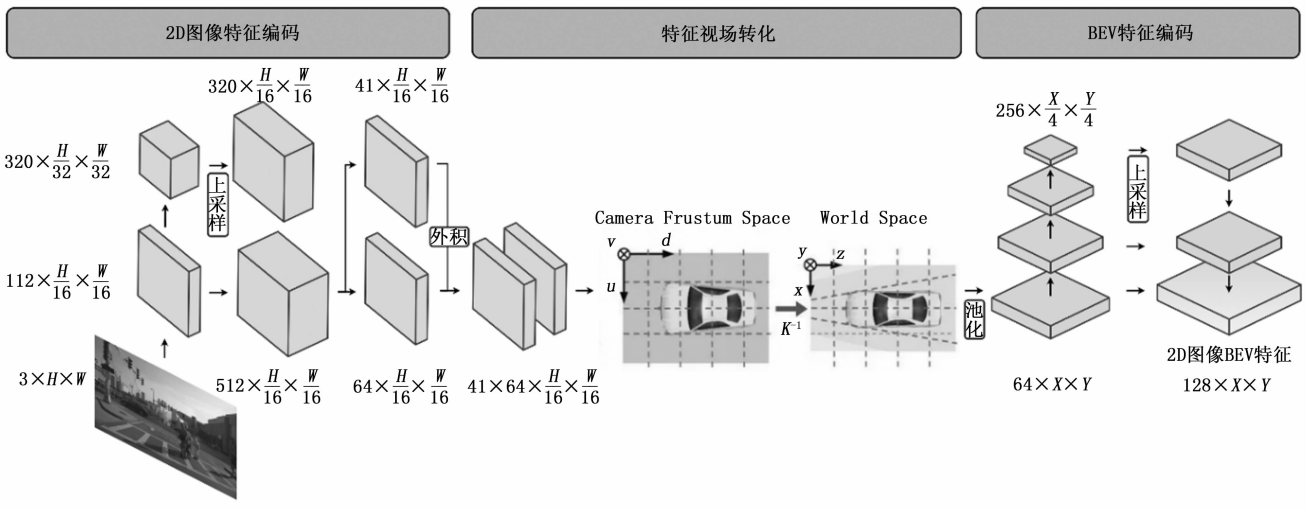


图 5　2D 图像转 BEV 特征编码网络

了在 Imagenet[19] 上预训练的 EfficientNet-B0[20] 主干网络。最终得到一个$\left[128\times\frac{H}{16}\times\frac{W}{16}\right]$的原始特征。而对于 LSS 的离散深度估计，用 1×1 卷积层实现，从而得到$\left[41\times\frac{H}{16}\times\frac{W}{16}\right]$维的特征。最终图像特征与深度概率通过外积放缩后，输出特征图的尺寸从原先的$\left[\frac{H}{16}\times\frac{W}{16}\right]$升维为$\left[41\times\frac{H}{16}\times\frac{W}{16}\right]$，每个位置的语义特征大小为$\left[128\times\frac{H}{16}\times\frac{W}{16}\right]$。

对于 BEV 转化网络，利用求和池化操作，构建 BEV 特征。设置 BEV 网格的分辨率大小 $X\times Y$ 为 100×200（前视区域），单元大小为 0.5×0.5 的体素。使用 ResNet-18 的前 3 层以获得 3 个不同分辨率的鸟瞰图表示。最后上采样，以返回到原始输入 BEV 尺寸，输出 BEV 视角下的 2D 语义特征 f_{bev}^{2D}，f_{bev}^{2D}特征大小为 [128×100×200]。

2.2　激光雷达 3D 特征提取

激光雷达（LiDAR）能够为感知系统提供 3D 空间中的坐标数据，从而使感知系统对环境拥有更加精准的几何结构认知能力。为了更好地实现感知任务，对 3D 的几何特征提取至关重要。3D 点云特征提取的方法包括基于原始点云（Raw Point）的方法、基于体素（Voxel）的方法和基于 BEV 视角的方法。这些方法各有优势：直接从原始点云构建特征的方法能够保留最多的几何信息，但忽略了计算的复杂性；体素方法简化了数据复杂度但可能损失细节。将基于体素的方法和基于原始点云方法结合后，在 BEV 坐标系中得到 3D 点云特征的提取，以准备后续在 BEV 视角下将 3D 特征用于融合工作。

2.2.1　原点－体素 3D 关键点特征提取

1）关键点提取。关键点的获取使用最远点采样（FPS，farthest point sampling），以获取均匀分布在非空体素周围，并且可以代表整个场景的关键点。最远点采样（FPS）的迭代步骤如下：

（1）对于原始点云 P_{lidar}，随机选择一个点作为初始点加入至采样集中；

（2）在剩余的点云中，计算每个点与采样集 K 的欧氏距离 d（p，K）；

（3）选择距离值最大的点，并加入至采样集 K 中；

（4）重复（2）、（3）直至采样集达到预定的大小。

通过最远点采样能够使采样在空间上覆盖整个点集，同时使得采样点之间的空间分布尽可能均匀，缓解了需要太多体素来编码整个场景的问题，最终得到关键点 $K=\{p_1,\cdots,p_n\}$。

2）原点－体素编码。原始点信息具有更加灵活的感受野保留了准确的位置信息，而体素信息有效地编码了多尺度特征表示，生成高质量的 3D 建议。用这些关键点，统一地将原始点与体素信息进行有效编码至关重要。因此需要得到原始点与体素的几何对应关系，称之为原点－体素编码操作。

关键点 $K=\{p_1,\cdots,p_n\}$ 来自于原始点云，因此原始点的特征信息来源亦可以使用关键点 K 进行获取。而对于体素，需要进行原始－体素编码操作。

首先，对体素特征进行关键点位置关联编码。体素特征采用基于 CNN 的体素特征编码进行提取。得到以下定义，在第 k 层中体素的特征向量为 $\boldsymbol{F}^{l_k}=\{f_1^{l_k},\cdots,f_{N_k}^{(l_k)}\}$，对应体素的坐标集合为 $\boldsymbol{V}^{l_k}=\{v_1^{l_k},\cdots,v_{N_k}^{(l_k)}\}$，其中 N_k 是第 k 层中的非空体素的数量。对于关键点 K 中的每个点 p_i，设置判定区域的半径为 r_k，将原本的体素特征 $\boldsymbol{F}^{(l_k)}$ 聚合至集合 $\boldsymbol{S}^{(l_k)}$ 中，如式（7）：

$$\boldsymbol{S}_i^{(l_k)}=\left\{[\boldsymbol{f}_j^{(l_k)};v_j^{(l_k)}-p_i]^T\middle|\begin{array}{l}|v_j^{(l_k)}-p_i|^2<r_k\\ \forall\, v_{v_j}^{(l_k)}\in\boldsymbol{V}^{(l_k)}\\ \forall\, f_j^{(l_k)}\in\boldsymbol{F}^{(l_k)}\end{array}\right\}\tag{7}$$

其中：使用 $|v_j^{(l_k)}-p_i|^2$ 表示体素坐标与关键点坐标的关联，保留相对 p_i 满足 $|v_j^{l_j}-p_i|^2<r_k$ 的距离条件的体

素。完成了将体素与关键点之间，在几何空间上的关联编码，得到每个关键点所分配到的体素区间。

其次，对体素特征进行关键点特征聚合。对于得到与关键点匹配的集合 $\boldsymbol{S}_i^{(I_k)}$，使用 PointNet 对其进行特征提取操作，从而使各个体素的特征聚合至与关键点维度相统一的聚合体素特征，得到 $\boldsymbol{f}_i^{(pv_k)}$，如式（8）所示：

$$\boldsymbol{f}_i^{(pv_k)} = max\{G(M(S_i^{(I_k)}))\} \tag{8}$$

其中：M 为采样操作，通过在 $S_i^{(I_k)}$ 采样 T_k 个体素，减少计算成本，G 为 MLP 层，最终采用最大池化操作。

经过以上两步的操作后完成了原点－体素编码操作，至此得到了与关键点维度结构相统一的体素特征 $\boldsymbol{f}_i^{(pv_k)}$。

3）集合抽象模块。为了捕捉从局部到全局的多尺度几何特征，从而进一步提高 3D 特征在后续处理任务的性能。使用集合抽象模块（SA，set abstraction）对单一特征进行分层采样和特征抽象。本节实现了集合抽象模块的两种变体对原始点与体素进行分层采样和特征抽象，分别为点集合抽象模块（PSA）和体素集合抽象模块（VSA）。

点集合抽象模块（PSA）通过对点的采样和分组，实现。首先，使用最远点采样（FPS）得到的关键点 $K=\{p_1, \cdots, p_n\}$ 作为采样点；随后，根据采样点，生成 n 个邻域，使用球形区域查询，得到半径内的固定个数的点；最终，使用 PointNet 对以上结果进行特征提取，对于每一个 p_i 得到特征 $\boldsymbol{f}_i^{(p)}$。

体素集合抽象模块（VSA）使用 3D 稀疏卷积进行下采样，减少了特征图的空间尺寸，得到更抽象、更丰富、更高维的特征表示。经过 4 次下采样后，分别得到 4 个维度不同的特征 $[\boldsymbol{f}_i^{(pv_1)}, \boldsymbol{f}_i^{(pv_2)}, \boldsymbol{f}_i^{(pv_3)}, \boldsymbol{f}_i^{(pv_4)}]$，随后通过特征拼接得到最终的特征向量 $\boldsymbol{f}_i^{(vw)}$，如式（9）：

$$\boldsymbol{f}_i^{(vw)} = [\boldsymbol{f}_i^{(pv_1)}, \boldsymbol{f}_i^{(pv_2)}, \boldsymbol{f}_i^{(pv_3)}, \boldsymbol{f}_i^{(pv_4)}], \quad \text{for } i=1,\cdots,n \tag{9}$$

但是由于集合抽象操作会使得特征空间存在一定的稀疏性和不均匀性，导致部分稀疏点云存在特征丢失。使用多尺度分组，通过设置多个邻域范围，在多个尺度上对点云的进行分组处理，如图 6。MSG 能够同时捕获大尺度的全局结构信息和小尺度的局部细节信息，从而提高点云分类、分割等任务的性能。

2.2.2 关键点特征转 BEV 特征

得到关键点 $K=\{p_1, \cdots, p_n\}$，其中每个点 p_i 都链接了关联的原始点特征 $\boldsymbol{f}_i^{(p)}$ 以及体素特征 $\boldsymbol{f}_i^{(vw)}$。为了便于后续的融合，本小节将对特征进行归纳与映射，进行 BEV 空间的投影转换。定义体素空间感受野 $RF=[0, -50, -10, 50, 50, 1]$，体素尺寸为 $[0.5, 0.5, 0.2]$。将该体素空间得到的特征，执行求和池化以在每一个 z 轴上创建一个张量，从而将体素最终投影至 BEV 视角。

2.2.3 3D 特征网络结构

所以设计的 3D 特征网络结构如图 6 所示。对于激光雷达点云，存在以下超参，最远点采样（FPS）设置的关键点个数 $n=2\ 048$。对原始点进行 PSA 操作，设置 $r_{k1}=0.4$、$r_{k2}=0.8$ 代表分别使用 0.4 m 与 0.8 m 两个半径尺度的采集

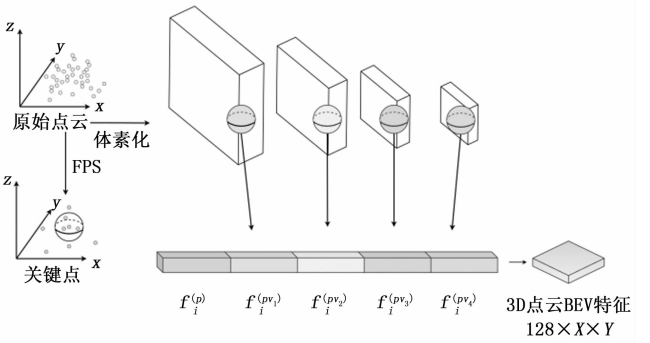


图 6 3D 点云转 BEV 特征编码网络

范围，获取特征进行多尺度分组得到尺寸为 2 048×16 的特征向量 $\boldsymbol{f}_i^{(p)}$。对体素进行 VSA 操作，初次下采样比例因子为 $scale=1$，设置 $r_{k1}=0.4$、$r_{k2}=0.8$ 表明第一层下采样感受野与原始点保持不变，使用相同的采集范围得到尺寸为 2 048×16 的特征向量 $\boldsymbol{f}_i^{(vw)}$；第二次下采样比例因子 $scale=2$，设置 $r_{k1}=0.8$、$r_{k2}=1.2$ 表明第二层下采样感受野扩大，扩大采集范围得到尺寸为 2 048×32 的特征向量 $\boldsymbol{f}_i^{pv_2}$；第三次下采样比例因子 $scale=4$，设置 $r_{k1}=1.2$、$r_{k2}=2.4$ 表明第三层下采样感受野扩大，继续增大采集范围得到尺寸为 2 048×64 的特征向量 $\boldsymbol{f}_i^{pv_3}$；第四次下采样比例因子 $scale=8$，设置 $r_{k1}=2.4$、$r_{k2}=4.8$，同理扩大采集范围得到尺寸为 2 048×128 的特征向量 $\boldsymbol{f}_i^{(pv_4)}$。

将得到的特征进行拼接后得到 256 维特征，通过特征降维，得到式（10）中聚合后的特征，最终尺寸为 2 048×128 的原点－体素 3D 关键点特征 $\boldsymbol{f}_i^{\text{gathered}}$。

$$\boldsymbol{f}_i^{(pv)} = [\boldsymbol{f}_i^{(p)}, \boldsymbol{f}_i^{(pv_1)}, \boldsymbol{f}_i^{(pv_2)}, \boldsymbol{f}_i^{(pv_3)}, \boldsymbol{f}_i^{(pv_4)}], \quad \text{for } i=1,\cdots,n \tag{10}$$

设置 BEV 网格的分辨率大小 $X\times Y$ 为 200×200，单元大小为 0.25×0.50 的体素。最后通过集合 $\boldsymbol{S}_i^{(I_k)}$，将原点－体素 3D 关键点特征 $\boldsymbol{f}_i^{\text{gathered}}$ 按照与体素的欧式距离，返回到原始输入 BEV 尺寸，输出 BEV 视角下的 3D 语义特征 $\boldsymbol{f}_{\text{bev}}^{(3D)}$，$\boldsymbol{f}_{\text{bev}}^{(3D)}$ 特征大小为 $[128\times200\times200]$。

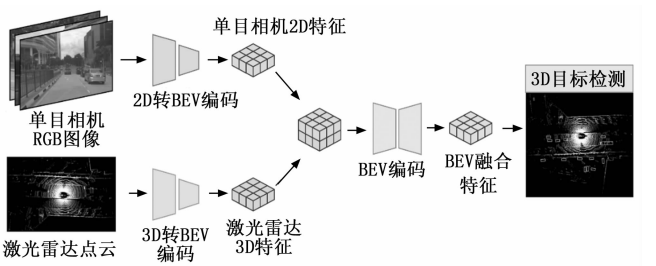


图 7 BEV 中的特征融合流程

2.3 融合相机与雷达的 3D 目标检测

为了解决数据映射过程中的几何失真问题，与映射后的数据优先级不对等问题，本节提出的 BEV 视场融合方案将 2D 图像特征与 3D 点云特征按各自的特性转换至共享的 BEV 视场，这使得图像与点云拥有平等的优先级，如图 7

所示。最终将图像的语义特征与点云的几何特征进一步融合，得到的融合特征能够通过 BEV 结构的多任务头，实现 3D 目标检测功能。

在 BEV 空间对 $\boldsymbol{f}_{\text{bev}}^{(2D)}$ 与 $\boldsymbol{f}_{\text{bev}}^{(3D)}$ 拼接后，表示为统一的 BEV 视角特征得到融合特征。$\boldsymbol{f}_{\text{bev}}^{(2D)}$ 提供了丰富的语义信息，以及在 BEV 空间下的粗略几何分布，$\boldsymbol{f}_{\text{bev}}^{(3D)}$ 提供了丰富的几何信息，以及在 BEV 空间下的精准几何分布。为了解决由于深度几何分布不准确导致的相机 BEV 特征和激光雷达 BEV 特征在空间上的错位问题，本节设计一个基于卷积神经网络（CNN）的编码器。首先，$\boldsymbol{f}_{\text{bev}}^{(2D)}$ 将和 $\boldsymbol{f}_{\text{bev}}^{(3D)}$ 的进行空间和通道上的静态融合。随后使用 SE 模块，在 Squeeze 阶段对每个通道使用全局平均池化，以此捕捉通道级的全局信息；在 Excitation 阶段，使用一个 Sigmoid 激活函数用于学习通道之间的非线性相互作用和依赖关系输出每个通道的重要性权重。包含残差块来进行特征融合和局部失准的校正。该编码过程对于拼接的特征进行修正后输出最终的融合特征 $\boldsymbol{f}_{\text{bev}}^{\text{fused}}$。最终使用 TransFusion 作为 BEV 检测头，将得到的特征 $\boldsymbol{f}_{\text{bev}}^{\text{fused}}$ 输入后进行 3D 多目标检测。TransFusion 中，首先使用热图头（Heatmap Head）生成每个类别的热图，用于目标的定位和分类；随后使用了一系列 Transformer 解码器层来处理通过共享卷积层预处理过的特征。这些解码器层利用自注意力机制来捕获不同特征之间的复杂关系，并且能够处理来自不同源的特征融合问题；最终针对每个 Transformer 解码器层的输出，有一个对应的预测头用于生成最终的检测结果，包括目标的类别、位置、尺寸等信息。

3　实验设置与结果分析

3.1　实验流程设计

由于本实验涉及对深度神经网络的模型训练以及推理，因此采用桌面级的计算平台对算法展开验证，配置如表 1 所示。

表 1　训练平台参数

名称	配置
CPU	Intel i9-13900K
GPU	NVIDIA RTX 3090 24G
驱动环境	CUDA 11.8
Python 版本	3.8.18

使用主流开放数据集 Waymo 开放数据集中的 Perception 系列数据集[21]对算法性能进行实验验证。Waymo 开放数据集是高质量、大规模的自动驾驶数据集之一，数据集覆盖了多种驾驶环境（城市街道、郊区道路和高速公路）、不同的天气条件和多样的交通情况，对感知性能具有更高的挑战性。因此，通过在 Waymo 开放数据集的感知数据集 v1.3.0 测试集上运行，得到最终的检测结果。为了得到对不同的算法，能有公平客观的性能比较。因此，检测结果将转换为 Waymo 官方要求的数据格式，并在官方给出的评估平台上进行性能评估。使用 Waymo 评估本文所提出的算法的检测精度。

同时使用实验室采集的数据集 ZJUT _ 1A 进行实际场景测试。ZJUT _ 1A 的原始数据由实验室自建的自动驾驶感知平台采集，并以 ROS 包的形式保存在本地。ZJUT _ 1A 包含 1 个完整的 3D 栅格地图、1 个结构化校园道路场景、1 个停车场场景和 1 个非结构化校园场景。为了满足对传感器融合的研究，感知平台包括了激光雷达、相机、GPS 和 IMU，如图 8 所示。通过 ZJUT _ 1A 数据集，并在 NVIDIA Jetson AGX 嵌入式 GPU 平台上对本文算法的实际性能进行评估。

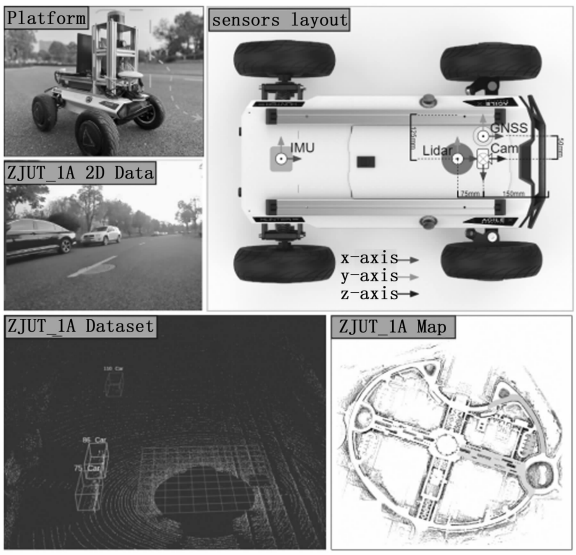


图 8　ZJUT _ 1A 数据集及采集平台

3.2　实验评价指标

为了测试所提出的融合方法的性能，使用 Waymo 开放数据集官方提供的评价指标，对提出的目标检测算法进行统一的评估。其中包含了在简单级别的 mAP/L1（L1 级总类平均精度），mAPH/L1（L1 级航向加权的总类平均精度），和在困难级别的 mAP/L2，mAPH/L2。由于 Waymo 开放数据集通过上传检测结果进行评估，因此对模型性能的评价，通过已知开源的模型参数量进行评价。

对于模型性能的评价，通过对比模型参数量给出了整体的运行性能评价。为了验证模型的实时性，设计实验在 Jetson AGX 嵌入式 GPU 平台上对部分算法的运算速度进行更具体的评估。

3.3　标定实验分析

通过实施单目相机与激光雷达的联合标定，准确地确定了二者的外参，相机图像与激光雷达测量数据映射在统一坐标系空间中。

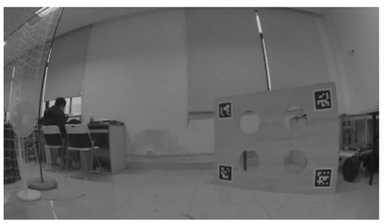
采用误差分析方法，将分别对不同的采集位姿数量的结果进行分析，如表 2 所示。标定过程中得到的单目相机与激光雷达的外部参数，角度与线性平移量，与它们在实际使用环境中的真实位姿进行对比，以此计算二者之间的误差值，如表 2。该步骤能够直观地展示标定准确度。

随后，执行视觉与雷达数据融合实验，通过将激光雷达扫描得到的点云数据映射到单目相机捕获的图像中，实

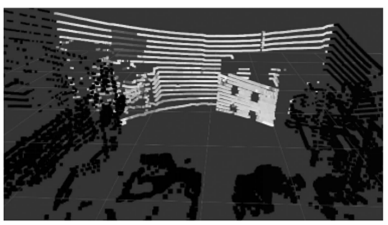
现点云的彩色化。这一过程不仅增强了点云数据的可解释性，而且有效地验证了联合标定的准确性与实用价值。通过视觉信息赋予点云色彩，能够更直观地评估相机与激光雷达之间的空间对应关系，进而验证标定结果的有效性，如图 9 所示。

表 2 不同目标位置数量的平均校准误差

位置数量	角度误差/rad	线性误差/cm
$N=1$	0.24	12.00
$N=2$	0.98	3.30
$N=3$	0.04	0.89
$N=4$	0.03	0.72



(a) 相机标定板图像



(b) 有色标定板点云

图 9 不同位姿采集数量的平均校准误差

3.4 3D 目标检测实验分析

1) Waymo 3D 目标检测验证结果。提出的融合方法用于目标检测，提交至 Waymo 开放数据集官方平台后，得到了最终的检测精度结果。表 3 显示了最终的结果。结果显示，相比于单纯基于激光雷达的检测方法的精度（mAP/L1：0.788 2），本文的融合方法拥有更高的检测精度（mAP/L1：0.820 3）。其次，通过对比困难难度与简单难度的精度结果，本文提出的融合方法具有一定的鲁棒性，在面对更加复杂的场景下也拥有更高的检测精度。最终，通过与传统的基于融合的方法相比（mAPH/L1：0.788 2），在基于航向角加权的精准性上，具有更高的性能（mAPH/L1：0.804 8）。证实了本文的融合方法在融合过程中，放弃了将相机或是激光雷达的数据进行单模态的投影，而是使用了统一的 BEV 空间进行融合的有效性。但本文 BEV 融合方法的精度相比于 SOTA 的融合方法，还存在差距，通过模型文件对比，可以看出所提出的方法拥有更加轻量化的体积（模型参数量为 SOTA 的 47.1%）。

表 3 WOD 数据集 3D 目标检测结果表

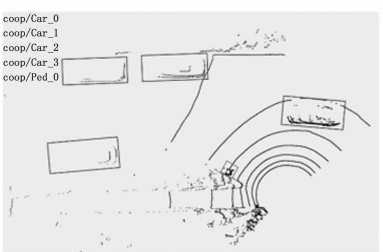
方法	传感器类型	mAP /L1	mAPH /L1	mAP /L2	mAPH /L2	模型参数量/MB
PointPillars[14]	L	0.455 2	0.412 0	0.406 3	0.367 2	18.45
CenterPoint[22]	L	0.787 1	0.771 8	0.733 8	0.719 3	35.3
PV-RCNN[18]	L	0.788 2	0.769 0	0.733 5	0.715 2	155
Deepfusion[23]	C+L	0.818 9	0.804 8	0.769 1	0.755 4	—
BEVfusion[24]	C+L	0.860 4	0.847 6	0.812 2	0.799 7	159
Ours	C+L	0.820 3	0.812 2	0.778 5	0.764 8	74.9

2) 目标检测实时性验证结果。为了对模型性能，拥有更加直观的对比，设计实验在英伟达 AGX 平台上，使用与模型训练相同的环境进行 3D 目标检测，如表 4 所示。在 ZJUT_1A 数据集下，CenterPoint 的平均运行帧率为 26 帧每秒，但本文的融合方法进行目标检测的帧率仅能达到 10 帧。为了满足实时需求，将融合最终的对齐编码简化，并将原先的 TransFusion 检测头替换为更加轻量化的 CenterPoint 检测头，保证在 NVIDIA Jetson AGX 嵌入式 GPU 平台上实现足够的目标检测精度，最终平均运行帧率达到 19 帧每秒。

表 4 嵌入式系统运行效率对比

Method	Speed/fps	Accuracy(ZJUT_1A)
CenterPoint	26	0.787 1
Ours-origin	10	0.820 3
Ours-simplify	19	0.809 7

(a) 相机视角



(b) BEV视角感知结果

图 10 Jetson AGX 嵌入式平台道路目标检测结果

3.5 消融实验分析

本文提出的 BEV 融合方法分为 3 个主要模块，2D 图像的 BEV 空间特征编码，3D 点云的 BEV 空间特征编码，以及 BEV 融合编码。通过对不同组件施行消融实验，验证每个组件的有效性。结果如表 5 所示。其中，单独使用 2D 图像在 BEV 空间的特征进行 3D 目标检测精度最低，由于 2D 图像主要包含语义特征，缺少几何特征。相对地单独使用 3D 点云在 BEV 空间的特征能够完成一定的 3D 目标检测任

务。表中使用 2D 与 3D 融合的方法结果优于单模态方法，表明 2D 与 3D 在统一 BEV 空间变换后进行融合的有效性。最终在 BEV 空间增加了编码操作，应对两组 BEV 特征在空间上的错位问题，能够一定程度地提高最终的检测精度。

表 5　消融实验表

LLS 2D encoder	PV-RCNN 3D encoder	BEV encoder	$mAPH$/L2
√			0.226 1
	√		0.715 2
√	√		0.755 4
√	√	√	0.764 8

通过消融实验，显示了本文的 3D 目标检测算法在不同传感器配置下的可行性。在复杂的交通场景下，当相机或雷达受到干扰失灵情况下，本文的方法依旧有效。

4　结束语

本文针对自动驾驶车辆对道路环境的 3D 目标检测，提出了一种基于 BEV 视角的多传感融合 3D 目标检测算法。通过将 2D 图像的语义特征与 3D 点云的几何特征转换至统一的 BEV 空间进行融合，处理传统融合过程中的几何失真与信息优先级不对问题。最终在 WOD 开放数据及上进行验证。实验结果表明，基于 BEV 视角的多传感融合 3D 目标检测算法相比于单模态的检测算法具有正优化，同时在与其他融合算法的横向比较中，具有一定的优势。本文的融合方案在面对传统相机与激光雷达传感器融合过程中的几何失真，以及信息优先级不对等，具有更高的鲁棒性。

参考文献：

[1] 任柯燕，谷美颖，袁正谦，等. 自动驾驶 3D 目标检测研究综述 [J]. 控制与决策，2023，38 (4)：865－889.

[2] 梅玲玲. 基于深度学习的道路车辆目标检测系统设计 [J]. 计算机测量与控制，2023，31 (2)：83－90.

[3] VORA S，LANG A H，HELOU B，et al. Pointpainting：Sequential fusion for 3d object detection [C] //IEEE Conference on Computer Vision and Pattern Recognition，2020：4604－4612.

[4] WANG C，MA C，ZHU M，et al. Point augmenting：cross-modal augmentation for 3d object detection [C] //IEEE Conference on Computer Vision and Pattern Recognition，2021：11794－11803.

[5] CHEN X，MA H，WAN J，et al. Multi-view 3d object detection network for autonomous driving [C] //IEEE Conference on Computer Vision and Pattern Recognition，2017：1907－1915.

[6] KU J，MOZIFIAN M，LEE J，et al. Joint 3d proposal generation and object detection from view aggregation [C] //IEEE Conference on Intelligent Robots and Systems，2018：1－8.

[7] PANG S，MORRIS D，RADHA H. CLOCs：Camera-LiDAR object candidates fusion for 3D object detection [C] //IEEE Conference on Intelligent Robots and Systems，2020：10386－10393.

[8] PANG S，MORRIS D，RADHA H. Fast－CLOCs：fast camera-LiDAR object candidates fusion for 3D object detection [C] // IEEE/CVF Winter Conference on Applications of Computer Vision，2022：187－196.

[9] 党亚南，田照星，郭利强. 车载激光雷达点云数据处理关键技术 [J]. 计算机测量与控制，2022，30 (1)：234－238.

[10] QI C R，SU H，MO K，et al. Pointnet：Deep learning on point sets for 3d classification and segmentation [C] //IEEE Conference on Computer Vision and Pattern Recognition，2017：652－660.

[11] QI C R，YI L，SU H，et al. Pointnet＋＋：Deep hierarchical feature learning on point sets in a metric space [J]. Advances in Neural Information Processing Systems，2017，30：5099－5108.

[12] SHI S，WANG X，LI H. Pointrcnn：3D object proposal generation and detection from point cloud [C] //IEEE Conference on Computer Vision and Pattern Recognition，2019：770－779.

[13] YAN Y，MAO Y，LI B. Second：sparsely embedded convolutional detection [J]. Sensors，2018，18 (10)：3337.

[14] LANG A H，VORA S，CAESAR H，et al. PointPillars：fast encoders for object detection from point clouds [C] //IEEE Conference on Computer Vision and Pattern Recognition，2019：12689－12697.

[15] HUANG J，HUANG G，ZHU Z，et al. Bevdet：High-performance multi-camera 3D object detection in bird-eye-view [EB/OL]. Arxiv Preprint Arxiv：2112. 11790，2021.

[16] HUANG J，HUANG G. Bevpoolv2：A cutting-edge implementation of BEVDET toward deployment [EB/OL]. Arxiv Preprint Arxiv：2211. 17111，2022.

[17] PHILION J，FIDLER S. Lift，splat，shoot：Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d [C] //European Conference on Computer Vision，2020：194－210.

[18] SHI S，GUO C，JIANG L，et al. Pv-RCNN：Point-voxel feature set abstraction for 3d object detection [C] //IEEE Conference on Computer Vision and Pattern Recognition，2020：10529－10538.

[19] DENG J，DONG W，SOCHER R，et al. Imagenet：a large-scale hierarchical image database [C] //IEEE Conference on Computer Vision and Pattern Recognition，2009：248－255.

[20] KOONCE B，KOONCE B. EfficientNet [J]. Convolutional Neural Networks with Swift for Tensorflow，2021：109－123.

[21] SUN P，KRETZSCHMAR H，DOTIWALLA X，et al. Scalability in perception for autonomous driving：Waymo open dataset [C] //IEEE Conference on Computer Vision and Pattern Recognition，2020：2446－2454.

[22] YIN T，ZHOU X，KRAHENBUHL P. Center-based 3D object detection and tracking [C] //IEEE Conference on Computer Vision and Pattern Recognition，2021：11784－11793.

[23] LI Y，YU A W，MENG T，et al. Deepfusion：lidar-camera deep fusion for multi-modal 3d object detection [C] //IEEE Conference on Computer Vision and Pattern Recognition，2022：17182－17191.

[24] LIU Z，TANG H，AMINI A，et al. Bevfusion：multi-task multi-sensor fusion with unified bird's-eye view representation [C] //International Conference on Robotics and Automation，2023：2774－2781.