计算机工程
Computer Engineering
ISSN 1000-3428,CN 31-1289/TP

《计算机工程》网络首发论文

题目：HDMapFusion：用于自动驾驶的多模态融合高清地图生成
作者：刘洋宏，付杨悠然，董性平
DOI：10.19678/j.issn.1000-3428.0070569
网络首发日期：2025-04-29
引用格式：刘洋宏，付杨悠然，董性平．HDMapFusion：用于自动驾驶的多模态融合高清地图生成[J/OL]．计算机工程.
https://doi.org/10.19678/j.issn.1000-3428.0070569

计 算 机 工 程
Computer Engineering

HDMapFusion：用于自动驾驶的多模态融合高清地图生成

刘洋宏，付杨悠然，董性平*

(武汉大学计算机学院，国家多媒体软件工程技术研究中心，人工智能学院，多媒体网络通信工程湖北省重点实验室，湖北 武汉 430072)

摘要：高清(HD)环境语义地图的生成是自动驾驶系统中不可或缺的关键技术。针对相机与激光雷达在感知任务中存在的模态差异问题，本文提出了一种创新的多模态融合范式 HDMapFusion。与传统的直接融合原始传感器数据方法不同，本方法通过将相机和激光雷达特征统一转化为鸟瞰图(BEV)表示，实现了多模态信息的物理可解释性融合。在 nuScenes 基准数据集上的实验结果表明，HDMapFusion 在 HD 地图生成精度方面显著优于现有基准模型，其中 IoU 得分提升了 23.0%，充分验证了该方法的有效性和优越性。

关键词：HD 地图生成；多模态融合；BEV 表示；自动驾驶；深度估计

DOI：10.19678/j.issn.1000-3428.0070569

HDMapFusion: HD Map Generation with Multi-Modality Fusion for Autonomous Driving

LIU Yanghong, FUYANG Youran, DONG Xingping*

(School of Computer Science, National Engineering Research Center for Multimedia Software, Institute of Artificial Intelligence, Hubei Key Laboratory of Multimedia and Network Communication Engineering, Wuhan University, Wuhan 430072, China)

【Abstract】The generation of high-definition (HD) environmental semantic maps plays a crucial and irreplaceable role in autonomous driving systems. To address the modality discrepancy between cameras and LiDAR in perception tasks, this paper proposes an innovative multi-modal fusion paradigm, HDMapFusion. Unlike traditional methods that directly fuse raw sensor data, this approach achieves physically interpretable fusion of multi-modal information by unifying camera and LiDAR features into a bird's-eye view (BEV) representation. Experimental results on the nuScenes benchmark dataset demonstrate that HDMapFusion significantly outperforms existing baseline models in HD map generation accuracy, with a 23.0% improvement in IoU score, fully validating the effectiveness and superiority of the proposed method.

【Key words】HD map generation; multi-modality fusion; BEV representation; autonomous driving; depth estimation

0 引言

自动驾驶技术的发展离不开包含海量语义信息的高清(HD)地图生成，相关研究显示了语义地图构建的良好应用前景[1,2,3]。早期研究[4]提出了另一种解决方案，构建用于处理自动驾驶任务(例如轨迹跟踪[5]和运动规划)的本地 HD 地图。然而，语义地图构建的性能对于安全高效的自动驾驶来说仍然不够准确。

要保证自动驾驶的安全性和高效性，一个至关重要的条件是能够实时准确感知自车的周围环境。这需要一个可靠的感知系统，能够在多样化和复杂的驾驶场景中，鲁棒地检测和识别各种动态和静态的道路元素，如行人、车辆和道路标记。高清地图生成可以近似理解为在鸟瞰(BEV)视图下对道路元素进行语义分割。BEV 地图提供了一个自上而下的场景视图，便于检测和跟踪物体，同时规划并控制车辆的轨迹。近几年特斯拉主导的纯视觉自动驾驶方案正在备受关注，

基金项目：中央高校基本科研业务费专项资金资助（2042024kf0036）；国家自然科学基金委项目（62471342）；澳门特别行政区科学技术发展基金（001/2024/SKL）、智慧城市物联网国家重点实验室（澳门大学）开放课题（编号：SKL-IoTSC(UM)-2024-2026/ORP/GA04/2023）。

通信作者 E-mail：xingpingdong@whu.edu.cn

然而在纯视觉的技术路线下，高清地图生成由于缺乏空间的几何信息仍然是一个具有挑战性的问题。

激光雷达和摄像头融合的技术路线具有以下研究意义：(1) 弥补摄像头的局限性，摄像头在光照变化（如强光、夜晚）和天气条件（如雨雾）下可能会出现性能下降，而激光雷达在这些场景中具有更强的鲁棒性；(2) 技术路线的互补性，纯视觉路线是对融合感知的简化方向，而融合技术可以为未来更高鲁棒性的感知系统提供技术储备。然而，由于传感器数据固有的稀疏性和差异性，如何有效利用激光雷达和摄像头感知得到的数据准确高效地分割语义地图仍然是一个具有挑战性的问题。

为了解决这个问题，最近的研究已经探索了多模态融合的实现，结合来自多个输入的信息，如激光雷达、相机和雷达传感器的信息，以提高 3D 检测性能的鲁棒性和准确性[6,7]。通过利用不同模态的互补优势，多模态融合可以有效增强感知系统处理复杂动态驾驶场景的能力。但是如何有效地建模、表征并融合多模态感知信息，从而获得全面的语义和空间感知特征成为目前算法仍需解决的问题。

通常多模态融合方法会存在以下两种问题：(1) 模态特征不一致性，不同模态的特征在分辨率、语义层次上存在差异，传统融合方法难以充分利用其互补性；(2) 信息冗余与噪声问题，直接融合可能引入冗余信息或噪声，需要设计动态权重分配机制以提升融合效果。本文通过设计基于注意力的动态融合机制，解决了上述问题，并实现了对不同模态特征的高效利用，提出了一种基于多模态融合机制的 HD 语义地图生成方法。具体来说，多模态输入利用注意力机制[8]将关于 3D 场景的全局上下文推理集成到多层特征提取中。与相机相比，激光雷达的主要优势之一是点云包含精确的 3D 位置信息。作者还建议将激光雷达视为有监督的训练，以协助相机进行深度估计和 BEV 转换。作者设计了一种策略，将图像从透视视图转换为 BEV 表示，总结为知识蒸馏，类似于“教师-学生”模式。在城市驾驶场景的大规模数据集上证明了所提出方法的有效性，并表明它在分割精度和效率方面优于最先进的方法。综上所述，总结出本文的主要贡献为三个方面。

-本文提出了一种端到端的高清地图生成范式，用于在 BEV 表示下绘制具有多模态融合的 HD 语义图，以利用相机和激光雷达得到的感知数据进行互补优势。

-本文受到知识蒸馏的启发，设计了一种 BEV 转化策略。通过点云数据所蕴含的深度信息将激光雷达作为监督老师，协助相机进行深度估计和 BEV 转换。

-本文在大规模自动驾驶 nuScenes[9]数据集上进行了定性和定量实验，验证了本文提出方法的有效性，并在[4]提出的综合评价指标上取得了显著提升。

1 相关研究

1.1 语义地图构建

近年来，语义地图构建领域提出了很多新方法，取得了重大进展。其中一种主流的方法是使用激光扫描仪或 LIDAR 传感器来生成周围环境的 3D 点云。Hao 等人[10]提出了一种利用激光雷达（LIDAR）数据和卷积神经网络(CNN)将点云划分为车道、人行道、建筑物等不同类别的方法来构建三维语义地图。另一种方法是使用相机或 RGB-D 传感器捕捉环境的图像或深度数据。Sun 等人[11]提出了一种使用卷积神经网络和条件随机场相结合的深度学习框架对 RGB-D 图像进行语义分割的方法。然而，这些方法未能充分利用传感器数据之间的语义和空间特性，HD 地图生成的精度有待进一步提高。

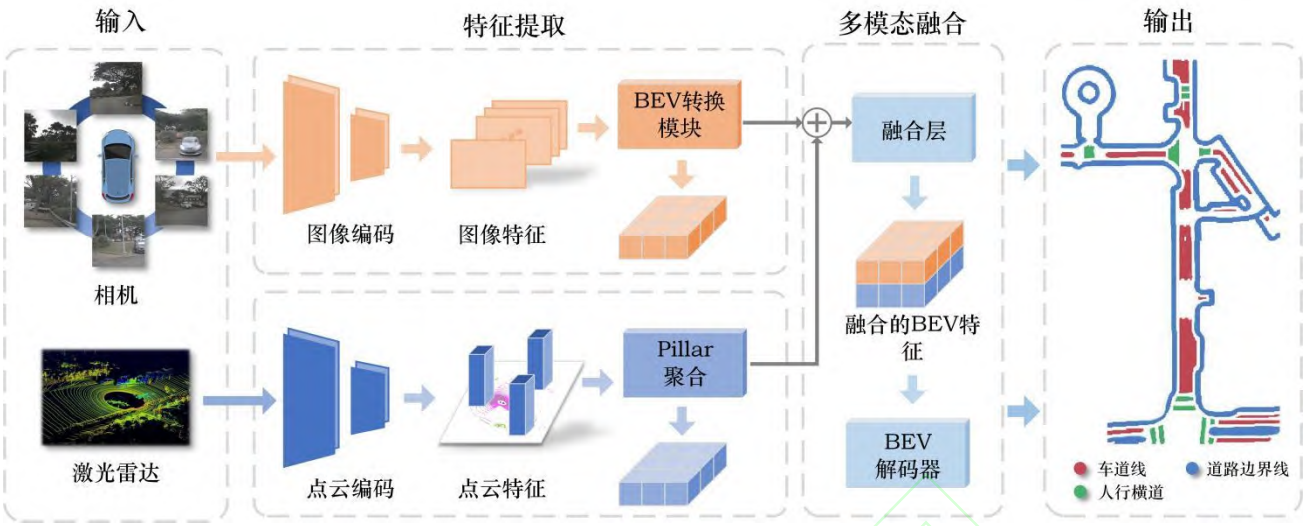

图 1　HDMapFusion 结构

Fig.1　Architecture of HDMapFusion

1.2 鸟瞰视角表示

鸟瞰图（BEV）是一种视野开阔的自上而下的视角，BEV 地图同样可以用作自动驾驶的驾驶地图。此外，BEV 为融合前统一多模态数据的表示提供了一种有效的方法。为了改进图像到 BEV 的表示，Hu 等人[12]开发了一种将过去的 3D 特征到当前自我视图的对齐特征的累积技术。Rhinehart 等人[13]提出了一种基于 LIDAR 的方法，将点云转换为 2D BEV 网格，根据目标来预测车辆轨迹。总的来说，BEV 提供了一种物理可解释的方法，用于融合来自不同视角和模式的信息，从而成为一种潜在的可扩展的自动驾驶解决方案。

1.3 多模态融合

近年来，多模态传感器融合方法[5,7]在感知任务中愈发普及。Xu 等人[14]提出了一种方法，使用多级融合算法组合来自多个传感器的数据，以创建一个未知环境的地图。Chen 等人[15]将激光雷达和 RGB 图像获得的多模态特征以多尺度方式融合，用于 3D 目标检测。然而，每个模态都拥有固有的数据表示，多模态直接操作融合机制可能导致分割性能不足。本文发现了这种限制，并在不同模态之间采用了统一的表示，以促进多模态融合的实现。

2 方法

如图 1 所示，本节详细介绍了一种基于 BEV 表示的多模态融合高清地图生成的新范式。模型详细设计如下：

2.1 输入和输出表示

输入表示：在这个任务中，给定可用的相机由全景图像 $\left(I_i, K_i, R_i, T_i\right)_{i=1}^{n}$ 组成。其中第 i 个相机的输入图像 $I_i \in \mathbb{R}^{H\times W\times 3}$，内部参数 $K_i \in \mathbb{R}^{3\times 3}$，外部旋转 $R_i \in \mathbb{R}^{3\times 3}$，相对于自我载体中心的平移 $T_i \in \mathbb{R}^{3}$。小车配备了 6 个不同朝向的摄像头(前、前左、前右、后、后左、后右)，以实现对周围环境的全面理解。激光雷达输入。原始激光雷达数据由具有稀疏特征的三维点云组成，这类数据缺乏图像固有天赋等语义信息。本文参考了之前的工作[13]，将 3D 激光雷达点云转换为图 2 所示的 2D BEV 网格上的 2-bin 直方图，以匹配相机输入。BEV 网格是一个 50 米形状的正方形区域 ×50m。在 LSS[16]操作之后，网格被分成 0.5m×0.5m 块，在 BEV 表示中提供 200×200 像素的分辨率。

输出表示：将多模态特征转移到 BEV 空间中进行融合机制，再由解码器模块学习，在正射影地图-视图坐标框架中预测多个二元语义分割掩码 $z \in \{0,1\}^{H_{bev}\times W_{bev}\times C_{class}}$。

2.2 多模态特征统一表示

特征提取：要从周围的 N 个摄像机中提取图像特征 $C=\left\{I_1,\ldots,I_N \middle| I_i\in\mathbb{R}^{H\times W\times 3}, i\in 1,\ldots,N\right\}$，本文递归地利用在 ImageNet[17]上预训练的高效 CNN 骨干 ResNet[18]来创建修订分辨率的特征图 $f\in\mathbb{R}^{H_f\times W_f\times C}$。每个相机都嵌入在一个参数共享编码器中。这些中间特征包含用于其他层的潜在显著因子。令 $P=\left\{(x,y,z,r)_{i\in 1,\ldots N}\right\}$ 表示一组无序点云，(x,y,z) 反映三维位置，r 表示反射强度。将三维激光雷达点云输入按照 x 轴和 y 轴分成规则网格，将具有相同网格的点云作为 PointPillars 之后维度为 (C,P,N) 的柱子[19]。所有的柱子都被 PointNet[20]编码以提取点特征，然后量化成网格大小为 $H_v\times W_v\times 1$ 的体素箱。这些体素被 pillar backbone 编码，生成体素特征 $f_v\in\mathbb{R}^{H_v\times W_v\times C}$。由于本文的研究重点是如何高效融合不同模态间的特征来提高高清地图的生成质量，所以在对图像和点云数据进行特征提取时的 backbone 选择，我们参考并使用先前工作中已被广泛使用证实的 ResNet 和 PointPillars 作为图像和点云特征提取的骨干网络。

2.3 基于点云监督的相机 BEV 转化

隐式监督深度估计：在这项研究中，本文提出了一种基于点云的监督策略，旨在完成 2D 图像的 BEV 投影转换。遵循 CroMA[21]中概述的方法，为每个输入图像构建了场景的 3D 体素表示。为了实现这一点，本文利用深度头来估计输入图像中每个像素的深度分布。将深度轴离散化为 N_d bins，图像的每个像素被扩展为多个体素，导致输入图像的每个像素被转换为 3D 坐标 $(h,w,d)\in\mathbb{R}^3\,|\,d\in N_d$ 表示的体素。为了获得给定像素 $x=(h,w)$ 在不同深度 bins 下的一组 N_d 体素，本文使用以下公式进行计算：

$$V=\left\{v_i=M^{-1}\left[d_i p_h, d_i p_w, d_i\right]^T \middle| i\in\left\{1,\ldots,N_i\right\}\right\} \quad (1)$$

公式中，M 表示相机矩阵，d_i 是第 i 个深度箱的深度值，p_h 和 p_w 分别是给定像素的高度和宽度值。深度值 d_i 可以从激光雷达输入或深度估计中获得。需要注意的是，V 代表一个较大的点云。

在获得每个像素的体素特征 v_i 后，本文利用 PointPillars[19]将体素向量 V 投影到 BEV 上，并对所有特征进行聚合生成 BEV 特征。这些 BEV 特征会通过一个包含基于 cnn 的架构的解码器，用于创建 BEV 地图。BEV 地图中的每个元素对应于一个 2D 网格 (x,y)。为了构建每个网格的特征，本文对投射到其中的所有 3D 体素的特征应用平均池化。

该模块经过训练，可以在激光雷达输入的指导下绘制精确的深度信息，就像老师一样。深度估计头利用图像特征来估计深度，然后在点云的监督下与激光雷达提供的地面真实值进行比较。一旦获得每个像素的深度，本文将其与嵌入的特征结合起来，将其投影到 BEV 表示中。虽然激光雷达有助于深度估计的中间过程，但本文的主要目标是将多模态特征与 BEV 表示相结合。为了提高从视角到 BEV 空间的图像投影精度，本文提出使用预训练的激光雷达 oracle 模型来监督生成的 BEV 特征图。这个 BEV 解码器在最后的特征嵌入阶段操作，这导致更接近最终目标的映射。

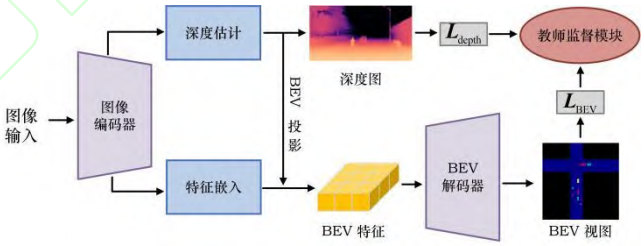

图 2　BEV 转换流程图

Fig.2　Overview the procedure of BEV transformation

2.4 多模态特征融合

考虑到模型输入多模态传感器中的数据并将其转换为统一的 BEV 表示，只要适当地合并，就可以充分理解两模态之间的潜在信息，从而有效地生成 BEV 语义图。由于相机和激光雷达数据特征的互补性，为了结合两者的优势，本文利用 transformer 的注意力机制[8]来融合多模态特征的全局语境。transformer 的输入被视为由图像和激光雷达特征向量组成的 token 序列。在每个 token 中添加一个可学习的位置编码，以记住每个传感器输入中的位置信息。采用自然语言

处理(NLP)任务中大火的 GPT[22]作为多模态特征融合模块，所有的融合层都是基于本文在多个尺度上设计的分层结构。GPT 则是由多个 transformer_block 堆叠而成，其融合层内部的详细网络架构如图 3 所示。

2.5 BEV 解码器

融合的 BEV 特征压缩了图像语义信息和点云几何信息，因此需要对其进行解码以生成自上而下表示的高清语义地图，也就是 BEV 视图。与 ResNet[18]等常用主干来组合不同分辨率的特征[16]不同，本文认为没有必要采用这样的主干来进行复杂的多特征提取计算。本文设计了一种基于一致的上采样和卷积的架构，以便最终的地图大小可以恢复到 $H_{bev}\times W_{bev}\times C_{class}$。通过残差连接，可以高精度地感知一些在 BEV 空间中定义的关键线索，如规模、类别和位置。此外，本文添加了另外两个预测头来对语义 BEV 地图进行矢量化构建[4]，包括实例嵌入预测和车道方向预测。

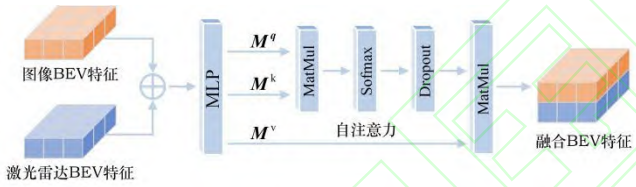

图 3 融合层模型结构

Fig.3 The details of fusion layer

2.6 损失函数

监督损失：激光雷达点云的输入提供了准确的深度信息，并作为教师监督引导图像学习深度映射。此外，为了对齐投影的 BEV 特征，本文利用头部生成 BEV 地图与激光雷达 BEV 地图的比较。因此，监督损失由深度损失和 BEV 损失两部分组成。深度损失 L_{dp} 由 L_1 准则计算，如公式(2)。BEV 损失 L_{bev} 可以表示为公式(3)中学习到的 BEV 投影 $\hat{I}$ 与地面真实激光雷达 BEV 地图 I 之间的交叉熵损失。

$$L_{\mathrm{dp}}=\sum_{n=1}^{N}\left\|V_n-V_n^{gt}\right\|_1 \tag{2}$$

$$L_{bev}=-\sum_{i=1}^{H}\sum_{j=1}^{W}\sum_{k=1}^{C}\sum_{n=1}^{N}I_{(i,j,k)}^{n}\log\hat{I}_{(i,j,k)}^{n} \tag{3}$$

辅助损失：遵循 HDMapNet[4]，本文将实例嵌入预测的损失 L_{ins} 定义为方差和距离损失 L_{dist}：

$$L_{\mathrm{var}}=\frac{1}{C}\sum_{c=1}^{C}\frac{1}{N_c}\sum_{j=1}^{N_c}\left[\left\|\mu_c-f_j^{\mathrm{instance}}\right\|-\delta_v\right]_+^2 \tag{4}$$

$$L_{dist}=\frac{1}{C(C-1)}\sum_{c_A\neq c_B\in C}\left[2\delta_d-\left\|\mu_{c_A}-\mu_{c_B}\right\|\right]_+^2 \tag{5}$$

$$L_{ins}=\alpha L_{var}+\beta L_{dist} \tag{6}$$

其中 C 测量聚类的数量，N_c 表示聚类 C 中的元素数量，μ_c 对应于均值嵌入。f_j^{instance} 是 c 中的第 j 个元素的嵌入，$\|\cdot\|$ 是 L_2 范数，$[x]_+=\max(0,x)$，δ_v 和 δ_d 是方差和距离损失的间隔。对于方向预测，本文将方向离散为一个圆上均匀的 36 个类，并将损失 L_{dir} 定义为交叉熵损失。本文只对那些位于车道上具有有效方向的像素进行反向传播。

总损失：模型通过加权结合监督损失和辅助损失进行端到端的训练，因此本文得出总训练损失如下：

$$L_{total}=\lambda_{dp}L_{dp}+\lambda_{bev}L_{bev}+\lambda_{ins}L_{ins}+\lambda_{dir}L_{dir} \tag{7}$$

其中 λ_{dp}、λ_{bev}、λ_{ins} 和 λ_{dir} 为权重系数，具体数据在 4.1 节中设置。

3 实验

3.1 基准模型

为了全面对比本文提出的多模态融合方法的有效性，我们设置了三组不同的方法作为基准模型，分别有：仅相机输入、仅激光雷达输入和多模态输入[4]。具体选择的基准模型有：

IPM [23]：这是一种直接将语义分割结果映射到 BEV 视角下的方法。LSS [16]：该方法在 2D 图像空间下对每个像素进行深度估计，将 2D 特征转为 3D 特征再映射到 BEV 空间下做道路语义分割。VPN [24]：提出了一种简单的视图转换模块，包含视图关系模块用于对任意两个像素之间的关系进行建模以及视图融

合模块用于融合像素的特征。PointPillars [19]：是一种将点云数据转换为伪图片的 3D 目标检测模型。HDMapNet [4]：将 2D 图像和 3D 点云分别编码后进行特征融合，通过 BEV 解码器得到语义分割地图。

3.2 实验设置

实验软件平台基于Ubuntu18.04 和 python3.8 编写。网络的多模态输入主要包括全景相机和激光雷达与巨大的 3D 点云。本文用 pointpillar[19]处理点云，并利用 PointNet[20]将点聚合成一个柱子。所有图像数据都使用备选图像主干 ResNet[18]进行特征提取。为了使本文的输入更通用和鲁棒，本文随机地围绕自车将输入数据旋转 10° 以进行数据增强[25]。以自车 $[-50,50,0.5]\times[-50,50,0.5]$ 的范围为中心设置语义 BEV 地图，生成的分辨率为 $H_{bev}\times W_{bev}=200\times 200$。本文主要集中在三种类型的地图元素 $C_{class}=3$，包括车道分割线、人行横道以及车道边界线。模型使用 AdamW 优化器[26]进行训练，batch 大小为 16，初始学习率为 1×10^{-3}，在单个 3090 GPU 上进行 200 个 epoch。对于损失系数的权重，本文设置 $\lambda_{dp}=1.0$，$\lambda_{bev}=1.0$，$\lambda_{ins}=0.5$ 和 $\lambda_{dir}=0.5$。另外，变量和系数与[4]一致，本文设置 $\delta_v=0.5$，$\delta_d=3.0$ 和 $\alpha=1.0$，$\beta=1.0$。

表 1　nuScenes 数据集上语义 BEV 地图的交并比分数(%)

Table 1　IoU scores (%) of semantic BEV map on nuScenes dataset

方法	模态		交并比			平均交并比
	相机	激光雷达	车道分界线	人行横道	车道边界	所有对象
IPM [23]	√	-	14.4	9.5	18.4	14.1
LSS [16]	√	-	38.3	14.9	39.3	30.8
VPN [24]	√	-	36.5	15.8	35.6	29.3
PointPillars [19]	-	√	41.5	26.4	53.6	40.5
HDMapNet [4]	√	√	46.1	31.4	56.0	44.5
SuperFusion [10]	√	√	47.9	37.4	58.4	47.9
HDMapFusion (Ours)	√	√	**58.0**	**52.7**	**66.0**	**58.9**

表 2　nuScenes 数据集上的实例检测结果

Table 2　Instance detection results on nuScenes dataset

方法	模态		交并比			平均交并比
	相机	激光雷达	车道分界线	人行横道	车道边界	所有对象
IPM [23]	√	-	19.6	7.8	23.7	17.0
LSS [16]	√	-	35.9	8.9	41.2	28.7
VPN [24]	√	-	34.9	9.0	42.7	28.9
PointPillars [19]	-	√	40.6	18.7	49.3	36.2
HDMapNet [4]	√	√	46.0	17.8	**65.6**	43.1
SuperFusion [10]	√	√	48.2	30.3	58.0	45.5
HDMapFusion (Ours)	√	√	**52.3**	**47.2**	60.6	**53.4**

3.3 数据集

本文在大规模自动驾驶数据集 nuScenes[9]上进行了全面的实验，该基准 nuScenes 包含来自真实交通场景的高质量传感器数据（超过 1000 个场景，每个场景使用各种传感器捕获，包括激光雷达、毫米波雷达和摄像头）。这些场景被记录在各种天气和照明条件下，以及在不同类型的城市环境中，如城市中心、高速公路和住宅区。

3.4 评价指标

交并比(IoU)。为了直观地反映地图视图下模型的语义分割性能，生成的地图 M_{pred} 和 M_{gt} 之间的

每一个像素按如下方式计算:

$$\mathrm{IoU}\left(M_{pred},M_{gt}\right)=\frac{\left|M_{pred}\cap M_{gt}\right|}{\left|M_{pred}\cup M_{gt}\right|} \tag{8}$$

平均精度(AP)。在目标检测中广泛使用的平均精度(average precision, AP)[27]衡量的是实例检测能力，定义为:

$$\mathrm{AP}=\frac{1}{10}\sum_{r\in\{0.1,0.2,\ldots,1.0\}}\mathrm{AP}_r \tag{9}$$

$$\mathrm{CD}=\frac{1}{C_{pred}}\sum_{a\in C_{pred}}\min_{b\in C_{gt}}\left\|a-b\right\|_2 \tag{10}$$

给定预测值C_{pred}和地面真值C_{gt}的两条曲线，本文使用 Chamfer distance (CD)来计算两条曲线之间的空间距离。如文献[4]中介绍作者使用 CD 来选择真正实例。本文最初为 CD 设置一个阈值，并与计算出的 CD 值进行比较，如果结果高，则它是一个真阳性实例，否则是一个假阳性实例。

3.5 定量结果

为了进一步证明本文提出的方法的有效性，本文进行了实验，将 HDMapFusion 与 nuScenes 基准上 IoU 的表 1 和 AP 的表 2 中的基线进行比较。语义 BEV 地图的元素主要包括静态对象(如车道分界线、人行横道、车道边界)和动态对象(如车辆)。由于本文的方法利用了多模态融合，因此本文比较的基线不仅包含了仅使用相机或仅使用激光雷达的方法，而且还包含了两者的融合。一个有趣的现象是，涉及激光雷达的方法通常比仅使用相机的方法取得更好的性能。与同类方法相比，本文的方法实现了优越的性能和显著的提升。HDMapNet 作为强大的对手，在 HD 地图构建上表现出了很高的精度，但本文的方法在一类 IoU 上的得分为 58.0，在二类 IoU 上的得分为 52.7，在三类 IoU 上的得分为 66.0。平均 IoU 分数比最先进的 SuperFusion 方法提高了近 11.0 分，这意味着本文提出的方法具有强大的环境感知能力。表明本文提出的新型多模态融合机制方法的有效性，而不是直接合并原始传感器数据，本文将不同模态转化到 BEV 空间中进行统一表示，以保持尽可能多的语义和几何信息。

3.6 定性结果

图 4 展示了各种驾驶场景的可视化结果。第一列描绘了来自 6 个摄像头的原始输入全景图像。第二和第三列从预测和地面真实值进行可视化。每个语义 BEV 地图主要包含三种类型的道路元素，并以三种不同的颜色绘制。第二列可视化的预测保持了稳定和令人印象深刻的结果。更重要的是，本文的方法即使在光照条件较暗的情况下也可以生成第四排所示的完整车道线，本文推测，在道路不完全可见的情况下，模型可以根据部分观测来预测车道线的形状。

3.7 消融研究

本文在表 3 所示的 nuScenes 数据集上进行消融实验，以分离每个模块对最终性能的影响。重点解释了所提出模型的两个主要部分的效果，即输入和功能模块，以及不同 backbones 组合的效果。在只有摄像机作为输入的条件下能获取足够的语义信息来生成高清地图，但效果并不理想。原因是语义信息不能够弥补在 2D 图像特征转 3D 向 BEV 下投影过程中缺乏的深度信息。仅激光雷达作为输入在生成地图元素车道边界上比仅相机作为输入表现更好，但在车道分界线和人行横道方面表现更差，这表明单靠一种输入方式受限于平等识别不同的地图元素。本文提出的相机和激光雷达同时作为输入，模型结合了语义和几何信息来识别细粒度的地图特征，这种方式可以在高清地图生成时获得更好的精度。

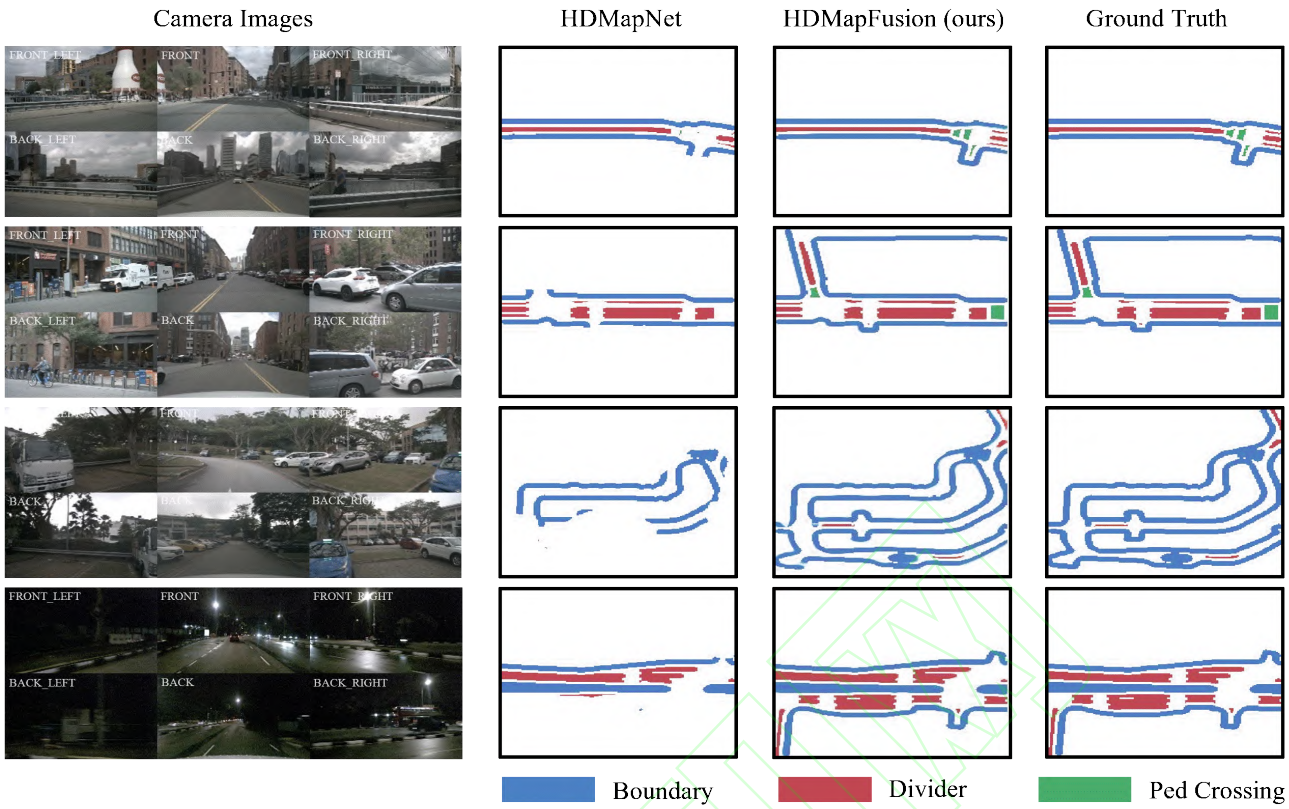

图 4　模型 HDMapFusion 在 nuScenes 数据集上的定性结果

Fig.4　Quantitative results of HDMapFusion on nuScenes benchmark

表 3　消融实验结果

Table 3　The results of ablation studies

		平均交并比（Avg IoU）		
		车道分界线	人行横道	车道边界
输入	仅相机	33.2	25.9	38.0
	仅激光雷达	30.3	19.2	40.6
	C&L Fusion	48.4	42.6	51.7
Function	w/o Depth Supervision	41.2	36.6	49.4
	w/o Fusion Mechanism	46.4	38.9	52.5
	w/o BEV Alignment	48.1	41.2	55.0
Backbones	ResNet18+PointPillars	56.0	50.4	63.2
	ResNet34+PointPillars	58.0	52.7	66.0
	ResNet50+PointPillars	59.1	52.4	65.6

表 4　对比知识蒸馏的不同影响

Table 4　Compare Different Impacts of Knowledge Distillation

方法	参数（M）	推理速度（FPS）	地图生成性能（IoU）			平均
			车道分界线	人行横道	道路边界	
无知识蒸馏	32.2	25	55.2	50.6	63.3	56.4

知识蒸馏	26.8	30	58.0	52.7	66.0	58.9

本文设计了一种类似知识蒸馏中的“教师-学生”模式，利用激光雷达所蕴含的丰富几何信息来监督图 2 中 BEV 转换的过程。在 camera-to-BEV 转换过程中，如果没有 LiDAR 作为老师来监督深度估计，仅靠图像特征来进行深度估计在复杂环境下是不可靠的，从而产生较差的结果，说明受监督的深度估计模块是有必要的。表 4 对比教师模型和学生模型在参数量、推理速度以及地图生成精度（如 IoU）上的表现。通过表中数据对比，可以清晰地展示知识蒸馏在保持较高精度的同时，显著降低了模型的复杂度和计算开销。从表 4 的实验结果可以看出引入知识蒸馏后模型参数从 32.2M 减少到 26.8M，并且推理速度从 25FPS 提升到 30FPS，在高清地图生成的性能上也提升了 4.4%。由于缺乏多模态融合机制，简单的特征拼接操作无法充分利用不同模态之间的互补性导致模型不能充分理解不同模态间的潜在语音信息以高效生成 BEV 语义图。

4 结论

本文提出了一种基于统一 BEV 表示的多模态融合的 HD 地图生成方法用于城市道路的自动驾驶场景。通过本文的研究，发现不同的 HD 地图元素（车道线、人行横道和道路边界线）在不同模态（相机和激光雷达）的特征向量中存在联系。而 BEV 提供了一种物理可解释的方式融合来自多个模态的信息。此外，多模态融合相较于单个传感器输入提供了一种更全面的感知方式来整合关于 3D 场景的全局上下文推理。在 nuScenes 数据集上的实验结果证明了本文方法的有效性，并表明 HDMapFusion 对高清地图生成精度的优于所对比的基准模型。

本文工作主要基于开源的大规模自动驾驶数据集 nuScenes 开展的，并且这些数据集对自动驾驶场景涵盖范围有限，集中体现密集、复杂的多种代理交互的城市道路环境。未来，本文将继续探究在不同地理和交通场景下（如高速公路、乡村道路）的高清地图生成。

参考文献

[1] Liao, B., Chen, S., Wang, X., Cheng, T., Zhang, Q., Liu, W., Huang, C.: Maptr: Structured modeling and learning for online vectorized hd map construction[C]//Proceedings of International Conference on Learning Representations (ICLR), 2023.

[2] 赵南南，高翡晨. 基于改进 YOLOv8 的交通场景实例分割算法[J]. 计算机工程, 2025, 51(1): 198-207.
ZHAO Nannan, GAO Feichen. Improved YOLOv8-based Algorithm for Instance Segmentation in Traffic Scenes[J]. Computer Engineering, 2025, 51(1): 198-207.

[3] 秦严严. 交通流分析理论[M]. 人民交通出版社, 2023.
Qin Yanyan. Theory of Traffic Flow Analysis[M]. China Communications Press, 2023.

[4] Li, Q., Wang, Y., Wang, Y., Zhao, H.: Hdmapnet: An online hd map construction and evaluation framework[C]//Proceedings of International Conference on Robotics and Automation (ICRA), 2021: 4628-4634.

[5] 阳钧，鲍泓，梁军，马楠. 一种基于高精度地图的路径跟踪方法[J]. 计算机工程, 2018, 44(7): 8-13.
YANG Jun, BAO Hong, LIANG Jun, MA Nan. A Path Tracking Method Based on High Precision Map[J]. Computer Engineering, 2018, 44(7): 8-13.

[6] 刘宏纬，邵东恒，杨剑，魏宪，李科，游雄. 基于鸟瞰图融合的多级旋转等变目标检测网络[J]. 计算机工程, 2024, 50(11): 246-257.
LIU Hongwei, SHAO Dongheng, YANG Jian, WEI Xian, LI Ke, YOU Xiong. Multi-Level Rotational Equivariant Object Detection Network Based on BEV Fusion[J]. Computer Engineering, 2024, 50(11): 246-257.

[7] Liu Z, Tang H, Amini A, et al. Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation[C]//Proceedings of IEEE international conference on robotics and automation (ICRA). IEEE, 2023: 2774-2781.

[8] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. Advances in neural information processing systems[J]. Advances in neural information processing systems, 2017, 30.

[9] Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E.,

Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O.: nuscenes: A multimodal dataset for autonomous driving[C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019: 11618-11628.

[10] Dong H, Gu W, Zhang X, et al. Superfusion: Multilevel lidar-camera fusion for long-range hd map generation[C]//Proceedings of IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2024: 9056-9062

[11] Sun, L., Yang, K., Hu, X., Hu, W., Wang, K.: Real-time fusion network for rgb-d semantic segmentation incorporating unexpected obstacle detection for road driving images. IEEE Robotics and Automation Letters 5, 2020: 5558-5565.

[12] Hu S, Chen L, Wu P, et al. St-p3: End-to-end vision-based autonomous driving via spatial-temporal feature learning[C]//Proceedings of European Conference on Computer Vision. 2022: 533-549

[13] Rhinehart, N., McAllister, R., Kitani, K., Levine, S.: Precog: Prediction conditioned on goals in visual multi-agent settings[C]//Proceedings of IEEE/CVF International Conference on Computer Vision (ICCV), 2019: 2821-2830.

[14] Xu, H., Yang, C., Li, Z.: Od-slam: Real-time localization and mapping in dynamic environment through multi-sensor fusion[C]//Proceedings of International Conference on Advanced Robotics and Mechatronics (ICARM), 2020: 172-177.

[15] Chen, J., Li, X., Xie, J., Li, J., Qian, J., Yang, J.: Cbi-gnn: Crossscale bilateral graph neural network for 3d object detection[J]. IEEE Transactions on Intelligent Transportation Systems, 2022, 23(12), 23124-23135.

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

[17] Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database[C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2009: 248-255.

[18] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition[C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2016: 770-778

[19] Lang A H, Vora S, Caesar H, et al. Pointpillars: Fast encoders for object detection from point clouds[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019: 12697-12705.

[20] Qi, C., Su, H., Mo, K., Guibas, L.J.: Pointnet: Deep learning on point sets for 3d classification and segmentation[C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2016: 77-85.

[21] Man, Y., Gui, L., Wang, Y.X.: CroMA: Cross-modality adaptation for monocular BEV perception[C]//Proceedings of International Conference on Learning Representations (ICLR) (2023)

[22] Radford, A., Narasimhan, K.: Improving language understanding by generative pre-training[C]//Proceedings of International Conference on Learning Representations (ICLR), 2018.

[23] Deng, L., Yang, M., Li, H., Li, T., Hu, B., Wang, C.: Restricted deformable convolution-based road scene semantic segmentation using surround view cameras[J]. IEEE Transactions on Intelligent Transportation Systems 21,2018: 4350-4362.

[24] Pan, B., Sun, J., Leung, H.Y.T., Andonian, A., Zhou, B.: Cross-view semantic segmentation for sensing surroundings[J]. IEEE Robotics and Automation Letters 5, 2019: 4867-4873.

[25] 吴永庆, 姜正宇. 基于解耦合动态时空卷积循环网络的交通流预测[J]. 计算机工程, doi: 10.19678/j.issn.1000-3428.0070319.
WU Yongqing, JIANG Zhengyu. Traffic Flow Prediction Based on Decoupled Dynamic Spatial-Temporal Convolutional Recurrent Network[J]. Computer Engineering, doi: 10.19678/j.issn.1000-3428.0070319.

[26] Loshchilov, I., Hutter, F.: Decoupled weight decay regularization[C]//Proceedings of International Conference on Learning Representations (ICLR), 2017.

[27] Lin T Y, Maire M, Belongie S, et al. Microsoft coco: Common objects in context[C]//Proceedings of European Conference on Computer Vision. 2014: 740-755.