计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
[doi: 10.15888/j.cnki.csa.009758] [CSTR: 32024.14.csa.009758]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

面向自动驾驶的高效视图转换①

刘家辉[1,3,4], 官敬超[2,3,4], 方鸿清[1,3,4], 巢建树[3,4]

[1](福建师范大学 计算机与网络空间安全学院, 福州 350117)
[2](福州大学 先进制造学院, 泉州 362251)
[3](中国科学院 福建物质结构研究所 泉州装备制造研究中心, 泉州 362216)
[4](中国科学院大学 福建学院, 福州 350002)
通信作者: 巢建树, E-mail: jchao@fjirsm.ac.cn

摘　要: 在自动驾驶技术的领域中, 利用鸟瞰图 (bird's eye view, BEV) 进行 3D 目标检测任务已经引起了广泛的关注. 针对现有相机至鸟瞰视图转换方法, 实时性不足、部署复杂度较高的难题, 提出了一种简单高效、无需任何特殊工程操作即可部署的视图转换方法. 首先, 针对完整图像特征存在大量冗余信息, 引入宽度特征提取器并辅以单目 3D 检测任务, 提炼图像的关键特征, 确保过程中信息损失的最小化; 其次, 提出一种特征引导的极坐标位置编码方法, 增强相机视角与鸟瞰图表示之间的映射关系与模型空间理解能力; 最后, 通过单层交叉注意力机制实现可学习 BEV 嵌入与宽度图像特征的交互, 从而生成高质量的 BEV 特征. 实验结果表明: 在 nuScenes 验证集上该网络架构与 LSS (lift, splat, shoot) 相比 *mAP* 从 29.5% 提升到 32.0%, 提升了 8.5%, *NDS* 从 37.1% 提升到 38.0%, 提升了 2.4%, 表明该模型在自动驾驶场景下的 3D 目标检测任务的有效性. 同时相比于 LSS 在延迟上降低了 41.12 %.

关键词: 自动驾驶; 鸟瞰图; 视图转换; 目标检测; 交叉注意力

引用格式: 刘家辉,官敬超,方鸿清,巢建树.面向自动驾驶的高效视图转换.计算机系统应用. http://www.c-s-a.org.cn/1003-3254/9758.html

Efficient View Transformation for Autonomous Driving

LIU Jia-Hui[1,3,4], GUAN Jing-Chao[2,3,4], FANG Hong-Qing[1,3,4], CHAO Jian-Shu[3,4]

[1](College of Computer and Cyber Security, Fujian Normal University, Fuzhou 350117, China)
[2](School of Advanced Manufacturing, Fuzhou University, Quanzhou 362251, China)
[3](Quanzhou Institute of Equipment Manufacturing, Fujian Institute of Research on the Structure of Matter, Chinese Academy of Sciences, Quanzhou 362216, China)
[4](Fujian College, University of Chinese Academy of Sciences, Fuzhou 350002, China)

Abstract: In autonomous driving, the task of using bird's eye view (BEV) for 3D object detection has attracted significant attention. Existing camera-to-BEV transformation methods are facing challenges of insufficient real-time performance and high deployment complexity. To address these issues, this study proposes a simple and efficient view transformation method that can be deployed without any special engineering operations. First, to address the redundancy in complete image features, a width feature extractor is introduced and supplemented by a monocular 3D detection task to refine the key features of the image. In this way, the minimal information loss in the process can be ensured. Second, a feature-guided polar coordinate positional encoding method is proposed to enhance the mapping relationship between the camera view and the BEV representation, as well as the spatial understanding of the model. Lastly, the study has achieved the interaction between learnable BEV embeddings and width image features through a single-layer cross-attention mechanism, thus generating high-quality BEV features. Experimental results show that, compared to lift, splat, shoot (LSS), on the nuScenes validation set, this network structure improves mAP from 29.5% to 32.0%, an increase of 8.5%,

① 基金项目: 福厦泉国家自主创新示范区协同创新平台项目 (2023FX0002)
收稿时间: 2024-07-12; 修改时间: 2024-08-01; 采用时间: 2024-08-20; csa 在线出版时间: 2024-11-15

and NDS from 37.1% to 38.0%, an increase of 2.4%. This demonstrates the effectiveness of the model in 3D object detection tasks in autonomous driving scenarios. Additionally, compared to LSS, it reduces latency by 41.12%.

Key words: autonomous driving; bird's eye view (BEV); view transformation; object detection; cross-attention

在自动驾驶领域, 3D 物体检测是一项至关重要的感知技术[1]. 相较于依赖昂贵的激光雷达[2]或多模态方法[3,4], 利用多摄像头获取的 2D 图像进行 3D 物体检测不仅可以与激光雷达系统协同工作, 提供丰富的视觉信息, 增强系统的整体感知能力, 还可以独立部署以保持低成本. 与直接从图像特征中检测物体相比, 从统一的鸟瞰图表示中识别 3D 物体更直观地符合人类感知. 鸟瞰图视角的优势在于其能够在一个统一的空间中整合不同视角、传感器数据乃至时间序列信息, 为特征融合提供了便利[5]. 这种统一的表示方法不仅促进了信息整合, 也能够应用于各种下游任务[6,7](例如物体检测、地图分割、运动规划等).

视图转换是将多视图特征转换为鸟瞰图的关键技术, 在先前工作中得到广泛研究. 这些研究主要分为基于 Lift-Splat[8−10]和基于 Transformer[11]的两种方法. 基于 Lift-Splat 方法通过预测每个像素的分类深度分布, 将特征"Lift"至 3D 空间, 并在垂直方向上将投影图像特征"Splat"到预定义网格上, 构建鸟瞰图特征. 尽管类 Lift-Splat 方法非常有效, 但"Splat"的实现要么依赖于效率极低的累加技巧[9], 要么需要特定设备的定制运算符[12], 增加了鸟瞰图感知方法的应用成本. 另一方面, 基于 Transformer 的方法通过注意力机制[13]查询图像特征, 导出 BEV 表示. 但这些方法使用多层 Transformer 解码器计算 BEV 表示, 严重影响了处理速度, 限制了它们在实时应用中的部署. 此外, 可变形注意力操作[14]也需要 CUDA 支持. 因此, 这使得这些方法难以应用于现实世界复杂的驾驶环境.

考虑到这些因素, 本文提出了一种简单高效的视图转换方法, 且无需复杂的工程工作即可部署. 通过在 nuScenes[15]验证集上对本方法进行评估, 并与现有视图转换方法进行对比. 结果表明, 本视图转换方法在保持高效实时检测速度的同时, 具有更高的检测精度.

1 相关工作

在自动驾驶技术领域, 基于视觉的三维目标检测技术是众多下游应用的核心. 单目 3D 检测技术, 如 Wang 等人[16]提出的 FCOS3D, 通过从单个图像中回归 3D 边界框, 将 2D 检测技术[17]拓展到三维空间. 而多视图 3D 检测技术则进一步融合多个视角的图像, 以实现更精确的几何推断. 受 2D 检测工作的启发, 一些方法提出了不同的三维物体检测框架, 以直接实现稀疏物体级特征提取. Wang 等人[14]提出的 DETR3D 率先使用多视角图像输入范式来增强 3D 检测的性能. Liu 等人[18]提出的 PETR 通过集成 3D 位置编码, 扩展了 DETR[14]这一稀疏检测器的能力. Liu 等人[19]提出的 PETRv2 进一步通过时间建模的融合, 对 PETR 进行了改进. Chen 等人[20]提出的 PolarDETR 提出了用于 3D 检测的极坐标参数化方法, 在极坐标系中重新制定了边界框参数化、网络预测和损失计算. Xiong 等人[21]提出的 CAPE 则在局部摄像机坐标系中创建位置编码, 优化了 PETR 的编码策略. 在自动驾驶的新趋势中, 鸟瞰图空间因其在感知、预测、多任务学习和规划等多个方面的优势而受到广泛关注. 因此, 一些前沿的方法开始利用 BEV 空间上的表示来进行 3D 物体检测.

1.1 基于 Lift-Splat 方法

在 Lift-Splat 方法中, BEV 特征是通过卷积、柱状池化 (pillar pooling) 或体素池化 (voxel pooling) 压缩投影点云特征, 并结合预测的深度信息进行加权计算得到的. 这一方法已成为后续研究的基石. Huang 等人[9]提出的 BEVDet 将此技术应用于多视图 3D 检测领域. Reading 等人[22]提出的 CaDDN 引入 LiDAR 点云以生成精确的深度真值, 以监督深度预测模块. Li 等人[12]提出的 BEVDepth 验证了深度预测的准确性对模型性能的显著提升. Liu 等人[23]提出的 BEVFusion 通过多线程技术加速池化过程来加快推理速度. Xie 等人[24]提出的 M2BEV 通过均匀深度假设有效降低了内存占用. Zhou 等人[25]提出 MatrixVT 通过垂直维度的特征压缩和极坐标变换, 进一步提升了 BEV 特征计算的效率.

1.2 基于 Transformer 方法

基于 Transformer 的视图转换方法利用注意机制直接输出 BEV 表示. Yang 等人[26]提出的 PYVA 通过

交叉注意力机制学习 BEV 特征, 并通过周期一致性对模型进行正则化以增强性能. Zhou 等人[27]提出的 CVT 网络采用 BEV 查询和交叉注意力机制, 结合添加由相机参数计算得到位置嵌入的图像特征, 以实现视图转换. Li 等人[11]提出的 BEVFormer 引入可变形注意力机制, 专注于 BEV 重投影透视图中的关键区域, 有效降低计算成本, 加速模型训练. Chen 等人[28]的 GKT 方法利用几何先验引导注意力, 聚焦于 2D 参考点的核区域, 以生成 BEV 表征. 金祖亮等人[29]提出通过构建 BEV 查询和图像特征之间的局部窗口交叉注意力, 完成对跨相机透视图之间的特征查询. 本文提出的算法, 是在 Transformer 结构的基础上构建的.

2 算法设计与实现

本文所提模型的网络结构如图 1 所示. 该模型以图像特征作为输入, 并输出用于下游感知任务的鸟瞰图特征. 整体流程如下: 首先, 利用一个宽度特征提取器 (width feature extractor, WFE) 来从原始图像特征中压缩并提取宽度特征. 这一步骤有助于减少输入数据的复杂度, 使得后续处理更高效. 本文使用交叉注意力机制获取鸟瞰图特征. 具体来说, 将可学习的 BEV 嵌入作为查询 (Query), 而宽度特征则用作键 (Key) 和值 (Value). 在这一过程中, BEV 嵌入仅通过单个交叉注意力层与宽度特征进行交互, 从而提取出高质量的 BEV 特征. 这种基于 Transformer 的计算方式有助于捕捉图像特征在鸟瞰图空间中的关系和表征. 为有效编码图像特征与 BEV 特征之间的空间关系, 本文设计了特征引导的极坐标位置编码模块 (feature guided polar position encoder, FPPE). 这一模块通过极坐标系统将特征的位置信息嵌入到特征表示中, 从而增强模型对空间结构的感知能力. 在训练阶段, 本文采用单目 3D 目标检测作为辅助任务来补充缺失的信息.

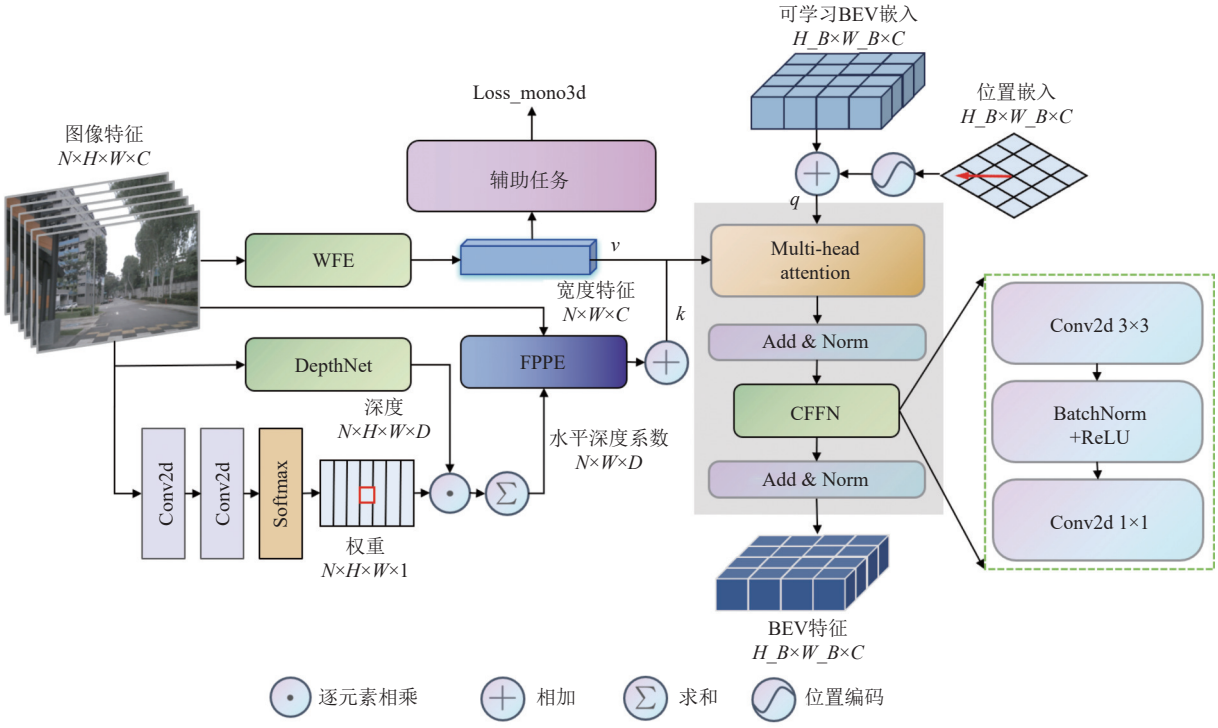


图 1 视图转换算法网络结构

2.1 特征引导的极坐标位置编码

本文设计了一种特征引导的极坐标位置编码方式它通过隐式引入视觉先验, 将图像特征融入 3D 位置嵌入的生成过程中, 网络结构如图 2 所示. 首先对相机视锥空间进行离散化处理, 以创建一个维度为(H,W,D)的网格结构. 在这个网格中, 每个点可以被表示为$\{p_{i,j,k}=(u_{i,j}\times d_k,\ v_{i,j}\times d_k,\ d_k)^{\mathrm{T}}|i\in|H|,\ j\in|W|,\ k\in|D|\}$, 其中, $(u_{i,j},v_{i,j})$表示图像上的像素坐标, 而d_k表示与图像平面垂直的轴的深度值, H 表示图像特征高度, W 表示图像特征宽度, D 表示深度预测值. 通过相机的内参和外参, 将这些视锥坐标转换到 3D 的激光雷达坐标系中. 然后根据预定义网格范围, 将 3D 坐标转换至体素坐标系

中. 计算公式为:

$$C_{i,j,k} = K^n \cdot I^{-1} \cdot P_{i,j,k} + T^n \tag{1}$$

$$bound : [min, max, interval] \tag{2}$$

$$coor = (C_{i,j,k} - min)/interval \tag{3}$$

其中, $C_{i,j,k} = [x_{i,j,k}, y_{i,j,k}, z_{i,j,k}]^{\mathrm{T}}$ 为 3D 坐标, $K \in \mathbb{R}^{3\times3}$ 和 $T \in \mathbb{R}^{3\times1}$ 分别是将第 n 个视图的坐标转换为统一的激光雷达坐标系的旋转矩阵与平移向量, $I \in \mathbb{R}^{3\times3}$ 为相机内参 (由相机焦距、光学中心等参数组成的矩阵, 将坐标由相机坐标系转化为像素坐标系), $\cdot$ 为矩阵乘法, *bound* 为 X、Y 轴范围, *interval* 为每个网格的间隔, *coor* 为体素坐标.

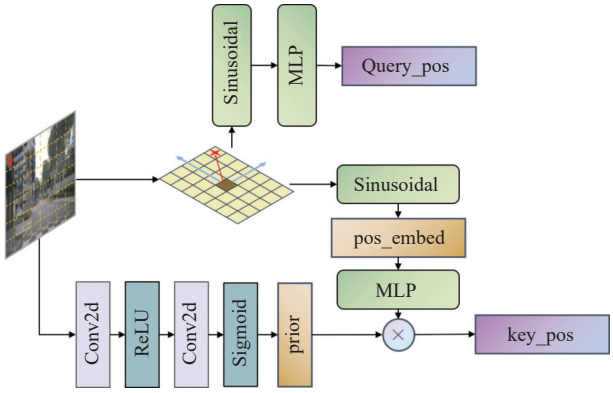


图 2 特征引导的极坐标位置编码

为了简化计算, 本文只考虑体素坐标的 X、Y 轴, 而 Z 轴上的特征则将自动聚合, 同时, 沿着 H 方向上对 3D 位置坐标进行平均处理以适应宽度特征. 本文采用极坐标系进行位置编码的计算, 并将自车设置为极坐标系原点. 根据图像特征对应的体素坐标计算其在极坐标系下与自车的径向距离与正余弦值, 并采用正余弦位置编码器将这些空间坐标信息编码, 计算过程如式 (4)、式 (5) 所示:

$$d_{j,k} = \sqrt{x_{j,k}^2 + y_{j,k}^2},\ \cos\theta_{j,k} = \frac{x_{j,k}}{d_{j,k}},\ \sin\theta_{j,k} = \frac{y_{j,k}}{d_{j,k}} \tag{4}$$

$$\varphi_{j,k} = Concat(\psi(d_{j,k}), \psi(\cos\theta_{j,k}), \psi(\sin\theta_{j,k})) \tag{5}$$

其中, $\varphi_{j,k} = \mathbb{R}^{N\times W\times D\times3C}$, ψ 为正余弦位置编码, $j \in |W|$, $k \in |D|$, *Concat* 为沿通道维度聚合. 随后, 利用水平深度系数 (coefficient) 对位置编码进行加权, 并通过多层感知机 (MLP) 对加权后的位置编码进行聚合. 最后, 引入一个多层感知机获取 2D 图像特征的注意权重, 并与位置编码相乘以获取特征引导的极坐标位置编码. 计算过程如式 (6)–式 (11) 所示:

$$d_{\mathrm{pre}} = Conv(fea) \tag{6}$$

$$h_{\mathrm{pre}} = Conv_{3\times3}(ReLU(Norm(Conv_{3\times3}(fea)))) \tag{7}$$

$$coef = \mathrm{sum}(Softmax(h_{\mathrm{pre}}) \times Softmax(d_{\mathrm{pre}}), dim_h) \tag{8}$$

$$\varphi_j = \mathrm{MLP}\left(\sum_{k=1}^{D}(\mathrm{coef} \times \varphi_{j,k})\right) \tag{9}$$

$$F_j = Sigmoid(Conv(\mathrm{ReLU}(Conv(fea)))) \tag{10}$$

$$pos = \varphi_j \times F_j \tag{11}$$

其中, *fea* 为图像特征, $Conv_{3\times3}$ 为 3×3 卷积, *Norm* 为 BatchNorm, d_{pre} 为预测的深度信息, h_{pre} 为预测的高度信息, sum 为沿 dim_h 维度求和, $\varphi_j \in \mathbb{R}^{N\times W\times C}$, $coef \in \mathbb{R}^{N\times W\times D}$ 为水平深度系数, $fea \in \mathbb{R}^{N\times H\times W\times C}$ 为 2D 图像特征, *Conv* 表示 1×1 卷积网络, ×表示逐元素相乘.

2.2 宽度特征提取

直接使用多视图特征与 BEV 查询进行交互会带来巨大的计算开销, 导致效率低下. 而图像数据中充斥着大量的背景信息, 因此, 对图像特征进行高度方向上的压缩是一种合理的选择. 模块结构如图 3 所示, 计算过程如下.

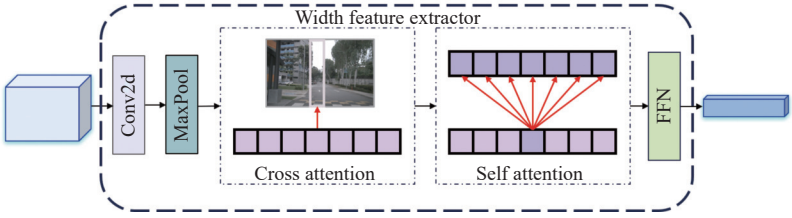


图 3 宽度特征提取模块结构

1) 使用最大池化层 (MaxPool) 对图像特征进行垂直压缩, 以提取主要宽度特征并保留有效信息. 如式 (12) 所示:

$$F_{\mathrm{width}} = \mathrm{MaxPool}(Conv(fea)) \tag{12}$$

其中, $F_{\text{width}} \in \mathbb{R}^{N\times W\times C}$ 为宽度特征, $fea \in \mathbb{R}^{N\times H\times W\times C}$ 为原始图像特征, *Conv* 为 1×1 卷积网络.

2) 引入交叉注意力机制 (cross attention) 来聚合对应图像列上的信息, 丰富宽度特征的信息量.

$$F'_{\text{width}} = Cross\_Attention(F_{\text{width}(j)}, fea_j, fea_j) \tag{13}$$

其中, j 表示第 j 列特征, $F_{\text{width}(j)}$ 为查询, fea_j 为键和值. 交叉注意力的计算复杂度为 O($N\times W\times H\times C$), N 为视图数量.

3) 通过自注意力 (self attention) 操作对宽度特征进行增强.

$$F''_{\text{width}} = Self\_Attention(F'_{\text{width}}, F'_{\text{width}} + \varphi_j, F'_{\text{width}}) \tag{14}$$

其中, φ_j 为极坐标位置编码, F'_{width} 作为查询和值的输入, $F'_{\text{width}} + \varphi_j$ 作为键的输入. 自注意力的计算复杂度为 O($N\times W\times W\times C$).

2.3 BEV 特征提取

在自动驾驶感知中, BEV 特征提取是一个至关重要的环节. BEV 特征能够提供自车周围环境的全局视角, 有助于准确感知和理解车辆周围的空间布局和动态变化. 本节将介绍如何通过可学习 BEV 嵌入及其位置嵌入, 结合交叉注意力机制来计算和提取 BEV 特征.

2.3.1 可学习 BEV 嵌入及其位置嵌入

本文预先定义一组具有栅格形状的可学习参数 $Q_B \in \mathbb{R}^{H_B\times W_B\times C}$ 作为 BEV 查询, 用来捕获围绕自车的 BEV 特征. BEV 平面上的每个网格单元都对应于现实世界中大小为 $s\times s$ m^2 的区域. BEV 查询的高度 H_B 和宽度 W_B 与 BEV 平面在 X 轴和 Y 轴的栅格尺寸保持一致. 为确保查询位置嵌入与宽度特征的位置嵌入保持一致性, 本文同样采用了极坐标系统来计算查询的位置编码.

对于每个体素网格的中心坐标 $[x, y]^{\text{T}}$, 位置编码计算公式为:

$$d = \sqrt{x^2 + y^2},\ \cos\theta = \frac{x}{d},\ \sin\theta = \frac{y}{d} \tag{15}$$

$$\varphi = Concat(\psi(d), \psi(\cos\theta), \psi(\sin\theta)) \tag{16}$$

$$Q_{\text{pos}} = \text{MLP}(\varphi) \tag{17}$$

其中, $\varphi \in \mathbb{R}^{H_B\times W_B\times 3C}$, $Q_{\text{pos}} \in \mathbb{R}^{H_B\times W_B\times C}$, ψ 为正余弦位置编码, *Concat* 为沿通道维度聚合, MLP 为多层感知机.

2.3.2 BEV 特征计算

本文引入了交叉注意力机制来计算 BEV 特征. 在此框架下, 定义了可学习的 BEV 嵌入 $Q_B \in \mathbb{R}^{H_B\times W_B\times C}$ 及其位置嵌入 $Q_{\text{pos}} \in \mathbb{R}^{H_B\times W_B\times C}$ 作为查询输入, 宽度特征 $F''_{\text{width}} \in \mathbb{R}^{N\times W\times C}$ 及其位置嵌入 $pos \in \mathbb{R}^{N\times W\times C}$ 作为键输入, 宽度特征本身则作为值输入. 为了捕捉局部空间特征并增强模型的表达能力, 本文采用了卷积前馈网络 (convolutional feed-forward network, CFFN) 进行特征提取. BEV 特征计算过程如下所示:

$$Q'_B = Q_B + MHA(Q_B + Q_{\text{pos}}, F''_{\text{width}} + pos, F''_{\text{width}}) \tag{18}$$

$$CFFN \to Conv2d_{1\times1}(\sigma(Norm(Conv2d_{3\times3}(x)))) \tag{19}$$

$$BEV\_fea = Q'_B + CFFN(Q'_B) \tag{20}$$

其中, *MHA* 为多头注意力 (multi-head attention), *CFFN* 由 3×3 卷积函数 ($Conv2d_{3\times3}$)、批量归一化 (*Norm*, BatchNorm)、激活函数 (σ, *ReLU*) 和 1×1 卷积函数 ($Conv2d_{1\times1}$) 组成.

相较于采用全局图像特征作为交叉注意力机制的查询 (计算量如式 (21) 所示), 本文利用宽度特征的方法显著降低了计算量 (如式 (22) 所示), 实现了 H 倍的效率提升.

$$F = N\times H\times W\times H_{\text{BEV}}\times W_{\text{BEV}}\times C \tag{21}$$

$$F = N\times W\times H_{\text{BEV}}\times W_{\text{BEV}}\times C \tag{22}$$

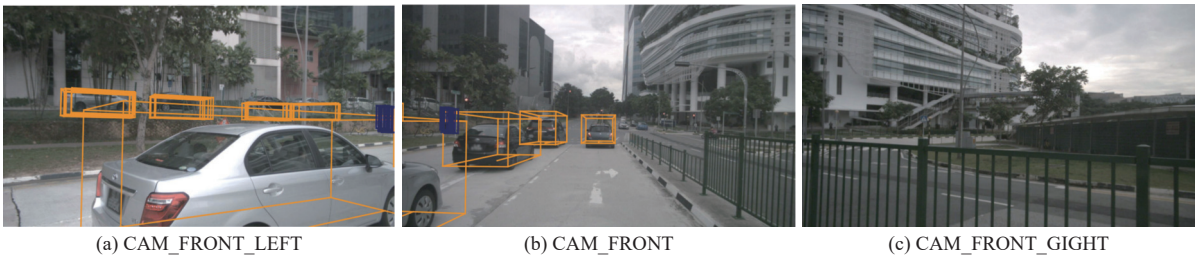
其中, F 为计算量, N 为视图数量, H 和 W 分别为图像特征的高度与宽度, H_{BEV} 和 W_{BEV} 分别为 BEV 查询的高度与宽度, C 为通道数.

3 实验

3.1 数据集

本文在公开的自动驾驶数据集 nuScenes 上对所提网络进行了测试. 该数据集涵盖了波士顿和新加坡的多种复杂城市驾驶场景, 共 1 000 个驾驶场景, 其中包含 700 个训练场景、150 个验证场景和 150 个测试场景. 每个场景约 20 s, 覆盖不同时间和天气条件下的城市环境. 每个场景由 6 个视角的图像构成, 覆盖整个周围环境. 数据集集成了 6 个摄像头、1 个前向激光雷达、5 个毫米波雷达、12 个超声波传感器以及 GPS 和 IMU 数据, 提供 360 度全方位的环境感知能力. 注释信息包括动态物体 (如车辆、行人、骑行者) 和静态物体 (如路标、交通信号灯、道路障碍) 的 3D 边界框标注、运动轨迹、道路布局和行驶路径等. 图 4 展示

了数据集实例与 3D 目标检测结果在输入数据集上的可视化. 模型以自车为中心的 6 个方向 (分别为左前、前、右前、左后、后、右后) 上的相机图像为输入, 预测 3D 目标 (车辆、行人等) 的目标值.



(a) CAM_FRONT_LEFT (b) CAM_FRONT (c) CAM_FRONT_GIGHT

(d) CAM_BACK_LEFT (e) CAM_BACK (f) CAM_BACK_GIGHT

图 4 数据集实例与 3D 检测结果可视化

3.2 评价指标

对于 3D 目标检测任务, 本文遵循 nuScenes 官方的评价标准. 除了常用的平均精度均值 (mean average precision, *mAP*), 评估指标还包括 nuScenes 真实阳性 (TP) 误差, 其中包含 5 个 TP 指标: 平均平移误差均值 (mean average translation error, mATE)、平均尺度误差均值 (mean average scale error, mASE)、平均方向误差均值 (mean average orientation error, mAOE)、平均速度误差均值 (mean average velocity error, mAVE) 和平均属性误差均值 (mean average attribute error, mAAE). 此外, nuScenes 数据集还提出了 nuScenes 检测分数 (nuScenes detection score, *NDS*), 用于评估网络的综合性能, 其计算公式为:

$$NDS = \frac{1}{10}\left[5mAP + \sum_{\{mTP\}} \max(1 - mTP, 0)\right] \tag{23}$$

3.3 实验设置

本文提出的网络架构在 nuScenes 训练集上进行模型训练, 并在 nuScenes 验证集上完成性能评估. 本文采用了 BEVDet 和 BEVDet4D 作为执行 3D 目标检测任务的基础框架. 以在 ImageNet 上预训练的 ResNet50 为主干网络, 特征金字塔网络 (FPN) 作为特征融合的颈部, 以及 CenterHead 作为检测头部. 输入图像的原始分辨率为 1600×900, 经过适当缩放至 704×256 作为模型输入. 对于 BEV 网格, 其分辨率设置为 128×128, 对应于 X 轴和 Y 轴方向上[−51.2 m, 51.2 m]的范围, 间隔为 0.8 米, 通道维度设定为 64.

本文利用 CBGS[30]进行类平衡采样增强, 对模型进行 24 epoch 的训练. 采用 AdamW 优化器, 初始学习率设置为 0.000 2, 权重衰减系数为 0.01, 每批次的样本数量 (batch size) 设置为 8. 此外, 采用分步学习率策略, 将第 16 和第 22 个 epoch 的学习率降低 0.1 倍. 在正式训练开始前, 进行 500 步的小学习率预热训练, 在预热阶段, 学习率从初始学习率的 0.1 倍开始, 线性增加至初始学习率. 实验过程使用的环境配置如表 1 所示.

表 1 实验配置

类型	型号
系统	Ubuntu 20.04.6
CPU	Intel(R) Xeon(R) Gold 6326
GPU	NVIDIA GeForce RTX 3090 24 GB
计算平台	CUDA 11.1
编程语言	Python 3.7.16
深度学习框架	Torch 1.9.0+cu 111

3.4 消融实验

本文在 WidthFormer[31]模型上分别添加可学习 BEV 嵌入与特征引导的极坐标位置编码以验证所提出模块的有效性, 如表 2 所示. 其中“√”表示在基准网络中加入该模块, “Fixed”表示固定 BEV 查询, “Learned”表示可学习 BEV 查询, “FPPE”表示特征引导的极坐标位置编码.

表 2 模块消融实验结果 (%)

模块			*mAP*	*NDS*	mATE	mASE	mAOE	mAVE	mAAE
Fixed	Learned	FPPE	↑	↑	↓	↓	↓	↓	↓
√	—	—	31.39	37.56	71.39	27.89	60.89	94.85	**26.34**
—	√	—	31.38	**37.79**	**72.38**	**27.77**	60.21	**90.77**	27.83
—	√	√	**31.71**	37.74	71.40	27.91	**59.50**	94.13	28.24

由表 2 可知, 当引入可学习 BEV 嵌入时, 在 *NDS* 指标上取得提升, 从 37.56% 提升至 37.79%, 提升了 0.6%. 这表明引入可学习 BEV 嵌入在一定程度上提升了模型的整体预测能力. 当引入 FPPE 后, *mAP* 从 31.38% 提升至 31.71%, 提升了 1.1%. 这表明, FPPE 对目标检测精度有显著的、积极的影响, 因为 FPPE 增强了模型在特征提取和目标区分上的能力, 使得检测更加准确. 总体而言, 结合可学习 BEV 嵌入和 FPPE 可以提高模型的 *mAP*, 且对 *NDS* 也有一定提升, 说明这两种改进策略在提升目标检测精度和整体性能方面具有一定的互补性和有效性.

3.5 对比实验

本文在 BEVDet 与 BEVDet4D 框架下将所提出的模型与主流的不同视图转换模型进行比较, 结果如表 3、表 4 所示. 为了确保比较的公正性, 仅采用这些视图转换方法来生成 BEV 特征, 同时保持了基准检测器的其他配置不变.

表 3 BEVDet 框架下 nuScenes 验证集上对比实验 (%)

视图转换方法	*mAP*↑	*NDS*↑	mATE↓	mASE↓	mAOE↓	mAVE↓	mAAE↓
IPM	25.3	34.5	78.5	27.6	62.5	85.9	26.6
LSS	29.5	37.1	73.9	**27.3**	61.2	88.1	**24.8**
MatrixVT	28.9	36.5	74.6	28.3	**60.0**	89.5	27.3
FastBEV	28.9	37.1	73.3	28.1	62.6	**82.6**	27.1
BEVFormer	29.1	34.1	76.1	28.3	71.8	97.2	30.0
WidthFormer	31.4	37.6	71.4	27.9	60.9	94.9	26.3
本文	**32.0**	**38.0**	**69.9**	28.1	64.6	89.2	28.4

表 4 BEVDet4D 框架下 nuScenes 验证集上对比实验 (%)

视图转换方法	*mAP*↑	*NDS*↑	mATE↓	mASE↓	mAOE↓	mAVE↓	mAAE↓
IPM	27.1	41.0	77.8	28.6	57.9	39.7	21.5
LSS	32.8	45.7	71.0	27.9	**51.2**	36.0	20.5
MatrixVT	32.4	45.8	69.6	**27.6**	51.9	36.3	**18.9**
FastBEV	30.8	42.4	73.7	28.1	53.7	51.6	22.4
BEVFormer	31.1	41.1	74.9	28.2	63.7	52.6	24.0
WidthFormer	34.6	46.4	71.0	28.1	54.1	35.0	21.1
本文	**34.9**	**47.2**	**68.5**	28.1	51.4	**34.6**	20.3

根据表 3 的实验结果, 本文提出的模型在关键性能指标 *mAP*、*NDS* 和 mATE 上的表现均优于其他视图转换方法. 具体来说, 该模型在 BEVDet 框架下 *mAP* 达到了 32.0%, *NDS* 达到了 38.0%, 均在所有参与比较的模型中排名第一. 相比之下, 最接近的是 WidthFormer 模型, 其 *mAP* 为 31.4%, 而其他模型均在 29.5% 以下. 这一比较结果不仅凸显了本文模型在目标检测精度上的优势, 也证明了其在综合性能上的出色表现.

本文使用 BEVDet4D 框架对视图转换模型进行了时间维度上的扩展. 根据表 4 的实验结果, 本文提出的视图转换模型依旧优于其他视图转换方法. 总体来看, 本文模型在目标检测和综合性能的主要指标 *mAP* 和 *NDS* 上均优于其他视图转换方法, 验证了本文所提出模型的有效性, 也表明了其在自动驾驶场景下的强大检测能力, 为自动驾驶技术的进一步发展提供了有力的技术支持.

为了进一步评估模型的实际应用性能, 本文在 3090 GPU 上测试了 LSS、WidthFormer 以及改进模型的延迟, 测试结果如表 4 所示.

根据表 5 的结果, 本文所提模型相对于 LSS 模型, 在延迟方面显著提升, 由 5.01 ms 优化至 2.95 ms, 优化了 41.12%. 相对于 WidthFormer 模型, 仅增加 0.1 ms 的延迟就获得更好的性能.

表 5 延迟对比

视图转换方法	延迟 (ms)
LSS	5.01
WidthFormer	2.85
本文	2.95

4 结论与展望

本文提出了一个视图转换模型, 旨在将多摄像头输入转化为具有丰富信息的鸟瞰图特征. 该模型的网络架构简洁而高效, 易于实现快速部署, 且避免了复杂的工程化操作, 确保了模型的实用性和灵活性. 宽度特征提取器有效提取主要特征, 同时排除了冗余信息, 从而提高了特征提取的质量. 引入了一种特征引导的极坐标位置编码机制, 结合可学习的 BEV 嵌入, 它促进了宽度图像特征与查询之间的有效融合, 并通过单层交叉注意力机制生成强大的 BEV 特征. 与先前的视图转换技术相比, 本文提出的模型在 3D 目标检测任务上展现良好的性能, 同时在性能与精度之间实现了有效的平衡.

参考文献

1 周松燃, 卢烨昊, 励雪巍, 等. 车路两端纯视觉鸟瞰图感知

研究综述. 中国图象图形学报, 2024, 29(5): 1169–1187.
2 Shen YC, Geng ZG, Yuan YH, *et al*. V-DETR: DETR with vertex relative position encoding for 3D object detection. Proceedings of the 12th International Conference on Learning Representations. Vienna: ICLR, 2024.
3 Lin ZW, Liu Z, Xia ZY, *et al*. RCBEVDet: Radar-camera fusion in bird's eye view for 3D object detection. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2024. 14928–14937.
4 刘宏纬, 邵东恒, 杨剑, 等. 基于鸟瞰图融合的多级旋转等变目标检测网络. 计算机工程, 2024. [doi: 10.19678/j.issn.1000-3428.0068696]
5 Wang BL, Zhang L, Wang ZZ, *et al*. Core: Cooperative reconstruction for multi-agent perception. Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision. Paris: IEEE, 2023. 8676–8686.
6 Hu SC, Chen L, Wu PH, *et al*. ST-P3: End-to-end vision-based autonomous driving via spatial-temporal feature learning. Proceedings of the 17th European Conference on Computer Vision. Tel Aviv: Springer, 2022. 533–549.
7 Doll S, Hanselmann N, Schneider L, *et al*. DualAD: Disentangling the dynamic and static world for end-to-end driving. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2024. 14728–14737.
8 Huang JJ, Huang G. BEVDet4D: Exploit temporal cues in multi-camera 3D object detection. arXiv:2203.17054, 2022.
9 Huang JJ, Huang G, Zhu Z, *et al*. BEVDet: High-performance multi-camera 3D object detection in bird-eye-view. arXiv: 2112.11790, 2022.
10 Philion J, Fidler S. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3D. Proceedings of the 16th European Conference on Computer Vision. Glasgow: Springer, 2020. 194–210.
11 Li ZQ, Wang WH, Li HY, *et al*. BEVFormer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. Proceedings of the 17th European Conference on Computer Vision. Cham: Springer, 2022. 1–18.
12 Li YH, Ge Z, Yu GY, *et al*. BEVDepth: Acquisition of reliable depth for multi-view 3D object detection. Proceedings of the 37th AAAI Conference on Artificial Intelligence. Washington: AAAI, 2023. 1477–1485.
13 Vaswani A, Shazeer N, Parmar N, *et al*. Attention is all you need. Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 6000–6010.
14 Wang Y, Guizilini VC, Zhang TY, *et al*. DETR3D: 3D object detection from multi-view images via 3D-to-2D queries. Proceedings of the 5th Conference on Robot Learning. London: PMLR, 2022. 180–191.
15 Caesar H, Bankiti V, Lang AH, *et al*. nuScenes: A multimodal dataset for autonomous driving. Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 11618–11628.
16 Wang T, Zhu XG, Pang JM, *et al*. FCOS3D: Fully convolutional one-stage monocular 3D object detection. Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision. Montreal: IEEE, 2021. 913–922.
17 Tian Z, Shen CH, Chen H, *et al*. FCOS: A simple and strong anchor-free object detector. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(4): 1922–1933.
18 Liu YF, Wang TC, Zhang XY, *et al*. PETR: Position embedding transformation for multi-view 3D object detection. Proceedings of the 17th European Conference on Computer Vision. Tel Aviv: Springer, 2022. 531–548.
19 Liu YF, Yan JJ, Jia F, *et al*. PETRv2: A unified framework for 3D perception from multi-camera images. Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision. Paris: IEEE, 2023. 3239–3249.
20 Chen SY, Wang XG, Cheng TH, *et al*. Polar parametrization for vision-based surround-view 3D detection. arXiv:2206. 10965, 2022.
21 Xiong KX, Gong S, Ye XQ, *et al*. CAPE: Camera view position embedding for multi-view 3D object detection. Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023. 21570–21579.
22 Reading C, Harakeh A, Chae J, *et al*. Categorical depth distribution network for monocular 3D object detection. Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 8551–8560.
23 Liu ZJ, Tang HT, Amini A, *et al*. BEVFusion: Multi-task multi-sensor fusion with unified bird's-eye view representation. Proceedings of the 2023 IEEE International Conference on Robotics and Automation. London: IEEE, 2023. 2774–2781.
24 Xie EZ, Yu ZD, Zhou DQ, *et al*. M^2BEV: Multi-camera joint 3D detection and segmentation with unified birds-eye view representation. arXiv:2204.05088, 2022.

25 Zhou HY, Ge Z, Li ZM, *et al*. MatrixVT: Efficient multi-camera to bev transformation for 3D perception. Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision. Paris: IEEE, 2022. 8514–8523.

26 Yang WX, Li Q, Liu WX, *et al*. Projecting your view attentively: Monocular road scene layout estimation via cross-view transformation. Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 15531–15540.

27 Zhou B, Krähenbühl P. Cross-view transformers for real-time map-view semantic segmentation. Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 13750–13759.

28 Chen SY, Cheng TH, Wang XG, *et al*. Efficient and robust 2D-to-BEV representation learning via geometry-guided kernel transformer. arXiv:2206.04584, 2022.

29 金祖亮, 隗寒冰, Zheng L, 等. 基于局部窗口交叉注意力的轻量型语义分割. 汽车工程, 2023, 45(9): 1617–1625.

30 Zhu BJ, Jiang ZK, Zhou XX, *et al*. Class-balanced grouping and sampling for point cloud 3D object detection. arXiv: 1908.09492, 2019.

31 Yang CHY, Lin TW, Huang LC, *et al*. WidthFormer: Toward efficient transformer-based bev view transformation. arXiv:2401.03836, 2024.

(校对责编: 王欣欣)