计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
2025,34(11):212–219 [doi: 10.15888/j.cnki.csa.009981] [CSTR: 32024.14.csa.009981]

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

联合高度感知稀疏化与任务解耦的自动驾驶全景占用预测①

姜彦吉[1,3], 张　潇[1,3], 董　浩[2,3]

[1](辽宁工程技术大学 软件学院, 葫芦岛 125105)
[2](清华大学苏州汽车研究院, 苏州 215134)
[3](优策 (江苏) 安全科技有限公司 OpenSafe 实验室, 苏州 215100)
通信作者: 董　浩, E-mail: eason@utcer.com

摘　要: 基于环视相机的全景占用预测是自动驾驶环境理解的核心任务, 然而, 现有方法在提高检测精度和兼顾计算效率方面面临挑战, 传统体素化方法因全高度密集计算导致冗余, 而特征压缩会丢失高度方向细粒度信息, 多任务耦合进一步降低小目标预测精度. 本文基于 Panoptic-FlashOcc 提出一种动态稀疏体素引导的轻量化网络: (1) 设计动态稀疏体素采样机制, 通过可学习掩码预测高度方向自适应采样点, 减少无效计算; (2) 提出高度感知补偿模块, 通过 LSTM 编码和残差融合恢复空间细节; (3) 构建多任务解耦金字塔, 采用可变形卷积分离语义/实例特征流. 在 Occ3D-nuScenes 数据集上, 本文方法较基线在 RayIoU 指标上提升 5.5%, 达到 41.2%. 实验结果表明, 本文方法显著提升了小目标检测精度和全景占用预测任务的实时性.

关键词: 自动驾驶; 全景占用预测; 动态稀疏体素采样; 高度感知特征补偿; 多任务解耦特征金字塔

引用格式: 姜彦吉,张潇,董浩.联合高度感知稀疏化与任务解耦的自动驾驶全景占用预测.计算机系统应用,2025,34(11):212–219. http://www.c-s-a.org.cn/1003-3254/9981.html

Integration of Height-aware Sparsity and Task Decoupling for Panoptic Occupancy Prediction in Autonomous Driving

JIANG Yan-Ji[1,3], ZHANG Xiao[1,3], DONG Hao[2,3]

[1](Software College, Liaoning Technical University, Huludao 125105, China)
[2](Suzhou Automotive Research Institute, Tsinghua University, Suzhou 215134, China)
[3](OpenSafe Laboratory, Youce (Jiangsu) Security Technology Co. Ltd., Suzhou 215100, China)

Abstract: Panoptic occupancy prediction using surround-view cameras is a core task in environmental understanding for autonomous driving. However, existing methods face challenges in simultaneously improving detection accuracy and maintaining computational efficiency. Traditional voxelization approaches introduce redundancy due to full-height dense computation, while feature compression may lead to the loss of fine-grained information along the height dimension. Moreover, multi-task coupling further degrades the prediction accuracy of small objects. To address this issue, this study proposes a lightweight network guided by dynamic sparse voxels, based on Panoptic-FlashOcc: (1) A dynamic sparse voxel sampling mechanism is designed to predict adaptive sampling points in the height dimension using a learnable mask, effectively reducing redundant computation. (2) A height-aware compensation module is introduced to recover spatial details via LSTM encoding and residual fusion. (3) A multi-task decoupled pyramid is constructed, employing deformable convolution to separate semantic and instance feature streams. On the Occ3D-nuScenes dataset, the proposed method improves the RayIoU metric by 5.5% over the baseline, achieving a score of 41.2%. Experimental results

① 基金项目: 广东省科技创新战略专项市县科技创新支撑项目 (STKJ2023071); 浙江省自然科学基金面上项目 (LMS25G010003); 葫芦岛市科技计划 (2023JH(1)4/02b)
收稿时间: 2025-03-25; 修改时间: 2025-04-14; 采用时间: 2025-05-06; csa 在线出版时间: 2025-09-18
CNKI 网络首发时间: 2025-09-22

demonstrate that this approach significantly enhances small object detection accuracy and improves the real-time performance of panoptic occupancy prediction.

Key words: autonomous driving; panoptic occupancy prediction; dynamic sparse voxel sampling; height-aware feature compensation; multi-task decoupled feature pyramid

自动驾驶技术的快速发展对环境感知系统提出了更高的要求. 全景占用预测作为自动驾驶中的关键任务之一, 能够将 3D 场景从视觉图像中划分为结构化的体素, 并为每个体素分配实例 ID 和语义类别. 这种细粒度的环境理解不仅包括物体的类别 (如车辆、行人、建筑物等), 还能够区分不同的实例 (如不同的车辆或行人), 为自动驾驶系统提供了更丰富的信息, 使其能够更准确地感知周围环境.

语义占用预测作为全景占用预测的子任务, 旨在估计围绕自车的 3D 体素占用状态, 为自动驾驶系统提供全面的 3D 场景理解[1–5]. 早期的方法[6–9]依赖于复杂的 3D 卷积操作, 导致计算资源消耗过大. 为了提高全景占用预测的效率, 一些研究开始探索更高效的模型架构. TPVFormer[10]则通过三视角视图表示来补充垂直结构信息, 简化了体素级表示. 这些方法在一定程度上提高了计算效率, 但仍依赖于 3D 体素级表示或 Transformer 模块. 全景占用预测是语义占用地扩展, 在关注场景语义理解的基础上, 增添了实例级别的区分. SparseOcc[11]是第 1 个专注于全景占用预测的研究, 提出了一种完全稀疏的占用预测方法, 显著减少了计算量. Panoptic-FlashOcc[12]模型使用 2D 卷积替换 3D 卷积, 有效地解决了三维 voxel-level 中高内存和计算量大的缺陷, 但在处理复杂场景时, 对远距离和小目标的检测精度仍有待提高.

为了解决上述问题, 本文提出了一种基于 Panoptic-FlashOcc 模型的改进方法, 在其基础上引入了 3 个改进模块, 本文贡献工作如下.

(1) 动态稀疏体素采样 (DSS): 受完全稀疏的占用网络 SparseOcc 启发, 该网络提出场景的固有稀疏性 (超过 90% 的体素是空气). 本文提出动态稀疏体素采样机制, 该机制实时预测高密度和低密度区域, 通过引入一个可学习的动态掩码生成器, 生成动态高度采样掩码矩阵, 进而动态地调整采样密度, 使模型在处理不同场景下的体素采样问题时更加灵活和高效.

(2) 高度感知特征补偿 (FCM): 通过 LSTM[13]编码, 将稀疏特征补偿到稠密 BEV 平面特征中, 增强了高度信息的利用, 提高了小目标预测的准确性和鲁棒性.

(3) 多任务解耦特征金字塔 (DFP): 通过底层共享几何特征和高层分离语义与实例占用特征流, 利用可变形卷积[14]动态调整卷积核实现特征解耦, 适应不同任务需求.

1 动态稀疏体素采样机制

Panoptic-FlashOcc 虽然继承了 FlashOcc[15], 采用 2D 卷积替代 3D 卷积, 一定程度上减少了计算开销, 但全高度体素采样仍保留大量无效区域 (如天空、空旷地面等) 的计算. 因此本文提出一种动态稀疏体素采样机制, 以解决全高度体素计算中的冗余问题, 实现高度方向的自适应采样. 该机制引入可学习的稀疏掩码生成器, 生成动态高度采样掩码矩阵 M, 针对不同密度区域进行稀疏采样.

1.1 动态掩码生成器

将 Panoptic-FlashOcc 模型中 BEV 生成模块输出的 BEV 特征 f_{bev} 通过动态掩码生成器, 该生成器由两层全连接层和一个 *Sigmoid* 函数构成. 第 1 个全连接层配合 *ReLU* 激活函数, 引入非线性表达能力, 捕捉高度方向的关键模式 (例如车辆密集区域与天空的垂直分布差异); 第 2 个全连接层, 进一步将特征向量的维度压缩至 K (K 为预设最大采样层数, 其值的选取见第 4.5 节), 将场景语义特征映射到高度方向的不同层级, 每个维度对应一个高度层, 应用 *Sigmoid* 函数, 将每个高度层的重要性映射到[0, 1]区间, 趋近于 0 表示低密度区域可降采样, 趋近于 1 则表示高密度区域需精细采样. 动态掩码生成器计算过程如下:

$$P = \sigma(W_2 \cdot ReLU(W_1 \cdot f_{\text{bev}} + b_1) + b_2) \tag{1}$$

$$M_{i,j,k} = \begin{cases} 1, & \text{if } P_{i,j,k} > \theta \cdot \max(P_{i,j,k}) \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

其中, W_1 和 b_1 表示第 1 个全连接层的权重和偏置, W_2 和 b_2 表示第 2 个全连接层的权重和偏置, σ 表示 *Sig-*

moid 激活函数, P 表示每个 BEV 位置 (i, j) 的概率分布, θ 为动态阈值系数, 用于控制采样密度 (阈值的选取见第 4.5 节).

1.2 稀疏体素采样策略

在三维空间中, 并非需要对所有体素都进行采样, 而是根据场景的不同区域和重要性, 选择性地进行采样. 这种稀疏采样可以在保持关键信息的同时, 减少无效体素存储.

根据动态掩码生成器生成的掩码 M 从 f_{bev} 中收集非零特征. 计算如下:

$$f_{\text{sparse}} = Gather(f_{\text{bev}} \otimes M, non\text{-}zero\ indices) \tag{3}$$

其中, $\otimes$表示逐元素相乘, 用于生成稀疏特征图, 通过 *Gather* 操作提取出非零元素, 得到最终的稀疏体素特征 f_{sparse}. 流程如图 1 所示.

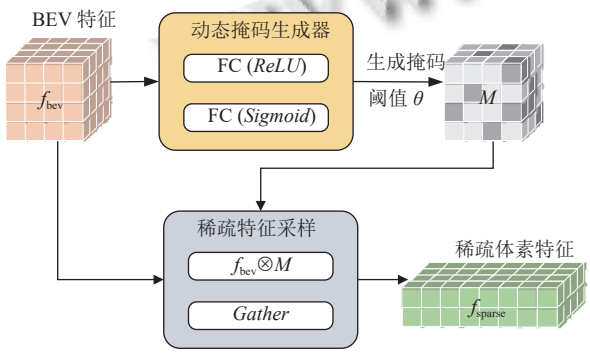


图 1 动态稀疏体素采样机制流程图

2 高度感知特征补偿

Panoptic-FlashOcc 中通道-高度变换模块将三维特征压缩到二维平面时, 可能会损失高度方向的细粒度信息. 为此, 本文通过高度方向残差融合补偿信息损失.

将稀疏体素 f_{sparse} 特征按照高度进行切片, 并将其作为序列输入到双向长短期记忆网络 (Bi-LSTM) 中, 每个时间 t 对应序列中的一个切片 $f_{\text{sparse}}^{(t)}$:

$$h_t = BiLSTM(f_{\text{sparse}}^{(t)}, h_{t-1}) \tag{4}$$

其中, h_t 表示在时间 t 的隐状态, 是通过 Bi-LSTM 处理当前切片 $f_{\text{sparse}}^{(t)}$和前一时间的隐状态 h_{t-1} 得到的.

然后将隐状态 h_t 映射到 BEV 平面, 生成一个与 BEV 特征图尺寸相同的特征图, 将时序信息转换为空间特征, 以便与 BEV 特征 f_{bev} 融合. 计算如下:

$$f_{\text{comp}} = f_{\text{bev}} + \alpha \cdot Conv_{1\times1}(h_t) \tag{5}$$

其中, α 为可学习参数, 在本文中, 将其设置为 0.5, f_{comp} 为补偿后的特征. 流程如图 2 所示.

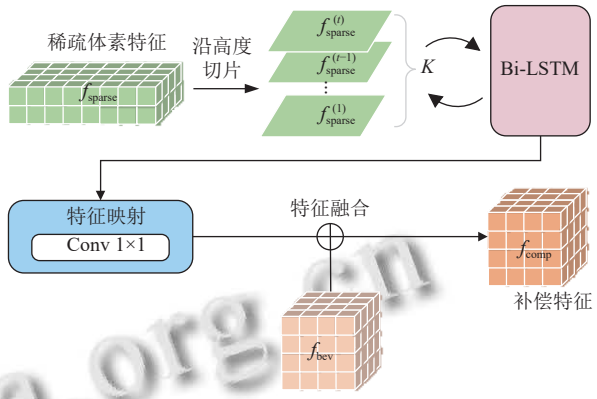


图 2 高度感知特征补偿流程图

3 多任务解耦特征金字塔

Panoptic-FlashOcc 的语义占用和实例占用预测头共享底层的 BEV 特征, 可能会导致交通标志等小物体的边缘细节模糊, 丢失关键信息. 为了更好地分离语义/实例占用任务的上下文特征并避免细节混淆, 本文提出了一种多任务解耦金字塔结构. 该结构包含多尺度特征偏移预测、任务特征解耦合和特征融合 3 个部分.

(1) 多尺度特征偏移预测: 针对不同尺度目标对偏移量的敏感性差异, 提出多尺度偏移量预测机制, 分别优化小目标 (细节敏感) 和大型物体 (结构敏感) 的偏移学习. 首先, 采用 3×3 卷积对对补偿体素特征 f_{comp} 进行处理, 捕捉局部边缘特征, 生成小尺度偏移量 Δp_{detail}. 其次, 采用 5×5 卷积, 扩大感受野以捕捉大物体轮廓特征, 生成大尺度偏移量 Δp_{struct}, 通过门控权重动态融合两类偏移, 进而预测语义/实例占用任务所需的卷积核偏移量 Δp_l, 用于指导后续的可变形卷积操作, 从而使得模型能够自适应地调整特征提取过程以适应不同任务需求.

$$\Delta p_l = \alpha \cdot \Delta p_{\text{detail}} + (1-\alpha) \cdot \Delta p_{\text{struct}} \tag{6}$$

其中, Δp_l 表示第 l 层 (本文设置可变形卷积层数 L 取值为 2) 的偏移量, 为每个任务分配 $2L$ 通道, 前 $2L$ 为语义偏移量 Δp_{seg}, 后 $2L$ 通道为实例偏移量 Δp_{ins}. 门控系数 α=*Sigmoid*(*MLP*(f_{comp})) 为自适应权重. 当输入特征对应大物体时, MLP 倾向于降低 α 值, 从而增强结构分支 (5×5 卷积) 的权重; 当输入特征对应小物体时, MLP 则提升 α 值, 偏重细节分支 (3×3 卷积). 本文采用

了轻量级的双层 MLP, 以保持较低的计算开销.

(2) 任务特征解耦: 通过可变形卷积 *DeformConv* 将补偿特征 f_{comp} 与预测的偏移量结合, 使其解耦为多个任务特定的特征表示 f_{task}, 使得每个任务都能获得最相关的信息.

$$f_{\text{task}} = \sum_{l=1}^{L} DeformConv(f_{\text{comp}}, \Delta p_l) \tag{7}$$

(3) 特征融合: 将解耦后的任务特征与原始补偿特征 f_{comp} 进行融合, 生成最终的特征表示.

$$f_{\text{final}} = f_{\text{comp}} + \beta \cdot f_{\text{task}} \tag{8}$$

其中, β 是用于平衡共享特征和任务特征的可学习参数, 在本文中, 将其取值为 0.2. 流程如图 3 所示.

本文提出的改进模型总体流程图如图 4 所示.

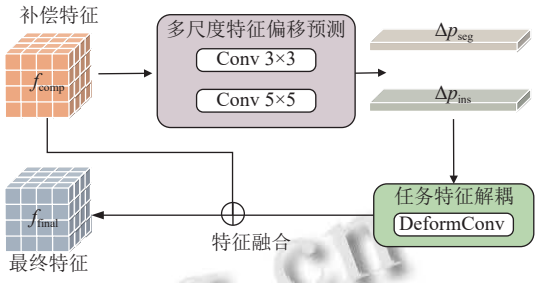


图 3 多任务解耦流程图

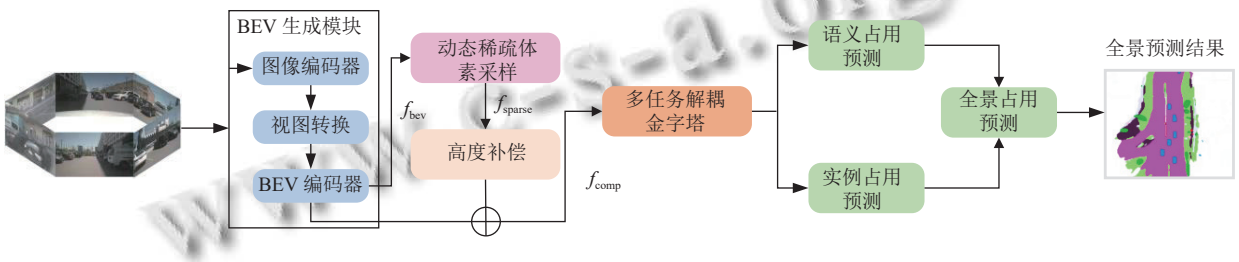


图 4 总体流程图

4 实验分析

4.1 数据集和评价指标

本文实验将基于 Occ3D-nuScenes 数据集进行, 该数据集由清华大学和英伟达等机构基于 nuScenes 数据集构建的一个大规模 3D 占用预测数据集. 该数据集包含 600 个训练场景、150 个验证场景和 150 个测试场景, 总计 40 000 帧. 其场景范围沿 X 轴和 Y 轴设置为 [−40, 40] m, 沿 Z 轴设置为[−1, 5.4] m. 该数据集通过半自动标注 pipeline 生成, 该 pipeline 利用现有的标记 3D 感知数据集, 并根据其可见性识别体素类型. Occ3D-nuScenes 数据集的复杂性和多样性使其成为评估 3D 占用预测方法性能的重要基准. 数据集示例如图 5 所示.

图 5 Occ3D-nuScenes 数据集示例

本文使用文献[13]提出的 RayIoU 和旧体素级 mIoU 指标评估所提模型在语义占用预测任务上的性能. 此外, 使用全景质量 (RayPQ) 作为全景占用预测任务的评价指标.

4.2 实验环境和训练细节

本文模型使用 PyTorch 框架实现, 在配备了 8 块 GeForce RTX 4090 显卡的服务器上进行分布式训练, batch_size 设置为 16, epoch 设置为 24, 采用 Adam 优化器, 初始学习率为 1×10^{-4}. 此外, FPS 是在 Tesla A100 GPU 上使用 PyTorch fp32 (1 个 batch_size) 测量的.

4.3 性能对比实验

本文对所提模型在语义占用预测任务中的表现进行了全面评估, 并与其他方法进行了对比分析. 如表 1 所示, 与基线模型 (Panoptic-FlashOcc) 相比, 本文方法在 RayIoU 指标上实现了 1.9%–5.5% 的显著提升, 尤其是在中距离 (2 m) 范围内性能提升最为明显, 从 39.3% 增至 42.0%. 在效率方面, 本文方法在单帧处理时的推理速度达到 39.8 f/s, 相较于基线模型的 38.7 f/s 有所提升, 即使在处理 8 帧时仍能保持 35.6 f/s, 满足了实时处理的要求 (>30 f/s).

表 2 展示了基线模型和所提方法在全景占用预测任务上的性能对比. 相较于基线模型, 本文方法单帧 (1 frame) 的 RayPQ 提升到 15.1% (增加了 1.9%), 远距离 (4 m) 从 16.8% 提升到 18.6% (增加了 1.8%), 同时小物

体检测显著提升, RayPQ 在 1 m 以下从 9.2% 提升到 11.5% (增加了 2.3%). 此外, 在效率方面, 该方法保持了优势, 单帧推理速度达到 37.5 f/s.

此外, 为了验证所提方法在提高小目标检测精度方面的有效性, 表 3 展示了基线模型与本文方法在 Occ3D-nuScenes 数据集上各类别的 RayIoU 性能对比. 从表中可以看出, 当扩展到 8 帧时, 相较于基线模型, 本文方法在交通锥 (tfc. cone)、自行车 (bicycle) 和行人 (pedes.) 等小目标的检测上取得了显著提升, 分别提升了 8.3%、6.6% 和 4.9%. 同时, 在公交车 (bus) 和卡车 (truck) 等大型目标的检测上也分别取得了 5.4% 和 6.3% 的明显提升.

表 1 在 Occ3D-nuScenes 数据集上 3D 语义占用预测性能对比

方法	Backbone	Input size	Epochs	RayIoU (%)	RayIoU_{1m} (%)	RayIoU_{2m} (%)	RayIoU_{4m} (%)	mIoU (%)	FPS (f/s)
BEVFormer[16]	R101	1600×900	24	32.4	26.1	32.9	38.0	39.3	3.0
BEVDetOcc	R50	704×256	90	29.6	23.6	30.0	35.1	36.1	2.6
BEVDetOcc (8f)	R50	704×384	90	32.6	26.6	33.1	38.2	39.3	0.8
FB-OCC (16f)[6]	R50	704×256	90	33.5	26.7	34.1	39.7	39.1	10.3
SpareOcc (8f)	R50	704×256	24	34.0	28.0	34.7	39.4	30.6	17.3
SpareOcc (16f)	R50	704×256	24	35.1	29.1	35.8	40.3	30.9	12.5
Panoptic-FlashOcc-tiny	R50	704×256	24	34.8	29.1	35.7	39.7	29.1	43.9
Panoptic-FlashOcc (1f)	R50	704×256	24	35.2	29.4	36.0	40.1	29.4	38.7
Panoptic-FlashOcc (2f)	R50	704×256	24	36.8	31.2	37.6	41.5	30.3	35.9
Panoptic-FlashOcc (8f)	R50	704×256	24	38.5	32.8	39.3	43.4	31.6	35.6
本文方法 (1f)	R50	704×256	24	37.1	31.0	37.6	42.0	31.5	39.8
本文方法 (2f)	R50	704×256	24	38.5	32.8	39.4	43.5	33.2	37.2
本文方法 (8f)	R50	704×256	24	41.2	35.2	42.0	46.3	34.5	35.6

表 2 在 Occ3D-nuScenes 数据集上全景占用预测性能对比

方法	Backbone	Input size	Epochs	RayPQ (%)	RayPQ_{1m} (%)	RayPQ_{2m} (%)	RayPQ_{4m} (%)	FPS (f/s)
Panoptic-FlashOcc-tiny	R50	704×256	24	12.9	8.8	13.4	16.5	39.8
Panoptic-FlashOcc (1f)	R50	704×256	24	13.2	9.2	13.5	16.8	35.2
Panoptic-FlashOcc (2f)	R50	704×256	24	14.5	10.6	15.0	18.0	30.4
Panoptic-FlashOcc (8f)	R50	704×256	24	16.0	11.9	16.3	19.7	30.2
本文方法 (1f)	R50	704×256	24	15.1	11.5	15.5	18.6	37.5
本文方法 (2f)	R50	704×256	24	16.9	13.1	17.6	20.6	33.1
本文方法 (8f)	R50	704×256	24	19.1	14.9	19.4	22.5	30.3

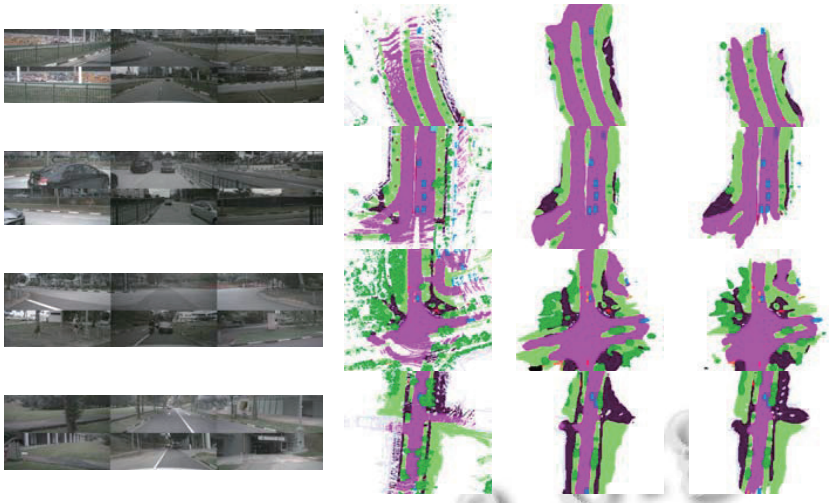
表 3 基于 Occ3D-nuScenes 数据集不同类别的 RayIoU 性能对比 (%)

方法	RayIoU	others	barrier	bicycle	bus	car	cons. veh.	motor	pedes.	tfc. cone	trailer	truck	drv. surf.	other flat	sidewalk	terrain	manmade	vegetation
基线 (1f)	35.2	28.3	42.1	18.7	54.3	52.8	36.5	21.5	25.6	25.3	33.2	46.7	68.5	65.2	61.8	58.3	55.6	60.1
基线 (2f)	36.8	29.1	43.6	19.3	56.1	54.2	38.1	22.8	26.9	26.7	34.8	48.2	69.3	66.4	62.9	59.6	56.7	61.5
基线 (8f)	38.5	30.5	45.2	**20.6**	**58.8**	56.7	40.3	24.1	**28.3**	**27.9**	36.5	**50.1**	70.8	67.9	64.2	61.1	58.4	63.0
本文方法 (1f)	37.1	30.1	43.8	22.3	58.2	56.2	38.5	24.5	29.2	30.8	35.3	49.8	69.8	67.8	65.4	60.7	57.5	62.3
本文方法 (2f)	38.5	31.5	46.2	24.8	60.1.	58.4	41.0	27.1	30.8	33.5	37.7	52.3	71.5	69.6	66.2	62.5	59.7	64.1
本文方法 (8f)	41.2	33.0	48.5	**27.2**	**64.2**	62.5	43.9	28.7	**33.2**	**36.2**	40.5	**56.4**	74.3	71.2	68.3	64.0	61.4	66.4

图 6 和图 7 展示了真实标签、Panoptic-FlashOcc 及本文方法的语义和全景占用预测结果的可视化对比. 可以看出, 本文方法在细节处理和边缘识别方面优于 Panoptic-FlashOcc, 有效减少了模糊和过度覆盖的现象, 与真实标签的匹配度更高.

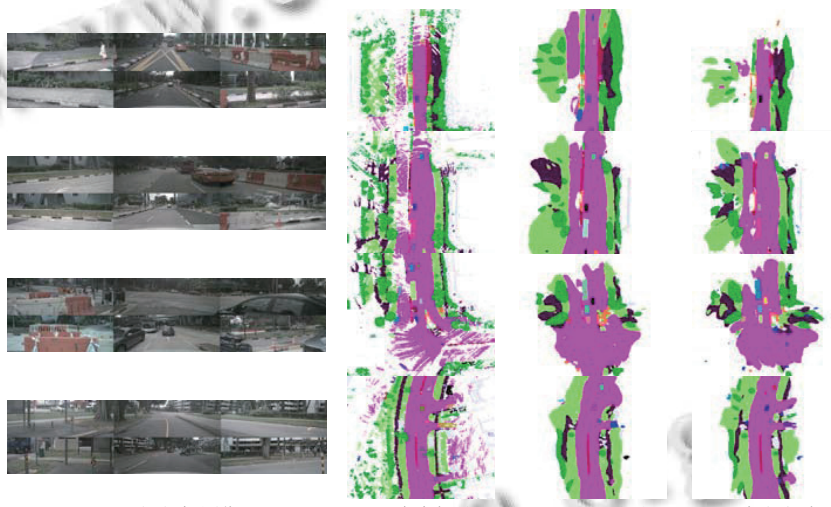
4.4 消融实验

为了验证各组件对本文所提模型的影响, 在 Occ3D-nuScenes 数据集上进行了消融实验 (下面实验均仅融合 1 帧的时间信息), 如表 4 所示. 动态采样模块 (DSS) 带来了 0.3% 的 RayIoU 提升, 同时由于计算量的减少, FPS 提升 2.7 f/s. 高度补偿模块 (FCM) 带来了 1.1% 的 RayIoU 增益. 多任务解耦特征金字塔模块 (DFP) 实现了语义和实例特征的分离, 使得 RayIoU 提升了 1.9%, mIoU 提升了 2.1%. 结果表明, 本文方法在提升预测精度的同时, 也优化了计算效率, 对于自动驾驶系统中的全景占用预测任务具有重要的实际应用.



(a) 环视相机图像 (b) 真实标签 (c) Panoptic-FlashOcc (d) 本文方法

图 6 语义占用预测结果



(a) 环视相机图像 (b) 真实标签 (c) Panoptic-FlashOcc (d) 本文方法

图 7 全景占用预测结果

表 4 消融实验

组件	RayIoU (%)	mIoU (%)	FPS (f/s)
基线 (Panoptic-FlashOcc (1f))	35.2	29.4	38.7
+DSS	35.5	29.8	41.4
+FCM	36.3	30.7	41.1
+DFP	37.1	31.5	39.8

4.5 参数选取

动态阈值 θ 和高度方向采样层数 K 是平衡全景语义占用预测任务精度和计算效率的关键. 为了合理选取这两个参数, 本文对 Occ3D-nuScenes 数据集进行了深入分析.

首先, 提取数据集中每个目标的高度信息, 并按照目标类别 (车辆、行人、交通锥等) 分组统计其在各高度区间的频数. 随后, 使用核密度估计生成连续密度曲线 (如图 8 所示), 并通过梯度法定位密度峰值区间, 结果如表 5 所示.

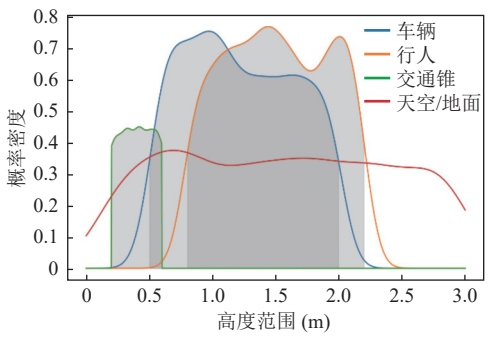


图 8 不同目标类别的概率密度曲线

由表 5 可知, 当动态阈值 θ 为 0.6 和 0.7 时, 接近行人和车辆的密度峰值, 可最大化保留有效目标信息. 而当 θ 为 0.4 时, 虽然能够覆盖交通锥等低矮物体, 但可能会引入噪声. 此外, 还应验证 θ 为 0.8 时, 过度稀疏化可能带来的影响. 基于经验法则, 以 0.1 为间隔覆盖典型操作点, 选取$\theta \in \{0.4, 0.5, 0.6, 0.7, 0.8\}$进行实验, 实验结果如表 6 所示. 当 θ 设置为 0.7 时, 模型的精度最高; 而当 θ 设置为 0.8 时, 虽然因过滤掉大部分体素导致计算量减少, 但交通锥等小目标物体几乎被全部过滤, 导致漏检增加, 精度明显下降.

表 5 Occ3D-nuScenes 数据集分析

目标类别	高度范围 (m)	概率密度峰值区间
车辆	0.5–2.0	0.65–0.75
行人	0.8–2.2	0.55–0.65
交通锥	0.2–0.6	0.35–0.45
天空/地面	>3.0或<0.1	<0.3

表 6 动态阈值选取实验

阈值	RayIoU (%)	mIoU (%)	FPS (f/s)
0.4	34.5	28.8	34.0
0.5	35.0	29.5	35.7
0.6	36..1	30.6	38.3
0.7	**37.1**	**31.5**	39.8
0.8	35.2	31.0	44.7

在高度方向采样层数 K 的选取上, 考虑到数据集 Occ3D-nuScenes 中大多数有效点云数据集中在地面至大约 4 m 的高度范围内, 本文设计了 3 种不同的采样层数 K 值进行实验. 当 K=4 时, 每层覆盖 1 m 的高度, 适合捕捉较大物体轮廓; 当 K=8 时, 每层覆盖 0.5 m, 能够更精细地捕捉行人和交通锥等较小物体的细节; 当 K=12 时, 每层覆盖 0.33 m, 虽然提高了采样密度, 但同时也可能引入噪声. 实验结果如表 7 所示, 当 K 值为 8 时, 模型性能最佳, 能够更准确地捕捉关键细节, 提高检测精度. 而当 K 增加到 12 时, 尽管采样密度进一步提升, 但同时引入了更多的噪声, 导致检测精度有所下降, 并且显存占用更高.

表 7 采样层数选取实验

采样层数	RayIoU (%)	mIoU (%)	显存 (GB)
4	34.2	29.8	5.2
8	**37.1**	**31.5**	6.9
12	35.5	30.6	8.5

综上所述, 动态阈值 θ=0.7 和采样层数 K=8 是平衡精度和计算效率的最佳选择.

5 结论与展望

本文提出一种基于动态稀疏体素采样的高精度轻量化全景占用预测框架, 通过可学习掩码实现高度方向自适应采样, 减少了密集体素的计算开销, 同时引入高度补偿机制缓解高度信息损失. 多任务解耦金字塔结构通过可变形卷积实现语义/实例特征解耦, 使 mIoU 提升 2.1%. 实验表明, 该方法在 nuScenes 数据集上以 41.2% RayIoU 达到 SOTA 性能, 且单帧推理速度达 39.8 f/s. 未来工作将探索时序特征与动态采样的协同优化, 进一步提升运动物体预测精度.

参考文献

1 Miao RH, Liu WZ, Chen MR, *et al*. OccDepth: A depth-aware method for 3D semantic scene completion. arXiv:2302.13540, 2023.

2 Tian XY, Jiang T, Yun LF, *et al*. Occ3D: A large-scale 3D occupancy prediction benchmark for autonomous driving. Proceedings of the 37th Conference on Neural Information Processing Systems. New Orleans: NeurIPS, 2023. 14365.

3 Tong WW, Sima CH, Wang T, *et al*. Scene as occupancy. Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision. Paris: IEEE, 2023. 8372–8381.

4 Wang XF, Zhu Z, Xu WB, *et al*. OpenOccupancy: A large scale benchmark for surrounding semantic occupancy perception. Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision. Paris: IEEE, 2023. 17804–17813.

5 Zhang YP, Zhu Z, Du DL. OccFormer: Dual-path Transformer for vision-based 3D semantic occupancy prediction. Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision. Paris: IEEE, 2023. 9399–9409.

6 Li ZQ, Yu ZD, Austin D, *et al*. FB-OCC: 3D occupancy prediction based on forward-backward view transformation. arXiv:2307.01492, 2023.

7 Pan MJ, Liu JM, Zhang RR, *et al*. RenderOcc: Vision-centric 3D occupancy prediction with 2D rendering supervision. Proceedings of the 2024 IEEE International Conference on Robotics and Automation. Yokohama: IEEE, 2024. 12404–12411.

8 Wang YQ, Chen YT, Liao XY, *et al*. PanoOcc: Unified occupancy representation for camera-based 3D panoptic segmentation. Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2023. 17158–17168.

9 Wei Y, Zhao LQ, Zheng WZ, *et al*. SurroundOcc: Multi-camera 3D occupancy prediction for autonomous driving. Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision. Paris: IEEE, 2023. 21672–21683.

10 Liu HS, Wang HG, Chen Y, *et al*. Fully sparse 3D panoptic occupancy prediction. arXiv:2312.17118v1, 2024.

11 Huang YH, Zheng WZ, Zhang YP, *et al*. Tri-perspective view for vision-based 3D semantic occupancy prediction. Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023. 9223–9232.

12 Yu ZC, Shu CY, Sun QP, *et al*. Panoptic-FlashOcc: An efficient baseline to marry semantic occupancy with panoptic via instance center. arXiv:2406.10527, 2024.

13 Hochreiter S, Schmidhuber J. Long short-term memory. Neural Computation, 1997, 9(8): 1735–1780. [doi: 10.1162/neco.1997.9.8.1735]

14 Dai JF, Qi HZ, Xiong YW, *et al*. Deformable convolutional networks. Proceedings of the 2017 IEEE International Conference on Computer Vision. Venice: IEEE, 2017. 764–773.

15 Yu ZC, Shu CY, Deng JJ, *et al*. FlashOcc: Fast and memory-efficient occupancy prediction via channel-to-height plugin. arXiv:2311.12058, 2023.

16 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. Tel Aviv: Springer, 2022. 1–18.

(校对责编: 张重毅)