# 基于占用预测的路端单目 3D 目标检测

任柄禄，尹建芹
（北京邮电大学人工智能学院，北京 100876）

**摘要：**近年来，基于车端传感器的自动驾驶感知算法得到了快速发展，而基于路边基础设施上的摄像头的感知算法仍处于起步阶段。这些放置在较高位置的相机对交通路口进行俯瞰，在视角上具有固有的优势，能够提供遮挡很少的视野。但是与其他基于单目相机的算法一样，由于 RGB 图像中缺乏深度信息，基于路边相机的 3D 目标检测仍然面临许多挑战。本文针对路端相机的视角特点，设计了一种新的基于占用预测的 3D 物体检测算法，称为 RoadOcc。具体而言，我们避免了使用深度图作为复杂的深度预测网络的监督信号，而是使用激光雷达点云来监督待感知区域内的占用预测，通过挖掘场景的几何和语义特征辅助 3D 目标检测网络的学习。基于公开真实世界数据集 DAIR-V2X-I 的实验表明，我们的方法可以高效地实现对复杂交通路口中感兴趣目标的检测，尤其是可以提高对行人、骑行者等小目标的检测精度。

**关键词：**模式识别 计算机视觉 3D 目标检测
**中图分类号：**TP391

# RoadOcc: Roadside Monocular 3D Object Detection via Occupancy Prediction

Ren Binglu, Yin Jianqin
(School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing 100876)

**Abstract:** In recent years, autonomous driving perception algorithms based on vehicle sensors have developed rapidly, while perception algorithms based on cameras on roadside infrastructure are still in their early stages. These cameras placed at higher positions have inherent advantages in terms of perspective when looking down at traffic intersections, providing a view with minimal occlusion. However, like other algorithms based on monocular cameras, 3D object detection based on roadside cameras still faces many challenges due to the lack of depth information in RGB images. In this paper we propose a new 3D object detection algorithm based on occupancy prediction, called RoadOcc, , tailored to the perspective characteristics of roadside cameras. Specifically, we avoid using depth maps as supervisory signals for complex depth prediction networks, and instead use LiDAR point clouds to supervise occupancy prediction within the perceived area, assisting the learning of 3D object detection networks by mining the geometric and semantic features of the scene. Experiments on the publicly available real-world dataset DAIR-V2X-I have shown that our method can efficiently detect target objects in complex traffic intersections, especially improving the detection accuracy of small targets such as pedestrians and cyclists.

**Key words:** Pattern Recognition; Computer Vision; 3D Object Detection

## 0 引言

近年来，基于视觉的自动驾驶算法在学术界和工业界都受到了极大的关注。其中，基于单目相机的 3D 目标检测是一种使用 RGB 图像作为输入来估计视场内感兴趣的目标物体

**基金项目：**国家自然科学基金资助项目（No. 62173045，62273054）；中央高校基本科研业务费专项资金资助(Grant No. 2020XD-A04-3)
**作者简介：**任柄禄(1999-)，男，硕士研究生，主要研究方向：计算机视觉、3D 目标检测
**通信联系人：**尹建芹(1978-)，女，教授、博士生导师，主要研究方向：服务机器人、计算机视觉. E-mail: jqyin@bupt.edu.cn

的 3D 信息的任务，包括目标在 3D 空间中的位置、大小和方位角。然而，大多数 3D 目标检测工作都集中在车端传感器上，而忽略了广泛分布的路端摄像头。这些摄像头安装在距离地面一定高度的基础设施上，因此视野开阔，障碍物很少，可以观察到大量路况信息。这些信息可以填补自动驾驶汽车的视觉盲区，从而提高智能交通系统的安全性。

与单车仅使用车端传感器相比，基于路端传感器的感知具有视角上的固有特点，具有较少遮挡的视角为使用轻量化的特征提升模块提供了可能。单目 3D 目标检测算法常用基于深度估计的 LSS[1]方法进行特征提升，将 2D 特征沿着射线放置在 3D 空间中的对应位置，而另一种轻量化的基于投影的特征提升方法[2]是通过体素中心点的投影直接采样得到 3D 特征。在车载设备的视角下，同一射线穿过的体素之间互相存在严重的遮挡情况，而路端感知视角的遮挡情况较少，因此本文提出使用轻量化的基于投影的特征提升方法，更适合路端感知场景的高效 3D 体素特征生成。

由于基于投影的特征提升方法是一种非参数化方法，本文进一步设计一种基于稀疏 3D 卷积层的占用预测网络用于挖掘 3D 空间中的几何和语义特征。在 DAIR-V2X-I 数据集[3]上的实验证明，本文提出的路端占用网络 RoadOcc 能够充分利用路端感知视角的优势，高效地增强体素特征对目标物体信息的表征能力。

# 1 相关工作

## 1.1 单目 3D 目标检测

给定一张 RGB 图像和相应的相机参数，基于单目图像的 3D 目标检测旨在对感兴趣的目标进行分类和定位，其关键问题和难点在于如何从 2D 数据中获得 3D 的结果。相关研究工作可以分为基于结果提升、数据提升和特征提升的方法。其中基于结果提升的方法首先估计 2D 目标候选框和目标深度，然后将其提升至 3D 空间。3DOP[4]和 Mono3D[5]使用几何先验计算候选框的置信度从而滤除低置信度候选框。CenterNet[6]提出将目标物体编码为单个关键点。但该类方法存在过于依赖 2D 目标检测网络而缺乏对 3D 空间建模能力的问题。基于数据提升的方法如 Pseudo-lidar[7]提出首先以图像为输入生成其对应的密集深度图，然后通过反投影生成伪云，进而应用其他基于 LiDAR 点云的方法，但容易受到深度估计准确性的影响。基于特征提升的方法将 2D 特征提升到 3D 体素特征，OFTNet[8]提出将每个体素的边界框进行投影，并对矩形投影区域内的特征进行求和以获得该体素的 3D 特征。LSS 提出将连续的深度空间的深度估计视为分类任务，通过深度向量与像素特征的外积得到视锥特征，最后使用体素池化生成 3D 体素特征，但这一过程会消耗较大的计算资源且体素池化操作较为复杂。

## 1.2 3D 占用预测

3D 占用预测任务目的是给定 2D 图像预测指定区域内全部体素的占用情况和语义标签。MonoScene[9]是相关工作中第一种基于相机的占用预测方法，它可以从单个图像中预测出占用语义标签。TPVFormer[10]提出了一种三视角视图表示来生成周围的占用预测，但由于采用车载 LiDAR 获取的稀疏点云进行监督，其占用输出也相对稀疏。OccFormer[11]基于 LSS 将 2D 图像特征提取到 3D 体素特征，然后使用双路径 Transformer[12]模块对其进行编码，减轻

占用预测的稀疏性和类不平衡问题。后续工作基于汽车环视 BEV 感知场景进行 3D 占用预测，如 SurroundOcc[13]、OpenOccupancy[14]，这些方法在车载视角下，主要通过构建密集的占用预测标签和 Transformer 结构来实现占用输出，而本文利用路端视角下遮挡情况少的优势，仅使用卷积层实现高效的占用预测。

# 2 方法

## 2.1 稀疏占用预测标签生成

在传统的计算机视觉任务中，例如基于图像的 2D 目标检测或分割，可以通过手动注释来获得真值标签。然而，在 3D 占用预测任务中，场景的真值标签由在 3D 空间中均匀分割的 X*Y*Z 体素表示，其中每个体素具有语义标签。如 SemanticKITTI[15]中所述，如果这些标签完全是手动注释的，则会消耗太多成本，因为 3D 场景的密集占用可能具有数百万个体素。

一些车端的占用预测工作提出了复杂的自动化或者半自动化方法用于生成密集的占用标签。然而，这些方法往往依赖时序上连续的多帧数据进行融合或者使用其他 3D 重建方法。这些方法的动机之一是车载激光雷达传感器往往是通过旋转发射平行于水平面的光束进行扫描的方式获取点云数据。因此，路面上的其他车辆、建筑等等物体都极容易将激光雷达的光束遮挡，形成大面积的视野盲区。同时，激光雷达点云的发散式扫描，使点云的稀疏性随着距离的增加而迅速加剧，即位于较远距离处的物体反射的激光光束十分少。我们分析了路端应用场景的特性发现，由于路端激光雷达安装在距离地面较高的位置，且其激光光束与地面形成较大的夹角，因此上述问题有所缓解。如图 1 所示，与车端视角相比，在一定的感知范围内路端场景深度的变化程度较小，遮挡情况较少，为 3D 占用预测标签的自动生成提供了有利的条件。

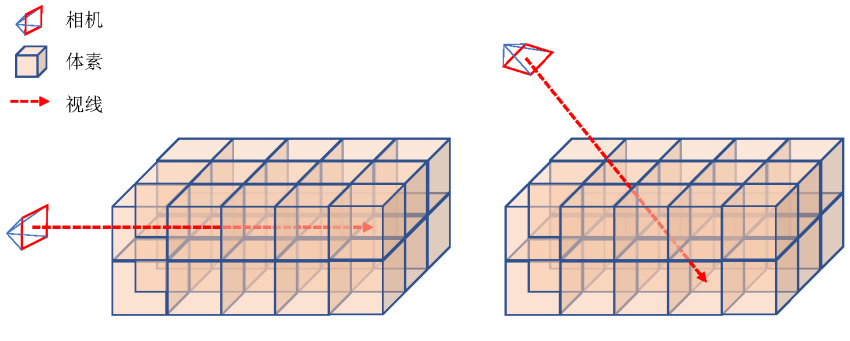


图 1 车端与路端视角对比示意图

Fig. 1 comparison between vehicle and road perspectives

出于上述考虑，我们仅使用单帧路端点云数据和 3D 目标检测标注框进行该场景的 3D 占用预测标签生成。具体来说，给定感知范围内包含 N 个点的点云数据和 K 个目标的标注框，我们首先为每个点分配语义标签。在标注框 $B_i$ 覆盖的长方体空间内的局部点云 $P_i$ 具有相同的语义标签 i，不在任何标注框范围内的点标记为背景类。然后我们使用动态体素化方法对附加了语义标签的点云进行体素化，位于同一个体素内的局部点云 $P_v$ 具有相同的体素索引 v。局部点云 $P_v$ 可能包含多种类别的点，出于效率考虑，我们使用散射最大值的方法为

该体素$V_v$分配一个语义标签作为占用预测的语义标签。即占用预测的类别包括空、背景类、目标类别，这一过程公式如下所示:

$$V = ScatterMax(P, Index) \tag{1}$$

其中 *Index* 为点云指向体素的索引。我们的方法可以高效地利用现有的适用于 3D 目标检测任务的多模态数据集进行全自动化占用预测标签生成，而不需要额外的补充数据，同时这一方法还可以拓展到其他类似的数据集。

## 2.2 路端占用网络

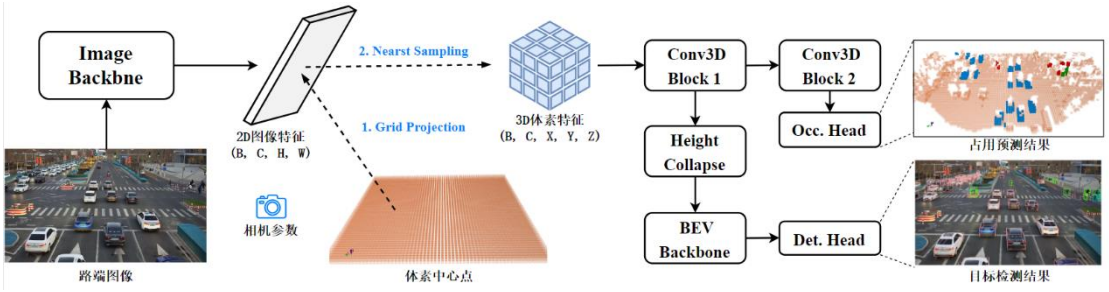


图 2 路端占用网络整体流程图

Fig. 2 Overview of RoadOcc network

我们提出的方法整体流程图如图 2 所示。给定一个路端基础设施拍摄的 RGB 图像 $I \in \mathbb{R}^{H\times W\times 3}$ 以及相应的相机内参 *Intr* 和外参 *Extr*，我们设计一个端到端的网络来实现占用预测辅助监督下的 3D 目标检测。我们首先利用常用的 2D 卷积骨干网络对输入图像进行特征提取，然后基于相机内外参将 2D 特征反向投影到 3D 空间，得到 3D 体素特征。接下来，我们分别设计了占用预测支路和目标检测支路，同时预测该场景中可见的指定区域内的目标检测结果以及占用预测结果：

$$B_{predict} = \{(x_i, y_i, z_i, l_i, w_i, h_i, \theta_i)\} \tag{2}$$

$$Occ_{predict} \in \mathbb{R}^{N_x\times N_y\times N_z\times C_{cls}} \tag{3}$$

其中生成的目标框包括物体中心坐标、表示长度、宽度、高度和围绕 z 轴的旋转角度，生成的占用预测结果中 $N_x$、$N_y$、$N_z$ 分别表示沿着 x、y、z 轴方向的体素的数量，$C_{cls}$ 表示体素占用的类别，在本实验所使用的数据集中，包含空白类、背景类、车、行人、骑行者 5 种类别。2D 特征投影方法和目标检测头部网络由其他工作提出，这里我们简单概述这些过程，然后介绍我们设计的路端占用预测整体网络结构。

RGB 图像本身不包含深度信息，因此要实现 3D 空间内的感知，如何将 2D 图像特征“提升”到 3D 空间是一个关键且困难的问题。一种常用的方法是基于深度的特征提升方法，例如 LSS，利用卷积网络预测每个像素的深度，通过外积将像素特征沿着射线按照一定的权重分布放置，形成视锥特征，进而使用体素池化技术将视锥特征转换为 3D 特征形式。CaDDN[16] 提出对其中的深度网络进行监督学习，通过提高深度估计的准确性来达到优化特征提升效果的目的。可见这些基于深度的方法依赖比较准确的深度估计过程，最近提出的 BEVHeight[17] 发现在路端感知场景使用基于深度的方法可能是次优的。由于路端场景下路面与摄像头光轴形成一定的夹角而不是平行，因此目标物体和地面之间的深度差较小，造成前景点与背景点之间特征差异很小，十分容易受到深度估计误差的影响。相比之下，另一种基于相机内外参数的特征提升方法是将预先定义的体素中心点投影到图像平面，通过插值的方式将对应的

2D 特征采样到 3D 空间，计算资源消耗较低。在车端应用场景中，由近到远距离的体素之间存在严重的交叠现象，这种方法容易产生特征模糊，即相同射线上的多个体素可能采样到相似的图像特征。而在路端场景中，摄像头的安装位置提供了具有极少遮挡现象的视角，因此使用这一特征提升方法具有显著优势。具体来说，我们遵循 Fast-BEV[18]中提出的快速射线转换方法(Fast-Ray Transformation)，将该环视图像反向投影特征提取方法应用到基于路端单目摄像头的场景，以实现高效的特征提升。

具体来说，给定图像特征 $I \in \mathbb{R}^{H \times W \times C}$ 、相机投影矩阵 $M \in \mathbb{R}^{3\times4}$，我们按照预先设定的体素尺寸以及待感知区域范围计算每个体素的齐次形式中心点坐标 $P_{3D} \in \mathbb{R}^{N_x \times N_y \times N_z \times 4}$，进而计算得到其在图像平面的投影点 $P_{2D} \in \mathbb{R}^{N_x \times N_y \times N_z \times 2}$：

$$P_{2D} = Trans(M \times P_{3D}) \tag{4}$$

其中 $Trans$ 表示由相机坐标系到图像特征像素坐标系的转换，每个体素特征根据该体素的投影点进行最近邻采样得到，最终形成 3D 体素特征图：

$$V_{3D} \in \mathbb{R}^{N_x \times N_y \times N_z \times C} \tag{5}$$

占用预测模块的目标可以定义为给定 3D 体素特征，输出待感知空间中每个体素的类别的置信度，由于体素数量较大，因此这一过程一般会消耗较大的计算资源。由于目前工作中一般使用 Transformer 进行隐式的特征提升，只需加入 Softmax 分类头就可以输出体素占用的类别置信度。本方法使用非参数化的 2D 到 3D 特征投影方式，因此针对其生成的 3D 体素特征使用 3D 卷积网络进行语义特征挖掘。

这一方法的难点在于如何克服 3D 卷积带来的计算资源消耗，我们使用两种策略来缓解这一问题，分别是使用稀疏 3D 卷积网络和特征下采样操作。受到路端摄像机视角范围限制，基于上述特征提升方法生成的 $N_x \times N_y \times N_z$ 个体素特征中存在全零特征，这是由于该部分体素并不在摄像机视野范围内，因此其投影点落在图像之外，由零值填充。待感知区域内的有效特征组成视锥形状，如果直接使用 3D 卷积网络，会带来较大的计算资源浪费。因此我们将 3D 体素特征输入基于子流形卷积核的稀疏 3D 卷积网络层，同时使用下采样层实现尺度缩小，以进一步节省 GPU 显存占用。需要注意的是，我们对占用预测标签进行同步的下采样预处理以适配该特征的尺度。

具体来说，我们首先使用输入层和步长为 1 的 3D 稀疏卷积模块得到原始尺寸的 3D 体素特征图 $V_1 \in \mathbb{R}^{N_x \times N_y \times N_z \times C_{in}}$：

$$V_1 = conv_1(conv_{input}(V_{3D})) \tag{6}$$

然后该特征输入到包含步长为 2 的卷积层的模块中实现下采样操作，并由 Softmax 激活函数输出，得到 3D 体素特征 $V_2 \in \mathbb{R}^{N_x' \times N_y' \times N_z' \times C_{out}}$，其中 $N_x' = \frac{1}{2}N_x$，$N_y' = \frac{1}{2}N_y$，$N_z' = \frac{1}{2}N_z$：

$$V_{occ} = Softmax(conv_2(V_1)) \tag{7}$$

常规的操作将 3D 空间中目标物体的检测任务转换为 BEV 空间中的二维检测，我们遵循 ImvoxelNet[19]网络的结构进行 3D 到 BEV 特征的转换，通过几层 3D 卷积层和下采样操作，在 Z 维度将 3D 体素特征压缩到 2D 的 BEV 特征 $V_{BEV} \in \mathbb{R}^{N_x \times N_y \times C}$，进而基于锚框的目标检测头分别使用 2D 卷积层输出目标类别、候选框以及水平朝向角度。

### 2.3 损失函数

我们提出的路端占用网络使用占用预测损失和目标检测损失同时监督网络学习。其中对于占用预测，我们使用交叉熵损失和 lovasz-softmax 损失，对于目标检测我们使用 PointPillars[20]中的损失函数：

$$\mathcal{L}_{occ} = \mathcal{L}_{ce} + \mathcal{L}_{ls} \tag{8}$$

$$\mathcal{L}_{\text{det}} = \frac{1}{N_{\text{pos}}}\left(\beta_{cls}\mathcal{L}_{cls} + \beta_{loc}\mathcal{L}_{loc} + \beta_{\text{dir}}\mathcal{L}_{\text{dir}}\right) \tag{9}$$

其中其中，$N_{\text{pos}}$ 是正锚框的数量，$\mathcal{L}_{cls}$ 是焦点损失(Focal Loss)，$\mathcal{L}_{loc}$ 为平滑 L1 损失(Smooth-L1 Loss)，$\mathcal{L}_{\text{dir}}$ 是交叉熵损失。

最终的损失为二者之和：

$$\mathcal{L} = \mathcal{L}_{occ} + \mathcal{L}_{det} \tag{10}$$

## 3 实验

### 3.1 实验细节

我们的网络结构以 ImvoxelNet 作为基线，使用特征金字塔网络融合不同尺度的特征，其中第一层的特征维度设置为 256，输出层维度设置为 64，尺寸为原图 4 倍下采样。对于路端占用网络中的 2D 到 3D 体素特征投影模块，我们预先设定体素尺寸长宽高为 $0.32\times0.32\times0.32$ 米，感知区域亦即点云范围设置为 X 轴 $[0m,+69.12m]$，Y 轴 $[-39.68m,+39.68m]$，Z 轴 $[-2.92m,+0.92m]$，因此 3D 体素特征尺寸为 $216\times248\times12$，占用预测尺寸为 $108\times124\times6$。对于目标检测头我们遵循 ImvoxelNet 和 PointPillars 等工作中的设定，损失函数的权重分别为 $\beta_{cls}=1$，$\beta_{loc}=2$，$\beta_{\text{dir}}=0.2$。

在训练过程中，我们使用 Adam 优化器，初始学习率设置为 0.0001，权重衰减为 0.0001。该网络在 2 块 Nvidia 3080 GPU 上训练 12 轮，批处理大小(batch-size)设置为 4，在第 8 和第 11 轮之后，学习率降低 10 倍。由于我们设计的网络涉及 3D 点到 2D 图像平面的投影以及占用预测监督，我们没有使用其他网络常用的数据增强方法。

### 3.2 数据集和评价指标

DAIR-V2X-I 路端 3D 检测数据集是首个同时具备图像和点云 3D 联合标注的大规模路侧多模态数据集，包括路侧 10084 帧图像数据 10084 帧点云数据。其中摄像头和激光雷达被安装在十字路口支架上的相同位置，并进行标定和图像去畸变。按照该数据集基准，DAIR-V2X-I 分别按照 50%，20%，30%的比例划分为训练集、验证集和测试集，由于测试集并未公开，我们在验证集上进行算法评估。该数据集提供的数据标注包括汽车、大货车、厢式货车、公交车、行人、自行车、三轮车、摩托车、手推车等，我们遵循 BEVHeight 将其中的二级类别重新划分，保留汽车、行人、骑行者三种类别。

遵循 ImvoxelNet、PointPillars 等相关工作，我们为汽车、行人和骑行者三种类别训练一个统一的网络，并使用具有 40 个召回采样位置的平均精度(mean Average Precision, mAP)作为目标检测结果的评价指标，其中对于三种类别的 IoU 阈值分别设定为 0.5、0.25、0.25。对

于 3D 语义占用预测，我们遵循常见做法，使用所有类别的占用体素的交并比（mean intersection over union, mIoU）：

$$\mathrm{mIoU}=\frac{1}{\mathrm{C}}\sum_{\mathrm{c}=1}^{\mathrm{C}}\frac{\mathrm{TP_c}}{\mathrm{TP_c}+\mathrm{FP_c}+\mathrm{FN_c}} \tag{11}$$

其中，$\mathrm{TP_c}$，$\mathrm{FP_c}$，$\mathrm{FN_c}$ 分别表示类别 c 的真正例、假正例、假负例的数量，C 表示总类别数。

### 3.3 与 SOTA 比较

在 DAIR-V2X-I 数据集的验证集上，我们比较了我们提出的基于占用预测的路端检测网络与其他前沿相关工作的性能。表 1 显示了汽车、行人、骑行者的平均检测精度，每个类别按照检测难度划分为简单、中等、困难三个等级。与基线工作 ImvoxelNet 相比，我们设计的路端占用网络提高了所有类别的检测精度，达到 41.39%的 3D 平均检测精度，尤其是提高了对行人、骑行者等小目标的检测精度，在中等难度的类别上分布达到 26.55%和 33.9%的 mAP。

表 1　在 DAIR-V2X-I 验证集上 3D 检测结果与 SOTA 方法的比较

Tab. 1　Comparing of 3D detection results on the DAIR-V2X-I val set with the state-of-the-art.

| Method | Modality | Auxiliary Data | Car.(IoU=0.5) | | | Ped.(IoU=0.25) | | | Cyc.(IoU=0.25) | | | mAP(%) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | *Easy* | *Mid.* | *Hard* | *Easy* | *Mid.* | *Hard* | *Easy* | *Mid.* | *Hard* | |
| PointPillars | LiDAR | - | 65.85 | 53.66 | 53.70 | 56.99 | 54.56 | 54.73 | 59.77 | 41.74 | 43.46 | 53.83 |
| ImvoxelNet* | LiDAR | - | 65.66 | 53.55 | 53.61 | 27.15 | 25.78 | 26.04 | 46.48 | 32.56 | 33.98 | 40.53 |
| ImvoxelNet | Camera | - | 63.14 | 53.19 | 53.26 | 21.83 | 20.75 | 21.20 | 44.51 | 30.72 | 32.55 | 37.91 |
| RoadOcc* | Camera | LiDAR | 65.71 | 53.57 | 53.65 | 27.99 | 26.55 | 26.70 | 46.19 | 32.36 | 33.74 | 40.72 |
| RoadOcc | Camera | LiDAR | 65.72 | 53.57 | 53.63 | 27.94 | 26.54 | 26.78 | 48.76 | 33.90 | 35.68 | 41.39 |

### 3.4 消融实验

我们对网络中 2D 体素特征的高度压缩方法进行了消融实验，表 1 中加*的实验为延续 ImvoxelNet 使用 3D 卷积进行高度压缩，而不加*的实验为仅使用 2D 卷积。在基线工作 ImvoxelNet 中，经过投影操作生成的 3D 体素特征 $\mathrm{V_{3D}}\in\mathbb{R}^{N_x\times N_y\times N_z\times C}$ 直接送入 3D 卷积网络挖掘几何空间特征，并通过其中的下采样层实现 Z 轴维度的压缩，输出为 $\mathrm{V}_{BEV}\in\mathbb{R}^{N_x\times N_y\times 1\times C}$ 的二维 BEV 特征。然而 3D 卷积层的引入造成了较多的计算资源占用，考虑到本方法已经引入了占用预测损失辅助监督网络同时学习语义特征和几何特征，我们进一步研究采用更加简单高效的高度压缩方法，并基于 2D 卷积网络进行 BEV 特征提取。具体来说，我们沿用 Fast-BEV 等相关工作中的“空间到通道”操作(Space-to-Channel, S2C)对 3D 体素特征进行降维压缩，通过对 4D 张量 $\mathrm{V_{3D}}$ 进行变形转换为 $\mathrm{V}_{BEV}\in\mathbb{R}^{N_x\times N_y\times(Z\times C)}$ 的 3D 张量，减少空间维度而增加了特征通道数，避免了引入复杂的 3D 卷积层，另外，我们在 BEV 骨干网络中使用多尺度融合结构，提高了网络对不同尺寸目标物体的适应性。

实验显示，由于 ImvoxelNet 依赖 3D 卷积网络中的 3D 卷积核提取空间几何信息，在更换为 S2C 操作后，3D 检测结果有较大程度的下降。然而我们提出的路端占用网络依然能够获得较高的检测精度，这验证了使用 3D 语义占用预测增强体素特征张量中的几何特征的有效性。

表 2　损失权重的消融实验

Tab. 2 Ablation experiment of weights of loss.

| Occ. Loss Weight | Det. Loss Weight | Car.(IoU=0.5) | Ped.(IoU=0.25) | Cyc.(IoU=0.25) | mAP(%) | mIoU(%) |
|---|---|---|---|---|---|---|
| 0.2 | 0.8 | 57.64 | 27.09 | 39.45 | 41.39 | 45.29 |
| 0.4 | 0.6 | 57.63 | 24.51 | 37.34 | 39.83 | 44.42 |
| 0.5 | 0.5 | 57.56 | 25.93 | 36.62 | 40.04 | 44.22 |
| 0.6 | 0.4 | 57.60 | 23.32 | 35.51 | 38.81 | 43.35 |
| 0.8 | 0.2 | 57.51 | 20.57 | 35.56 | 37.88 | 43.27 |

我们对占用预测和目标检测的损失权重进行了消融实验，如表 2 所示。占用预测任务面临的一个主要困难是类别的不均衡问题。路端交通场景的感知面积较大，在本实验中达到 $69.12\times79.36$ 米，因此中存在大量的空体素以及背景类体素，而汽车、行人等目标物体由于尺寸较小，其所占用的体素也较少。另外，真实环境中汽车数量居多，因此数据集中汽车类别的体素远远超过行人以及骑行者，严重的类别不均衡问题造成了网络优化的困难。本网络通过设置较大的路端占用预测损失权重，使整体网络学习的重点放在对场景中目标物体的体素占用细节的学习，因此该网络能够增强小尺寸目标的检测精度。

## 4 结论

本文提出了一种针对路端感知场景的基于占用预测的单目 3D 目标检测算法。该算法能够利用路端视角遮挡情况较少的优势，通过稀疏卷积网络高效地生成路端占用预测结果，并为目标检测提供包含丰富几何和语义信息的体素特征。在真实世界数据集 DAIR-V2X-I 上的实验结果表明，本文所提出的路端占用网络能够高效地实现对复杂交通路口中感兴趣目标的检测，尤其是可以提高对行人、骑行者等小目标的检测精度。

**[参考文献] (References)**

[1] Philion J, Fidler S. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d[A]. Proceedings of the Europeon Conference on Computer Vision[C]. Glasgow: Springer International Publishing, 2020. 194-210.

[2] Harley A W, Fang Z, Li J, et al. Simple-bev: What really matters for multi-sensor bev perception?[A]. IEEE International Conference on Robotics and Automation [C]. London: IEEE, 2023. 2759-2765.

[3] Yu H, Luo Y, Shu M, et al. Dair-v2x: A large-scale dataset for vehicle-infrastructure cooperative 3d object detection[A]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition[C]. New Orleans: IEEE, 2022. 21361-21370.

[4] Chen X, Kundu K, Zhu Y, et al. 3d object proposals for accurate object class detection[J]. Advances in neural information processing systems, 2015, 28.

[5] Chen X, Kundu K, Zhang Z, et al. Monocular 3d object detection for autonomous driving[A]. Proceedings of the IEEE conference on Computer Vision and Pattern Recognition[C]. Las Vegas: IEEE, 2016. 2147-2156.

[6] Duan K, Bai S, Xie L, et al. Centernet: Keypoint triplets for object detection[A]. Proceedings of the IEEE/CVF International Conference on Computer Vision[C]. Seoul: IEEE, 2019. 6569-6578.

[7] Wang Y, Chao W L, Garg D, et al. Pseudo-lidar from visual depth estimation: Bridging the gap in 3d object detection for autonomous driving[A]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition[C].Long Beach: IEEE, 2019. 8445-8453.

[8] Roddick T, Kendall A, Cipolla R. Orthographic feature transform for monocular 3d object detection[J]. arXiv preprint arXiv:1811.08188, 2018.

[9] Cao A Q, de Charette R. Monoscene: Monocular 3d semantic scene completion[A]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition[C]. Piscataway, NJ: IEEE, 2022. 3991-4001.

[10] Huang Y, Zheng W, Zhang Y, et al. Tri-perspective view for vision-based 3d semantic occupancy prediction[A]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition[C]. Piscataway, NJ: IEEE, 2023. 9223-9232.

[11] Zhang Y, Zhu Z, Du D. OccFormer: Dual-path Transformer for Vision-based 3D Semantic Occupancy Prediction[J]. arXiv preprint arXiv:2304.05316, 2023.

[12] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.
[13] Wei Y, Zhao L, Zheng W, et al. Surroundocc: Multi-camera 3d occupancy prediction for autonomous driving[A]. Proceedings of the IEEE/CVF International Conference on Computer Vision[C]. Piscataway, NJ: IEEE, 2023. 21729-21740.
[14] Wang X, Zhu Z, Xu W, et al. Openoccupancy: A large scale benchmark for surrounding semantic occupancy perception[J]. arXiv preprint arXiv:2303.03991, 2023.
[15] Behley J, Garbade M, Milioto A, et al. Semantickitti: A dataset for semantic scene understanding of lidar sequences[A]. Proceedings of the IEEE/CVF international conference on computer vision[C]. Piscataway, NJ: IEEE, 2019. 9297-9307.
[16] Reading C, Harakeh A, Chae J, et al. Categorical depth distribution network for monocular 3d object detection[A]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition[C]. Piscataway, NJ: IEEE, 2021. 8555-8564.
[17] Yang L, Yu K, Tang T, et al. BEVHeight: A Robust Framework for Vision-based Roadside 3D Object Detection[A]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition[C]. Piscataway, NJ: IEEE, 2023. 21611-21620.
[18] Li Y, Huang B, Chen Z, et al. Fast-BEV: A Fast and Strong Bird's-Eye View Perception Baseline[J]. arXiv preprint arXiv:2301.12511, 2023.
[19] Rukhovich D, Vorontsova A, Konushin A. Imvoxelnet: Image to voxels projection for monocular and multi-view general-purpose 3d object detection[A]. Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision[C]. Piscataway, NJ: IEEE, 2022. 2397-2406.
[20] Lang A H, Vora S, Caesar H, et al. Pointpillars: Fast encoders for object detection from point clouds[A]. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition[C]. Piscataway, NJ: IEEE, 2019. 12697-12705.