（论文解析2）SurroundOcc: Multi-Camera 3D Occupancy Prediction for Autonomous Driving

目录

背景

不同于基于深度的方法，3D重建的另一种主流是直接预测3D占用率。

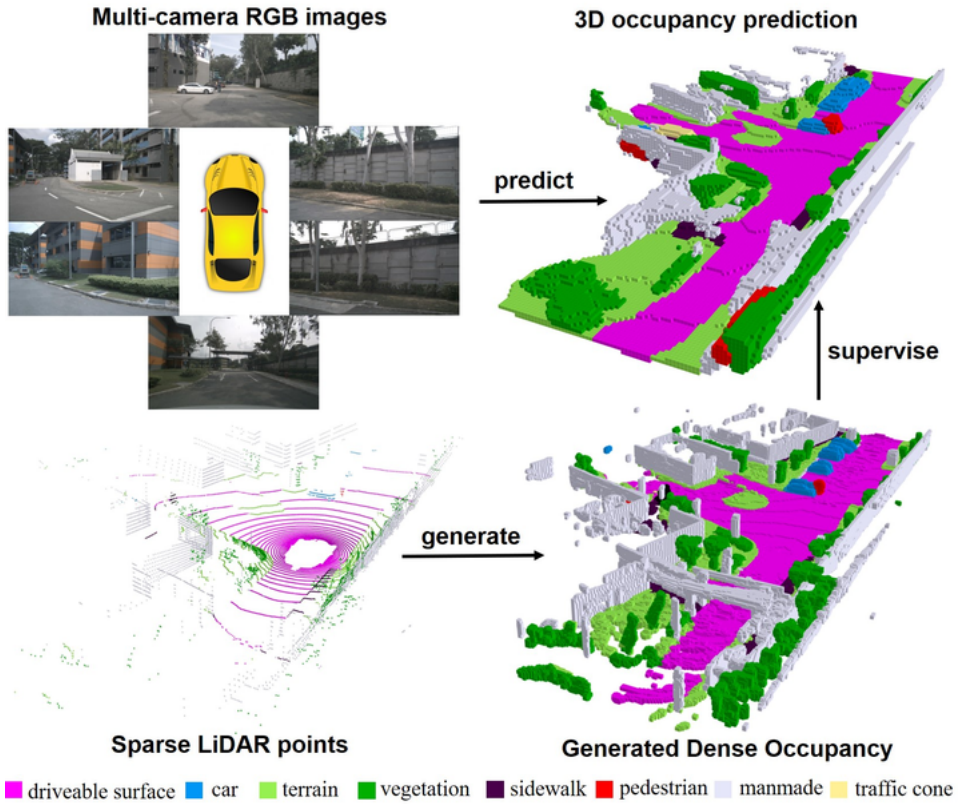
以往的occ模型都存在一些缺陷：

- MonoScene：由于单眼分析，跨相机后处理融合的方法导致了差的性能；
- TPVFormer：虽然使用的多路摄像机，但是使用稀疏点云导致occ结果也很稀疏；（有文章说本论文是该网络的延伸，已读）
  - 设计了一种TPV。相对于BEV（纯隐式保留），创造了三个平面（相互正交的俯视图、前视图以及侧视图），没有过度增大计算资源的同时，在一定程度上保留了更多视角的（包括深度）信息；相对于3维体素（显示保留），也极大地缩小计算成本；

因此，作者提出了SurroundOcc，ICCV 2023，作品链接：https://arxiv.org/abs/2303.09551

github：https://github.com/weiyithu/SurroundOcc

框架/流程

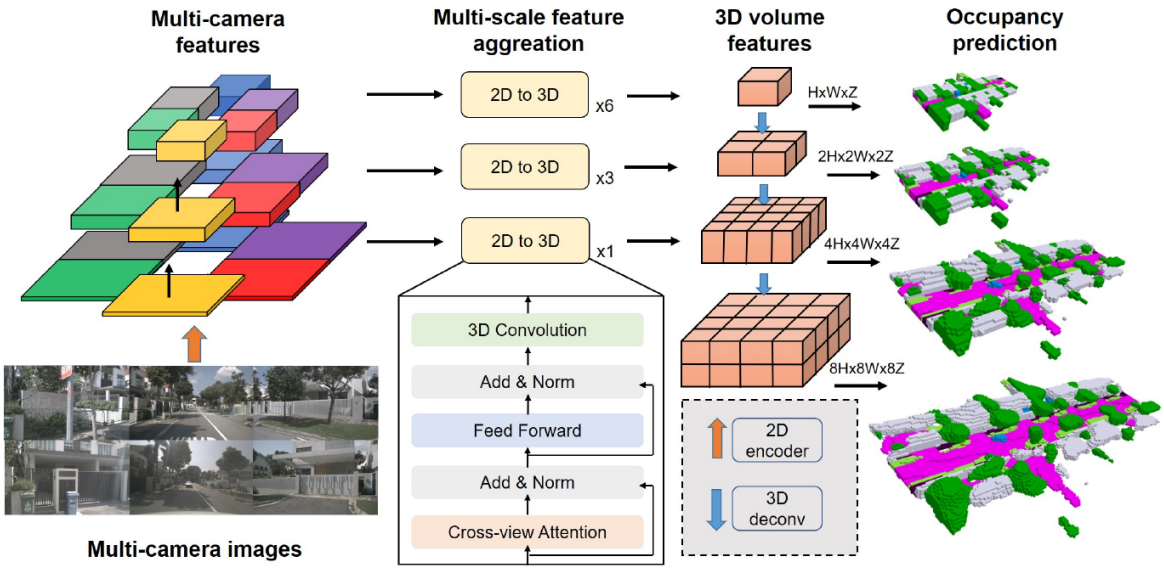


总的框架图；

模型输入：六路RGB图像　　模型输出：3D占用概率预测；

标签生成算法输入：稀疏点云图　　静态场景和动态物体独立处理；　输出：稠密占用图；

模型



1. 首先利用ResNet-101等backbone对多相机图像独立提取特征，并利用FPN对二维图像特征进行多尺度融合。
2. 之后在每个尺度上，均利用 spatial cross attention 对多相机特征进行提取，实现2D到3D的转换，得到三维体素特征。
3. 最后利用三维卷积网络对多尺度体素特征上采样和组合。

4. 不同尺度的特征会输出不同分辨率的三维占据预测，并利用 loss 权重衰减的稠密标签进行监督。

稠密标签生成算法

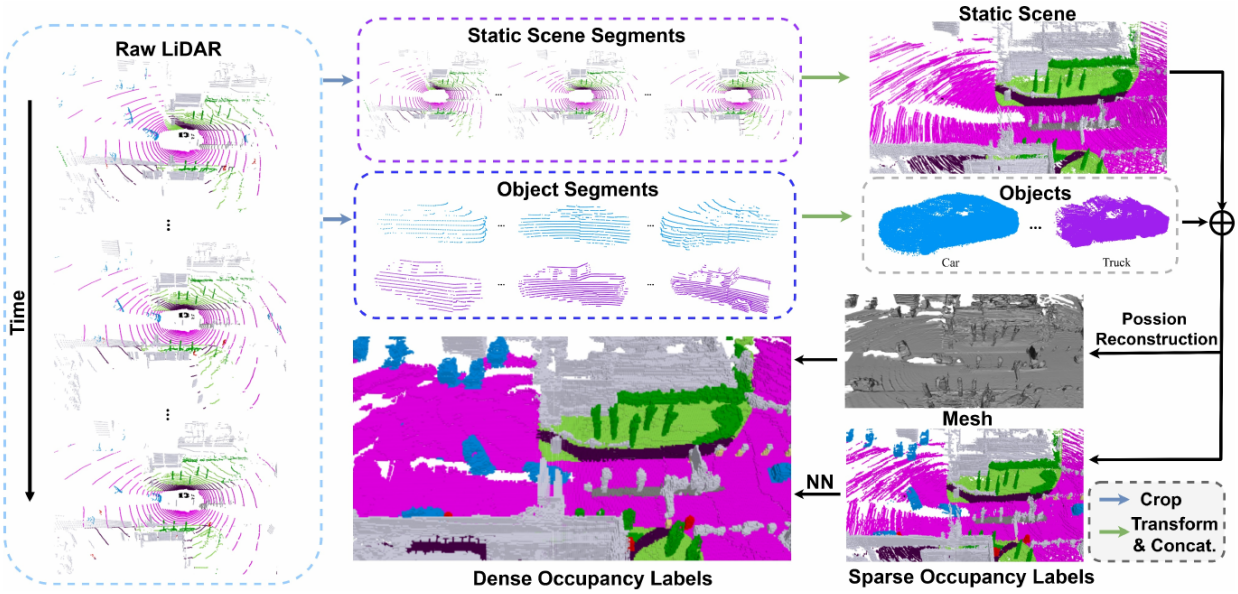


Figure 4. Dense occupancy ground truth generation. We first traverse all frames to stitch the multi-frame LiDAR points of dynamic objects and static scenes separately, and then merge them into a complete scene. Subsequently, we employ Poisson Reconstruction to densify the points and voxelize the resulting mesh to obtain a dense 3D occupancy. Finally, we use the Nearest Neighbor (NN) algorithm to assign semantic labels to dense voxels.

1. 根据目标检测label将点云分为static和dynamic，并分别累计**多帧静态场景（Scene）点云和动态物体（Object）点云**，然后利用位姿信息将场景和物体点云拼接起来。
2. 累计的多帧点云还存在hole，因此用**泊松重建**的方法得到更稠密的volumetric occupancy label。
3. 没有语义信息的grid选择距离最近的带有语义标注的grid作为自己的semantic label。（**最近邻思想**）

代码研究

Todo。。

ref

【1】 【论文阅读】SurroundOcc：面向自动驾驶的多相机3D占用预测

【2】 Occupancy感知算法如何获取数据标签？如何编码？任何解码？_自动驾驶之心的博客-CSDN博客