



# panoptic - fusionnet：基于摄像头-激光雷达融合的自动驾驶点云全景分割

Hamin Song <sup>a</sup>, Jieun Cho <sup>a</sup>, Jinsu Ha <sup>a</sup>, Jaehyun Park <sup>b</sup>, Kichun Jo <sup>b,\*</sup>

<sup>a</sup>建国大学智能汽车工程系,韩国首尔05029 <sup>b</sup>汉阳大学汽车工程系,韩国首尔

04763

## 条信息

关键词:激光雷达感知全视分割传感器融合特征图融合智能汽车

## 摘要

准确可靠的感知对于自动驾驶汽车的安全运行至关重要。光探测和测距(激光雷达)全光学分割任务在预测3D环境中的点级类别和实例id方面起着至关重要的作用，有助于全面了解车辆周围环境。现有的依赖于单个激光雷达传感器的方法经常面临诸如点稀疏和缺乏纹理信息等挑战所带来的限制。为了克服这些挑战，可以采用其他传感器(如相机)的融合。然而，利用传感器融合方法开发全光学分割模型仍然是一个未开发的领域。在本文中，我们提出了第一个用于全景分割的激光雷达-相机融合网络，名为“panoptic - fusionnet”。我们提出的网络通过一个特征融合模块来增强特征地图，该模块确保从两个传感器中提取的特征的精确几何对齐。我们创建了在多个尺度上具有精确对齐的点体素-像素对应关系的表，并通过查询表来融合特征图。融合的特征随后被用于自下而上的网络来预测语义标签、实例中心和偏移量。然后以类自适应的方式实现后处理，基于每个类的一般大小实现类智能聚类，从而改进最终的全景结果。我们评估了所提出的网络在SemanticKITTI和nuScenes数据集上的性能，这些数据集是提供同步激光雷达-相机数据的大型数据集。与仅使用lidar的基线相比，panoptic - fusionnet由于其几何精确的传感器融合方法、类别自适应处理和利用实例增强预训练权重，获得了更高的全光质量。

## 1. 介绍

对于智能车辆的可靠运行，感知周围环境至关重要。为了实现这一点，使用了摄像头、激光雷达、雷达等各种传感器。其中，激光雷达因其优越的几何精度和在不同照明条件下的一致性能而得到特别利用。激光雷达提供精确的周围环境三维(3D)数据。然而，需要补充有意义的信息，如物体类别，才能全面理解场景。

将有意义的信息分配给点云，目前正在研究使用深度学习方法来完成各种任务。首先，点云语义分割预测逐点标签，通过为点云中的每个点分配类来提供非常详细的分析。图1 (a)说明了语义分割的标签，包括汽车和骑自行车的人等对象以及建

筑物和道路等背景元素。然而，它无法区分相同类别的重叠对象。也就是说，它没有捕捉到单个对象的边界。其次，点云实例分割预测逐点类标签及其实例id，使其能够处理具有重叠或遮挡实例的复杂场景。然而，如图1 (b)所示，实例分割只关注前景感兴趣的点，如汽车和骑自行车的人，从而忽略了背景点，如停车场或建筑物。虽然这两项任务提供了详细的场景理解，但它们无法区分实例或倾向于忽略背景。为了克服这些问题，提出了一种将两者统一起来的新任务，即全视分割。

全视分割(Kirillov, He, Girshick, Rother, & Dollár, 2019)是一项集成了语义分割和实例分割概念的高级任务。该任务预测了两个主要的

\*Corresponding author.

E-mail addresses: [shm5069@konkuk.ac.kr](mailto:shm5069@konkuk.ac.kr) (H. Song), [dazzkaki@konkuk.ac.kr](mailto:dazzkaki@konkuk.ac.kr) (J. Cho), [soo4826@konkuk.ac.kr](mailto:soo4826@konkuk.ac.kr) (J. Ha), [chang8224@hanyang.ac.kr](mailto:chang8224@hanyang.ac.kr) (J. Park), [kichunjo@hanyang.ac.kr](mailto:kichunjo@hanyang.ac.kr) (K. Jo).

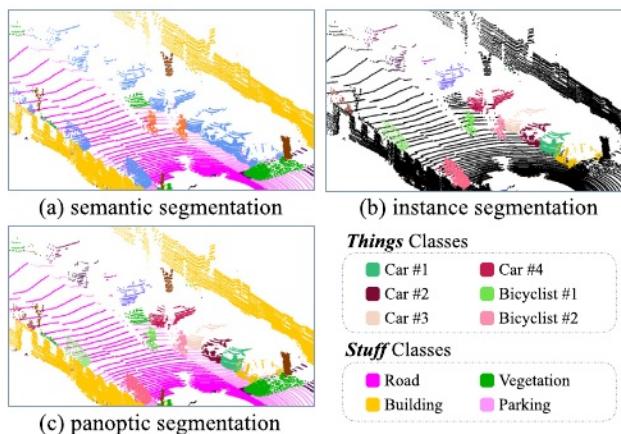


图1所示。(a)语义分割、(b)实例分割和(c)全视分割的基础真值说明。

类类型：(1)事物，(2)物质。事物表示感兴趣的可数对象，例如行人和汽车，与实例分割中的前景对象相同。*Stuff*表示背景元素，如建筑物和道路，这些在实例分割中被忽略。因此，全景分割不仅可以预测逐点语义类，还可以预测属于同一对象的唯一实例id，从而实现对整个场景的全面分析。图1 (c)很容易解释为什么全景分割是语义和实例分割的统一任务。

然而，仅基于单个激光雷达执行诸如全景分割之类的感知任务存在一些挑战。首先是点稀疏性，由于光束之间的距离随着与传感器的距离增加而增加，从而减少了从远处或小物体反射的点的数量。第二个挑战是缺乏纹理信息，使得区分形状相似的物体变得困难。预计这些挑战将降低仅使用激光雷达的全景分割性能。传感器与其他传感器的融合可以解决这些问题，相机是一个理想的候选者，因为它们能够用更密集的像素信息克服稀疏性，并通过RGB颜色提供纹理信息。将激光雷达与相机融合，可以潜在地提高远距离或小尺寸物体在全景分割中的性能，并区分形状相似的物体。然而，据我们所知，迄今为止还没有进行传感器融合以提高点云全景分割性能的研究。

本文提出了一种利用相机-激光雷达传感器融合的点云全景分割网络。这标志着基于相机-激光雷达融合的全景分割网络的开始。该网络的主要目标是通过解决使用单个激光雷达时出现的限制来提高全景分割的性能。在传感器融合策略方面，我们引入了基于点-体素-像素对应表的特征融合模块。与大多数以前的相机-激光雷达融合方法不同，这些方法将激光雷达数据投影到2D以方便特征地图匹配，我们直接利用原生3D形式的激光雷达数据。这种方法的基本原理在3.1节中进行了解释，其中我们保证其一致的几何精度。为了在模型的任何中间阶段精确对齐3D激光雷达特征与2D相机特征，我们建立了不同尺度的点体素-像素对应表，并通过查询和匹配中间层的特征进行特征融合。通过特征融合，它利用密集相机特征的辅助来补充激光雷达特征的稀疏属性，从而能够生成以更高维度的方式代表周围环境的上下文信息。关于所提出模型的设计，该架构由一个骨干网络、一个特征融合模块、一个语义头、一个实例头和一个全景头组成。为

了从每个传感器数据中提取特征，采用了不同的骨干网。然后在多个尺度上应用特征融合模块来聚合相机和激光雷达特征，生成增强的融合特征。随后利用融合特征来预测语义分割的结果以及每个实例的中心和偏移量。最后一步合并这些结果以生成逐点的类和实例id。为了实现合理的聚类，这个过程以类自适应的方式实现。

这种架构形成了一个高效的基于相机-激光雷达融合的全景分割网络，我们将其命名为panoptic - fusionnet。为了评估我们的panoptic - fusionnet，我们在SemanticKITTI (Behley 等人, 2019) 和 nuScenes (Caesar等人, 2020) 数据集上进行了实验。与纯激光雷达全景分割模型相比，本文提出的网络具有更高的性能，验证了其优越性。

我们的主要贡献可以概括如下：

- 我们提出panoptic - fusionnet，据我们所知，这是首个实现相机-激光雷达融合方法的点云全景分割任务。
- 我们引入了一个特征融合模块，确保几何上精确的特征对齐。我们建立点-体素-像素对应表，并在模型的中间层进行灵活的表查询，用于多尺度特征图融合。
- 我们在SemanticKITTI和nuScenes数据集上进行实验，验证panoptic - fusionnet，并将其性能与激光雷达全景分割方法进行比较，以评估融合方法的优越性。

本文包括以下内容：第2节回顾了前人关于panoptic分割和传感器融合方法的研究成果。在第3节中，我们解释了网络架构和网络训练策略的细节。然后，第4节给出实验结果，第5节为结论。第6节是缩略语和符号列表。

## 2. 相关工作

本节简要回顾了与我们提出的方法相关的先前研究。首先，我们调查了执行全景分割的模型，将它们分为仅相机方法和仅激光雷达方法。此外，我们回顾了相机-激光雷达融合被用于除全景分割以外的任务的案例，因为没有全景分割网络包含相机-激光雷达融合。

### 2.1. Panoptic segmentation: Camera-only methods

全景分割是计算机视觉领域对图像的初步研究。模型预测每个像素的语义标签和实例id。这些网络通常有三个组成部分：用于特征提取的主干，语义和实例头部，以及全景处理模块。

UPSN (Xiong et., 2019) 利用基于ResNet (He, Zhang, 任, 孙, 2016) 的共享骨干网进行多尺度特征提取。这些特征通过使用可变形卷积的语义头进行处理。基于Mask R-CNN (He, Gkioxari, Dollár, & Girshick, 2017) 的实例头预测掩码分割和边界框。panoptic head结合输出logits产生panoptic logit张量，通过softmax操作实现逐像素的类和实例ID预测。

EfficientPS (Mohan和Valada, 2021年) 采用改进的高效网(Tan和Le, 2019年)作为共享主干。语义头包括三个模块：用于捕获精细特征的大规模特征提取器，用于远程上下文的密集预测单元，以及用于减轻特征尺度差异的错配校正模块。实例头应用Mask R-CNN。在全景融合模块中，

使用sigmoid和Hadamard运算融合语义logit和边界框logit，从而预测最终结果。

Panoptic-DeepLab (Cheng et al., 2020)利用共享主干将最终编码器块拆分为语义和实例上下文模块。每个都使用单独的空间金字塔池(ASPP) (Chen, Papandreou, Kokkinos, Murphy, & Yuille, 2017)进行上下文提取。语义头采用DeepLab结构，实例头以类不可知论的方式操作，通过物体的质心来表示物体，以热图的形式预测。panoptic处理模块使用中心和偏移量对实例进行聚类，并将它们与语义逻辑融合。

axial - deeplab (Wang et al., 2020)是Panoptic-DeepLab的开发版本，集成了一个高效的注意力模块。axial - deeplab用轴向注意力代替ASPP。这是基于自注意力的，它通过在高度和宽度方向上沿一维轴施加注意力来使用低复杂度模块来增强性能。

## 2.2. 全视分割:仅激光雷达方法

随着技术的进步，现在不仅可以对图像进行全视分割，还可以对点云进行全视分割。考虑到仅激光雷达网络处理3D数据，点云的表示方法很重要。

EfficientLPS (Sirohi, Mohan, bscher, Burgard, & Valada, 2021)将点云编码为范围图像，并采用类似于EfficientPS的网络结构。它利用effenet拓扑作为共享骨干，并应用接近卷积模块来利用距离图像的深度特征。在语义头中，通过距离引导深度方向的亚特鲁斯可分离卷积来预测类。实例和全光融合模块分别采用Mask R-CNN和EfficientPS的方法。

Panoptic-PolarNet (Zhou, Zhang, & Foroosh, 2021)使用点云编码为鸟瞰(BEV)图像。它采用PolarNet (Zhang等人, 2020)作为共享骨干网，覆盖解码器层，而不仅仅是编码器部分，从而减少了计算需求。结构遵循Panoptic-DeepLab；语义头与PolarNet相同，实例头通过修改的PolarNet层预测BEV图像中的实例中心和偏移量。panoptic处理模块使用带有语义预测的预测中心和偏移量热图对实例进行聚类。

DS-Net (Hong, Zhou, Zhu, Li, & Liu, 2021)采用圆柱划分(Zhou等, 2020)从点云中高效提取3D体素特征。它将多层感知器(MLP)与语义头部的圆柱形卷积集成在一起，用于语义预测。实例头预测偏移向量，对事物执行中心回归。当集群实例时，DS-Net采用动态移位模块。它通过学习动态调整内核大小，处理不同的实例大小。

panopic - phnet (Li, He等, 2022)使用体素和BEV格式对点云进行编码。最初，体素编码提取3D特征，然后将其表示为2D BEV特征。语义头预测逐点类，而实例头利用knn-transformer模块对体素交互进行建模。它预测实例偏移量，并相应地替换与事物相关的体素，为实例中心生成基于密度的伪热图。最终结果是通过投票方法生成的。

## 2.3. 基于深度学习的融合方法

已经积极开展研究，通过深度学习，通过传感器融合来提高单一LiDAR使用之外的性能。由于没有传感器融合网络专门用于全光分割任务，我们回顾了在物体检测和语义分割中应用传感器融合的网络。

3D-CVF (Yoo, Kim, Kim, & Choi, 2020)是一种目标检测网络，通过单独的主干从相机和激光雷达数据中提取特征。通过自动标定特征投影，将相机特征转换为激光雷达BEV域的空间特征图。自适应门控融合网络识别相机数据增强Li-DAR的区域，生成引导特征融合的空间注意图。融合后的特征输入到区域建议网络中，从而得到最终的检测结果。

DeepFusion (Li, Yu et al., 2022)是一种目标检测网络，专注于在融合过程中对齐激光雷达和相机特征。它引入了两个模块：InverseA ug，它可以反转几何相关的增强以关联传感器数据；LearnableAlign，它使用交叉注意力来理解激光雷达和相机特征之间的相关性。由提出的模块导出的融合特征馈送到3D检测框架中，用于增强预测。

PMF (Zhuang et al., 2021)以语义分割为目标，旨在收敛图像和点云的感知信息。它利用具有传感器特定编码器-解码器结构的双流网络，通过基于残差的融合模块生成融合特征。感知损失测量模态之间的预测方差，使用可靠的预测来监督那些置信度较低的预测。

语义分割网络liff - seg (Zhao et al., 2021)采用多阶段融合。它集成了图像区域和激光雷达数据进行早期融合，并通过圆柱体3D提取3D特征。同时，图像骨干通过偏移校正提取与3D特征融合的2D特征，以解决同步问题。最后，在特征融合过程中，偏移学习阶段预测投影点和像素之间的偏移量。

此外，各种有效的传感器融合方法如(Bai et al., 2022; El Madawi et al., 2019; Pang, Morris, & Radha, 2020; Yan et al., 2022)已经开发出来，提高了跨多任务的性能。然而，目前还没有利用相机和激光雷达融合的全光学分割模型。

## 3. 模型

在这项工作中，我们提出了Panoptic-FusionNet，其概述如图2所示。该网络架构由五个主要组件组成：(1)骨干网，(2)特征融合模块，(3)语义头，(4)实例头，(5)Panoptic处理模块。特征分别通过它们的骨干网从输入点云和图像中提取。在特征提取过程中，在多尺度上将确保精确几何对齐的特征融合模块应用于每个传感器骨干的编码块，从而得到增强的融合特征。之后，通过共享解码器对这些特征进行处理，随后并行的语义和实例头预测语义信息、每个实例的中心以及对其相应中心的偏移量。最后，panoptic head通过类自适应聚类(class-adaptive clustering)生成具有语义类的逐点实例id，该聚类会考虑每个类的一般大小以进行适当的实例分组。

### 3.1. 骨干网络

点云网络有各种基于点云表示的框架：基于2d的框架，利用球面投影距离图像(Milioto, Vizzo, Behley, & Stachniss, 2019, Wang, Shi, Yun, Tai, & Liu, 2018)或BEV图像 (Lin & Wang, 2022; Mohapatra, Yogamani, Gotzig, Milz, & Mader, 2021)，采用体素等3D分区的基于3D的框架(Ciçek, Abdulkadir, Lienkamp, Brox, & Ronneberger, 2016; Zhou & Tuzel, 2018)，以及处理原始点云的基于点的mlp框架(Qi, Yi, Su, & gu, 2017, Hu et al., 2020)。为了在图像和点云之间实现精确的特征对齐，选择具有适当框架的点云骨干网络至关重要。以前大多数执行相机-激光雷达融合的工作都使用基于2d的框架，并且

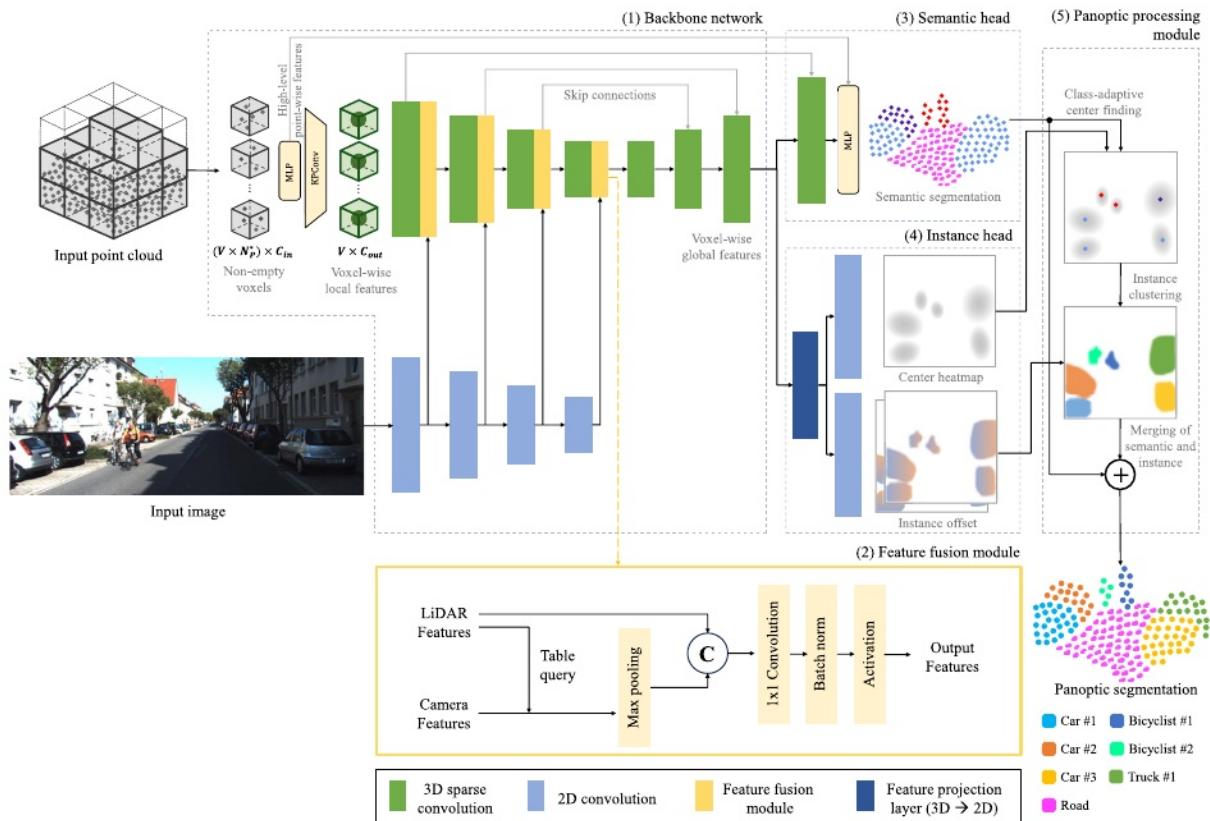


图2所示。整体架构由五个模块组成，具体执行如下：在(1)骨干网中提取每个传感器数据的特征。在每个中间编码层，应用(2)个特征融合模块来组合提取的特征，增强它们的上下文理解。然后将这些增强的特征输入到(3)语义头和(4)实例头中，分别预测语义分割和实例中心和偏移量。最后，在(5)全光处理模块中生成最终的全光分割结果。

该框架的优点是与图像特征映射的对应很简单。然而，在传感器融合中使用点云作为2D数据可能会导致特定情况下的极几何问题，如图3所示。在LiDAR二维平面和相机二维平面上分别显示了对齐的不同区域的LiDAR、Cin三维世界。虽然这些物体被投射到相机平面上的三个不同的点上，但它们在激光雷达二维平面上被投射到一个点上。因此，这种对应关系使确定相机2D平面上的哪个像素对应于激光雷达2D平面上的哪个像素变得复杂，可能导致传感器特征的不准确对齐。因此，为了避免极外几何问题，建议选择没有进行2D预处理的基于3d的框架作为点云骨干网络。以前的传感器融合方法并没有引起对epipolar几何问题的关注，这一发现突出了未来传感器融合方法的一个关键考虑因素。

我们采用PCSCNet (Park, Kim, Kim, & Jo, 2023)作为我们的点云骨干网络，设计用于基于3D体素划分的语义分割。PCSCNet在输入点云上利用体素化，并应用MLP层进行高级点向特征提取，并应用KPConv (Thomas et al., 2019)模块(一种点卷积形式)提取体素方向的局部特征。提取的特征通过采用3D稀疏卷积的编码器-解码器结构，从而得到体素级全局特征。随后，将另一个MLP层应用于解码器的输出，以及高级逐点特征，以预测每个点的类。PCSCNet通过仅利用非空体素进行操作来实现计算效率。此外，其对体素分辨率变化的鲁棒性能源于其通过将精细的逐点特征与解码器的粗全局特征集成来生成高维特征的能力。

此外，ResNet34 (He et al., 2016)架构被用作图像的骨干网络。我们的目标是提取密集的图像特征，并利用它们来增强点云的稀疏特征。具体来说，提取的整个2D特征集不被利用；而是采用通过体素查询确定的特定区域，详见下一节。我们不使用整个图像特征，因此我们更倾向于应用传统骨干网的简单版本ReseNet34，而不是其他高级网络。

### 3.2. 特征融合模块

为了在特征融合过程中充分利用相机和激光雷达数据的优势，必须准确对齐从每个传感器数据中提取的特征。实现这种对齐需要一个几何对应设置，以确保跨多尺度特征的准确性。如果不满足这一要求，就可能导致对融合特征的解释出现混乱，并导致性能下降。因此，我们引入了一个特征融合模块，旨在实现各种尺度上特征映射的精确对齐。在本节中，我们解释了(1)确保精确的几何对齐和(2)随后融合匹配特征的方法。

首先，我们讨论了确保跨多尺度特征的精确几何对应的方法。有必要描述相机与激光雷达之间的空间关系。这可以通过外部校准参数内的变换和旋转来封装。通过利用外部标定参数和摄像机的投影矩阵，可以获得每个3D点对应的图像平面上的具体坐标，满足精确的点对像素对齐。这可以用下面的公式推导出来

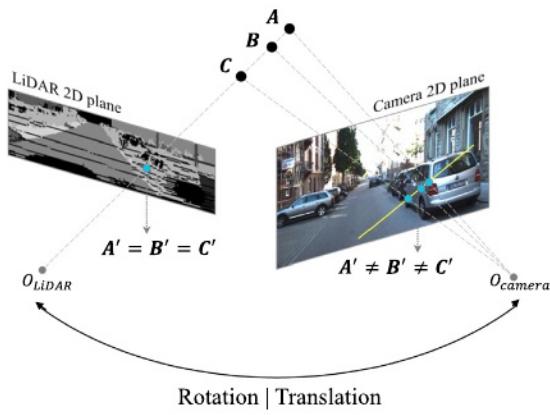


图3所示。使用2d投影格式的激光雷达时的极极几何挑战。

图4所示。使用对应表的体素到像素特征对齐。

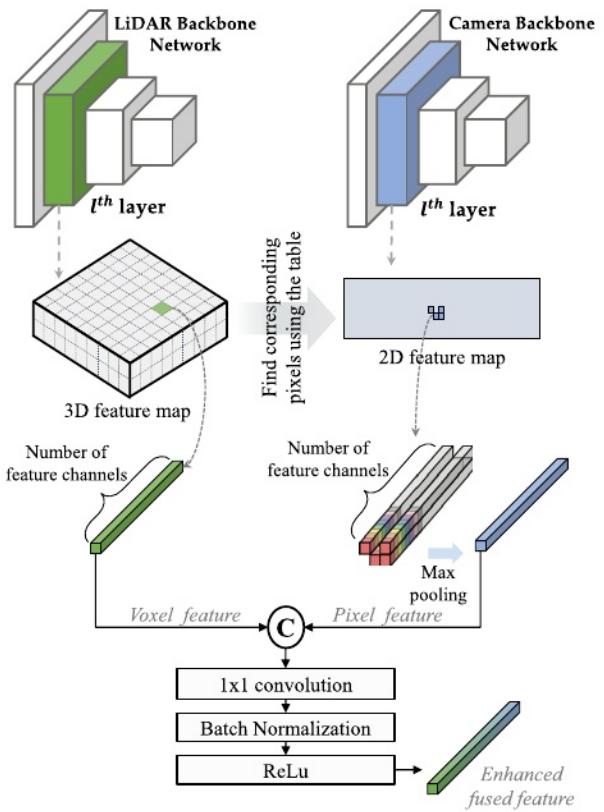
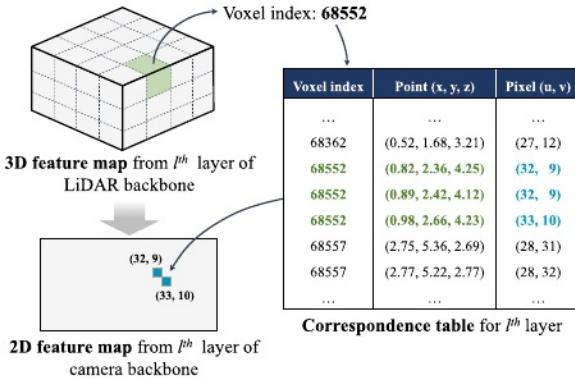


图5所示。特征融合模块，集成了用于增强融合特征的最大值像素和体素特征。

方程：

$$\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{pmatrix} f_u & 0 & c_u \\ 0 & f_v & c_v \\ 0 & 0 & 1 \end{pmatrix} [R_{LiDAR}^{camera} \quad T_{LiDAR}^{camera}] \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix} \quad (1)$$

$$\text{焦距 } F, \text{ 主点 } C, \quad (2)$$

$$\text{Rotation matrix } R_{LiDAR}^{camera} \in \mathbb{R}^{3 \times 3}, \quad (3)$$

$$\text{Translation matrix } T_{LiDAR}^{camera} \in \mathbb{R}^{3 \times 1} \quad (4)$$

当考虑到PCSCNet中体素化点云的利用时，特征融合期间的对应设置需要从点对像素的对齐扩展到包括体素对像素的对齐。因此，我们构建了一个对应表，如图4所示，其中包括3D点和像素之间的对应数据，并具有这些点所属的体素索引。这个表的使用过程如下：识别提取的体素特征的体素索引，并使用索引作为键进行查询。然后，识别共享相同体素索引的点，并获得这些点在图像平面上投影到的像素坐标。通过对应关系，便于从体素到像素的平滑过渡。此外，必须将此表应用于骨干网每个编码块内提取的多尺度特征图。因此，根据特征提取的缩小比例，生成四个版本的表，每个编码块对应一个版本。这个表可以搜索任何体素内的所有原始点，而不考虑特征图的大小，以及随后识别与这些点相连的像素坐标。这确保了体素和像素之间精确的几何对齐。

在表查询过程中，如图4所示，经常会出现单个体素对应多个像素区域的情况。从多个特征数组中选择有用的特征是有利的。在特征图中，可以理解为将更重要的语义信息压缩成更高的值。因此，在使用表提取多个像素的查询坐标对应的图像特征后，我们选择值最大的特征，如图5所示。

其次，我们通过揭示两个传感器的对应关系来解释融合特征的方法。对于特征融合的最后一步，必须将在相应区域内选择的最大值像素特征与体素特征进行集成，以生成增强的融合特征。特征融合模块的流程如图5所示，包括特征对齐。体素特征和max-pooled像素特征最初进行拼接。随后，使用 $1 \times 1$ 卷积操作恢复原始特征深度。最后，依次应用批处理归一化和ReLU函数。来自该模块的输出特征是一个增强的融合特征，取代点云骨干网内后续层的输入。这意味着通过融合模块结合来自图像特征的上下文信息来增强LiDAR特征。

### 3.3. 语义的头

Panoptic-FusionNet的结构包括一个用于逐点类预测的语义头和一个用于与类无关的实例识别的实例头，它们在骨干网的末端并行排列。该设计遵循自下而上的panoptic分割方法的流行结构；自下而上的方法通常是用没有对象检测器的语义分割网络生成实例信息，并通过分组的方式进行处理。

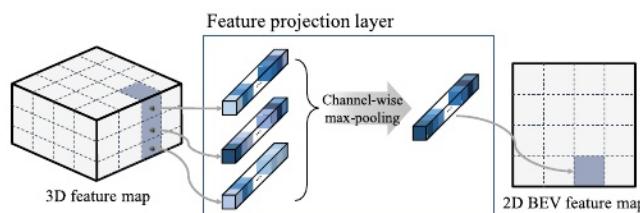


图6所示。特征投影层，对具有相同x和y指数的体素使用通道最大池化将3D特征映射投影到2D BEV特征映射。

语义头采用PCSCNet的最终解码块，以及逐点分类层。由于PCSCNet的所有编码器和解码器层都使用体素化的点云，因此最后的解码器块产生体素化的全局特征。为了便于逐点分类预测，这些逐体素的全局特征与编码器之前通过MLP层捕获的高级逐点特征相连接。最后，通过对连接的特征应用额外的MLP层，体素级特征被细分为逐点特征，从而能够预测每个点的语义类。

### 3.4. 实例的头

实例头部遵循Panoptic-PolarNet的拓扑结构(Zhou et al., 2021)，并以BEV格式为每个实例生成两个预测：实例中心和到其相应中心的偏移量。在预测实例中心和偏移量时，使用BEV图像而不是3D数据有几个优点。首先，处理2D数据可以减少计算需求。其次，事物的实例很少沿着Z-axis重叠，这使得预测BEV中的中心和偏移量更容易。然而，该模块之前使用的数据是体素化的3D数据，因此需要进行维度转换。因此，采用特征投影层将3D特征转换为BEV特征，如图6所示。这一层对BEV图像中单个像素对应的体素特征执行每个通道的最大池化，特别是那些具有相同XandYindices的体素。因此，体素特征 $\mathbf{V} \in \mathbb{R}^{V_N \times C_N}$ 转化为BEV特征 $\in \mathbf{RH} \times \mathbf{W} \times \mathbf{N}$ ，其中，BEV特征的个数分别为体素个数和特征通道个数。

对变换后的BEV特征应用两个平行层；其中一层预测实例中心。为此，我们通过将特征投影应用于初始特征融合模块中生成的融合特征，并将其用作跳跃连接来增强上卷积。实例中心以热图 $\mathbf{I}_c \in \mathbb{R}^{H \times W \times 1}$ 的形式进行预测。实例中心的基本真值为BEV形式，遵循以每个实例的质量中心为中心的高斯分布。热图地面真值生成公式如下：

$$I_c = \max_i \exp \left( -\frac{(p - c_i)^2}{2\sigma^2} \right) \quad (5)$$

式中 $P$ 是BEV图像中的每个像素， $C_i$ 是每个实例的质心。

另一层预测每个BEV像素最近的实例中心的偏移量。这一层还使用跳过连接来整合来自第一个融合特征的精细信息，然后进行上卷积。偏移量预测 $\mathbf{I}_o \in \mathbb{R}^{H \times W \times 2}$ 有两个通道，因为这一层预测XandY directions中的偏移量。偏移量的ground truth也以BEV形式生成，表示每个实例的质心偏移量。

### 3.5. Panoptic处理模块

Panoptic处理模块将来自语义头部的语义预测与实例中心合并，并对预测进行偏移

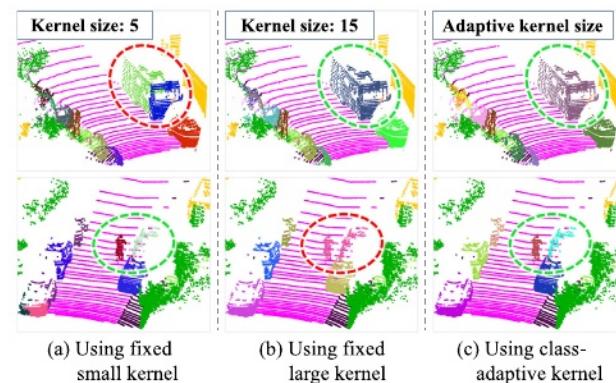


图7所示。类适应加工的必要性。该图说明了三种策略的结果：利用固定的大小和大小的内核大小，以及采用类自适应的内核大小进行精确的实例中心提取。

从实例头部为每个点分配类和实例id。在遵循Panoptic-PolarNet拓扑结构的同时，我们引入了一种额外的类自适应处理方法，该方法考虑了不同对象类的大小。

Panoptic-PolarNet的全光处理模块的第一步是通过对实例头部预测的中心热图应用非最大抑制(non-maximum suppression, NMS)来获取实例中心的像素坐标。但是，当为NMS使用固定的内核大小时，可能会产生不适当的结果。例如，如图7所示，使用较小的内核大小会导致较大的实例拥有多个中心，而使用较大的内核大小则会导致相邻的小实例共享单个中心。为了解决这个问题，我们的目标是通过类自适应处理对每个类应用不同的内核大小来提取类自适应中心。这种方法确保了无论实例大小或类别如何都能获得准确的结果，如图7的第三列所示。

Panoptic处理模块包括实例中心查找、实例聚类和语义与实例信息合并三个步骤。其中，在实例中心查找过程中采用了类自适应处理。这一步的输入由语义和实例中心热图预测组成，而输出则为每个实例生成不同的中心坐标。这一步需要与属于事物的类的典型大小成比例的值，并将它们作为参数存储在一个名为 $\mathbf{Th}_{size}$ 的预定义列表中。例如，我们在下面的SemanticKITTI数据集中给出了我们应用的每个事物的内核大小。

$$\begin{aligned} \mathbf{Th}_{size} = \{ &\text{car : 21, bicycle : 11, motorcycle : 11,} \\ &\text{truck : 25, other - vehicle : 25, person : 7,} \\ &\text{bicyclist : 13, motorcyclist : 13} \} \end{aligned}$$

类自适应实例中心查找的详细过程概述如下。首先，通过多数投票将语义预测 $\mathbf{S}$ 投影到BEV域，将其转化为 $\mathbf{S}_{BEV}$ 。随后，对于每一个属于事物的类，迭代以下三个步骤。(1)创建一个掩码，用于识别 $\mathbf{S}_{BEV}$ 内具有特定类的像素。(2)通过将内核大小设置为 $\mathbf{Th}_{size}$ 列表中特定类对应的参数来应用NMS算法。这个过程生成中心坐标的候选点。(3)候选中心坐标位置的语义预测与特定类匹配是很重要的。因此，在验证对应关系后，只有匹配的中心坐标才会被包含在最终的中心列表 $\mathbf{C}_{cor}$ 中。通过这些步骤，最终在BEV空间中导出实例中心坐标。这些步骤在算法1中有概述。

之后，实例聚类和语义与实例的合并遵循Panoptic-PolarNet。实例聚类步骤

表1

SemanticKITTI验证集上的定量全视分割结果。给出Full扫描上的性能是为了显示基于所使用的数据区域的性能趋势，而不是为了直接比较。这允许间接比较。

Point cloud region	Method	Input	PQ	$PQ^\dagger$	SQ	RQ	$PQ^{Th}$	$SQ^{Th}$	$RQ^{Th}$	$PQ^{St}$	$SQ^{St}$	$RQ^{St}$	mIoU
Full scans	Panoster	L	55.6	-	79.9	66.8	56.6	-	65.8	-	-	-	61.1
	DS-Net	L	57.7	63.4	77.6	68.0	61.8	78.2	68.8	54.8	77.1	67.3	63.5
	EfficientLPS	L	59.2	65.1	75.0	69.8	58.0	78.0	68.2	60.9	72.8	71.0	64.9
	Panoptic-PolarNet*	L	59.1	64.1	78.3	70.2	65.7	87.4	74.7	54.3	71.6	66.9	64.5
	Panoptic-PCSCNet*	L	61.3	65.3	80.0	70.4	67.2	90.0	72.9	57.0	72.7	68.5	64.9
	Panoptic-PHNet	L	61.7	-	-	-	69.3	-	-	-	-	-	65.7
	GP-S3Net	L	63.3	71.5	81.4	75.9	70.2	86.2	80.1	58.3	77.9	72.9	73.0
Cam FoV scans	PUPS	L	66.3	70.2	82.5	75.6	74.6	93.4	80.3	60.2	74.5	72.2	-
	Panoptic-PolarNet*	L	55.2	58.3	73.8	65.6	56.7	76.2	64.8	54.1	72.2	66.2	58.7
	Panoptic-PCSCNet*	L	59.5	62.6	80.6	70.7	61.9	89.1	71.0	57.7	74.3	70.4	62.6
<b>Panoptic-FusionNet (Ours)</b>		<b>L+C</b>	<b>64.3</b>	<b>66.6</b>	<b>82.1</b>	<b>73.6</b>	<b>70.9</b>	<b>92.4</b>	<b>76.9</b>	<b>59.5</b>	<b>74.6</b>	<b>71.2</b>	<b>65.4</b>

#### Algorithm 1 Class-adaptive instance center finding

**Input:**  $S \in \mathbb{R}^{N \times C}$ ,  $I_c \in \mathbb{R}^{H \times W \times 1}$ ,  $I_o \in \mathbb{R}^{H \times W \times 2}$ ;

**Output:** Class-adaptive searched centers  $C_{cor}$ ;

**Require:** List of the proportional value of typical sizes of classes belonging to *Things*  $\mathbf{Th}_{size}$ ;

```

 $C_{cor} := \{\}$ 
 $S_{BEV} \leftarrow projection(S)$ 
for all  $thing \in Things$  do
    if  $S_{BEV}(u, v) = thing$  then
         $mask(u, v) \leftarrow 1$ 
    else
         $mask(u, v) \leftarrow 0$ 
    end if
     $C_{thing} \leftarrow nms(I_c, KernelSize = \mathbf{Th}_{size}[thing])$ 
    for all  $center \in C_{thing}$  do
        if  $mask(center) = 1$  then
             $C_{cor}.append[center]$ 
        end if
    end for
end for

```

#### 4.1. 数据集

##### 4.1.1. SemanticKITTI

SemanticKITTI基于KITTI测程基准，使用Velodyne hd - 64e激光雷达提供总共43,441次扫描。由于它为训练和验证集提供了逐点语义和实例注释，因此可以用于泛光分割网络。类共由20个类组成，其中8个属于事物类，12个属于物质类。此外，它还提供了一个面向前方的单摄像头拍摄的同步图像。也就是说，每次激光雷达扫描都具有面向前方区域的相应图像。因此，出于数据融合的目的，Panoptic-FusionNet仅使用位于相机FoV(视场)中的点进行训练和评估，而不是整个360度。

##### 4.1.2. nuScenes

nuScenes提供了大约40,000个用Velodyne HDL-32E获得的点云扫描。对于训练集和验证集，它为每个点提供语义标签以及全景标签。nuScenes共包含16个类，其中10个属于事物，6个属于材料。在每个点云旁边，提供了6张同步图像，覆盖了完整的360度视角。这使得传感器融合不仅在视野(FoV)，而且在整个扫描。

#### 4.2. 实现细节

##### 4.2.1. 评价指标

Kirillov et al.(2019)定义的全光分割性能可以用全光质量(PQ)表示，如下式所示。PQ可以分解为语义质量(SQ)和识别质量(RQ)。SQ表示真阳性(TP)的平均IoU，RQ表示检测任务中的F1分数。在这种情况下，如果TP与地面真值的重叠面积超过50%，则视为匹配段。PQ、SQ和RQ这三个指标不仅适用于所有类，而且还分别适用于事物类和物质类。

$$PQ = SQ \times RQ \quad (6)$$

$$SQ = \frac{\sum_{TP} IoU}{|TP|}, RQ = \frac{|TP|}{|TP| + \frac{1}{2}|FP| + \frac{1}{2}|FN|} \quad (7)$$

将中心坐标与偏移量预测相结合，基于最小距离将共享相同最近中心的像素聚类。这为每个实例提供了BEV像素级数据的不同ID。最后，在合并语义和实例信息时，将语义预测与聚类实例相结合。对于属于the *Stuff*的区域，只分配一个语义类，而对于属于the *Things*的区域，则同时分配一个语义类和一个实例ID。这种综合的方法生成最终的panoptic预测。

#### 4. 实验

在本节中，我们展示了我们的Panoptic-FusionNet的实验结果。4.1节简要介绍了我们用于实验的数据集。第4.2节分享了培训Panoptic-FusionNet的详细配置。第4.3节提供了Panoptic-FusionNet的定量和定性演示，以验证传感器融合相对于单一激光雷达网络的优越性和效果。

我们还计算了另一个指标， $PQ^\dagger$ ，如Porzi, Bulo, Colovic和Kontschieder(2019)所述。 $PQ^\dagger$ 的设计是为了减少 $PQ$ 度量在物品分类错误时所带来的严重惩罚。为了计算 $PQ^\dagger$ ，对每个物品类别的IoU和每个物品类别的 $PQ$ 取平均值。这消除了对不需要大于0.5的IoU的内容进行性能测量的需要。此外，我们还提供了平均交联(mIoU)，这是所有类的语义分割指标。

#### 4.2.2. Pre-trained模型

在Panoptic-FusionNet的训练阶段，我们利用预训练的权重进行微调。我们的预训练过程旨在灌输Panoptic-FusionNet激光雷达流程部分的初步粗略理解。因此，我们独立构建了一个名为Panoptic-PCSCNet的基线模型，这是我们提出的融合模型的单一激光雷达版本。即该模型是在Panoptic-FusionNet中省去图像骨干和功能融合模块的一种形式。我们使用SemanticKITTI数据集对Panoptic-PCSCNet进行预训练，并将其权重作为激光雷达骨干和Panoptic-FusionNet的语义头和实例头的初始权重。

为了增强对事物的认识，我们在Panoptic-PCSCNet的预训练过程中实现了实例增强，Panoptic-PolarNet也采用了这种方法。实例增强的过程如下：在训练之前，对属于事物的实例对象进行采样和存储。在网络训练过程中，从采样数据中随机选择的实例被插入到扫描的未占用部分。由于这种方法增加了训练数据集本身中事物对象的数量，它可以潜在地提高事物的性能。因此，在激光雷达目标检测网络中被广泛采用。

#### 4.2.3. 损失函数

Panoptic-FusionNet从语义头生成一个预测，从实例头生成两个预测。这些预测中的每一个都相对于它们各自的基础真理进行计算，以获得单个损失，然后组合起来计算最终损失。首先，对于语义头预测的逐点类，使用交叉熵损失 $L_{ce}$ 和Lovasz-softmax损失 $L_{ls}$ 的总和。其次，对于实例头预测的中心热图，应用MSE(均方误差)损失。第三，将 $L_1$ 损失应用于实例头预测的偏移量。最终的损失如下面的等式所示：

$$L = w_{semantic}(L_{ce} + L_{ls}) + w_{center}L_{mse} + w_{offset}L_{l1} \quad (8)$$

它是三种不同损失的加权求和，权重系数设置为：WSEMANTC is 1, WCENTER is 100, WOFFSET is 10。

### 4.3. 评价

#### 4.3.1. 基线

为了比较所提出的模型的性能，我们使用了一个基线，包括仅依赖于激光雷达的各种全光分割方法。基线包括Panoptic-PCSCNet，Panoptic-FusionNet的激光雷达版本，以及初步开发的模型，如Panoster (Gasperini, Mahani, Marcos-Ramiro, Navab, & Tombari, 2021), DS-Net, EfficientLPS, Panoptic-PolarNet, panopti - phnet (Li, He等人, 2022), GP-S3Net (Razani等人, 2021)和PUPS (Su等人, 2023)。

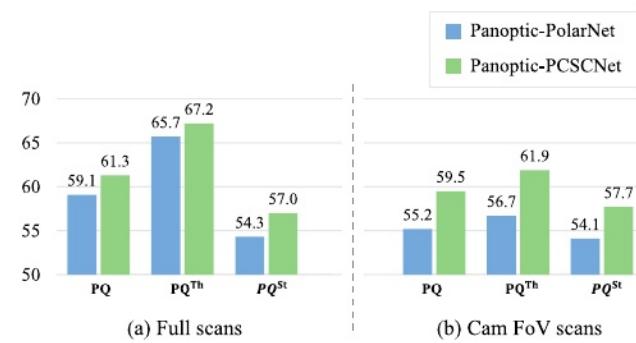


图8所示。该图显示了一般点云全视分割模型中基于数据利用区域的性能变化。

**SemanticKITTI基线。**在我们进行SemanticKITTI的比较之前，有一件事必须澄清。初步开发的基线模型提供的性能采用了覆盖360度的SemanticKITTI激光雷达数据的全面扫描，以进行训练和评估：为了方便起见，我们将这种情况称为全面扫描。然而，本文提出的融合模型仅使用位于相机视场(FoV)内的点云，旨在利用相机和激光雷达数据重叠的区域：这种情况很容易被称为Cam FoV扫描。在相同的条件下，比较Full扫描和Cam FoV扫描的性能不是一个公平的比较。因此，我们研究了在使用Cam FoV扫描训练和评估普通网络时，与使用Full扫描训练和评估网络时相比，性能是如何变化的。这种性能趋势使逻辑比较成为可能，即使基线的给定性能仅适用于完整扫描。

性能变化的趋势可以在图8中看到。我们使用Panoptic-PolarNet和Panoptic-PCSCNet进行分析。为此，有必要进行网络再训练和评估，因此需要一个可访问的代码库。因此，在现有的各种基线模型中，我们选择了Panoptic-PolarNet，该模型有代码可用，性能令人满意。首先，考虑到Panoptic-PolarNet，当使用Cam FoV扫描时，与使用Full扫描相比，Panoptic质量(PQ)，代表整体性能，下降了3.9%。进一步检查，虽然有类似的，但有显著下降9%的掉头掉头掉头掉头掉头掉头掉头掉头掉头掉头掉头掉头。这似乎是由于当使用Cam FoV扫描时，对应于事物的点明显减少而不是东西。第二个模型Panoptic-PCSCNet的趋势分析结果与此相似。PQ下降了约1.7%，这也被分析为在很大程度上受到了PQ指数显著下降的影响，即 $<Q??NGS>$ 。

因此，我们通过对上述两种模型的Full扫描和Cam FoV扫描结果进行比较得出的趋势是不可避免的性能下降。造成这种情况的主要原因是使用的数据量减少。这种减少是因为相机的FoV约为60度，导致与完全扫描相比数据减少了83%。分析特别显示了事物性能的显著下降。这些动态物体表现出由它们的角度和方向决定的不同形状。然而，由于只利用相机FoV内的数据，捕捉不同角度和方向的物体的能力受到限制，导致其泛化性能明显下降。

**nuScenes基线。**在nuScenes的情况下，由于它提供了覆盖整个360度视图的6个同步图像的图像数据，因此它可以对完整扫描进行训练和评估。

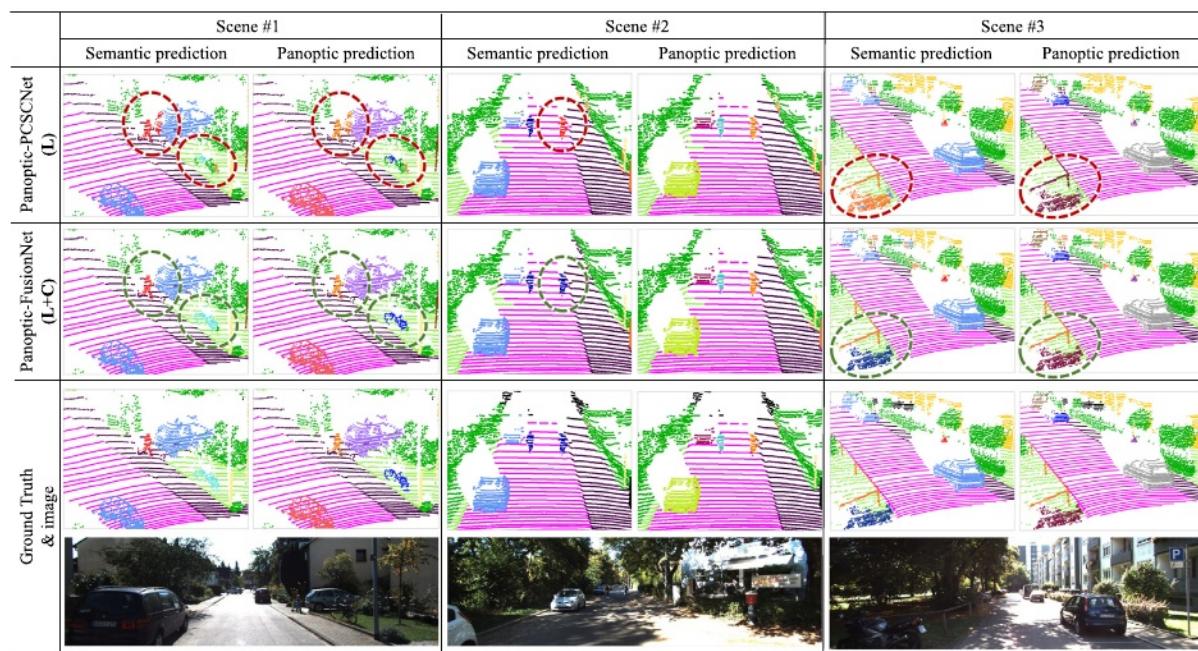


图9所示。这将使SemanticKITTI验证集的全视分割结果可视化。第一列显示了提出的Panoptic-FusionNet的预测，第二列显示了Panoptic-PCSCNet仅使用激光雷达的预测，第三列代表了地面真实值和相应的图像。

#### 4.3.2. 定量结果

**SemanticKITTI的定量结果。**表1比较了SemanticKITTI验证集与基线模型的性能。第一列表示点云的区域，第二列是包括基线模型和建议模型的方法。其中，标注了星号(\*)的两个模型，用于根据点云区域进行性能趋势分析。第三列是输入数据，由于Panoptic-FusionNet是第一个用于全景分割的相机-激光雷达融合网络，只有它具有多模态输入。其他列表示所有类的性能和每个事物的性能，mIoU表示语义分割。

与使用凸轮FoV扫描的基线相比，Panoptic-FusionNet实现了卓越的全光质量。首先，它的PQ比Panoptic-PolarNet高9.1%，后者似乎明显受益于 $PQ_{T\otimes INGS}$ 的14.2%的提高。相较于物，物的高性能对于智能车辆上的panoptic分割任务来说非常重要。这是因为物是动态对象，因此对驾驶策略的影响最大。因此，分类良好的事物提供了有价值的结果。

与使用凸轮FoV扫描的第二个基线模型相比，Panoptic-FusionNet的PQ提高了4.8%， $PQ_{T\otimes INGS}$ 比Panoptic-PCSCNet提高了9%。如前所述，Panoptic-PCSCNet是Panoptic-FusionNet的一个版本，只留下激光雷达流，这允许在相同条件下公平分析相机融合的效果。与Panoptic-PolarNet的比较结果相似，things类的性能比stuff类的性能提高得更多。为了分析性能改进的原因，我们在表2中导出了特定于类的PQ。虽然所有类的性能都有轻微的提升，但在摩托车、人、卡车等对象中，显著的提升尤为明显。这些都属于物品类，与物品类相比，它们的尺寸相对较小，因此在单独使用单个激光雷达时，它们会带来挑战。然而，当相机和激光雷达结合使用时，在特征提取阶段，相机相对密集的特征图补偿了激光雷达稀疏的特征图。因此，通过生成比仅使用激光雷达获得的特征在语义上更增强的特征，Panoptic-FusionNet可以实现更高的识别性能。

最后，这里是Panoptic-FusionNet与使用完整扫描的基线的间接比较。如前所述，当使用凸轮FoV扫描时，与使用Full扫描时相比，性能会下降。我们目前比较的所有使用完整扫描的基线的PQ都低于Panoptic-FusionNet。如果这些基线仅使用凸轮FoV扫描，我们可以预期性能将比完全扫描进一步下降。因此，在凸轮FoV扫描中，与Panoptic-FusionNet的性能差距将更大。因此，这种间接比较得出的结论是，所提出的模型比其他基线实现了更高的性能。

**nuScenes的定量结果。**表3提供了使用nuScenes验证集的Panoptic-FusionNet和基线模型的性能，每列的内容与表1一致。Panoptic-FusionNet在实验中表现出最高的平均PQ，与Panoptic-PCSCNet相比，PQ高出4.5%，量化了传感器融合的好处。Panoptic-FusionNet的这种优势不仅在SemanticKITTI数据集中很明显，而且在nuScenes中也很明显。这些结果表明，所提出的模型展示了性能改进，这些改进不局限于单个数据集，而是在不同的传感器配置和环境中进行了验证。

#### 4.3.3. 定性结果

为了定性地比较所提出模型的性能，我们在图9中展示了SemanticKITTI上的推理结果。总共有三个场景，其中第一行表示Panoptic-PCSCNet的预测，第二行表示Panoptic-FusionNet的预测，第三行表示ground truth和相应的图像。对于每个场景，我们不仅可视化了panoptic预测，还可可视化了语义预测。可视化的泛视预测可以直接反映事物的语义类，瞬间表示出来。然而，对于事物来说，每个实例都用一种不同的颜色来表示，从而隐藏了语义类。同时呈现两种结果有助于进行更精确的分析。

第一个场景是一个儿童行人和一辆停放在道路右侧的自行车。对于仅利用激光雷达的Panoptic-PCSCNet，自行车在语义预测中被部分分类为植被。此外，它错误地将行人和周围区域标记为行人。因此，泛视预测产生

表2

SemanticKITTI验证集上的类特定全视分割结果。

Method	Input	PQ per class (%)																		mean PQ (%)	
		car	bicycle	motorcycle	truck	bus	person	bicyclist	motorcyclist	road	parking	sidewalk	other-ground	building	fence	vegetation	trunk	terrain	pole	traffic-sign	
Panoptic-PCSCNet	L	91.4	59.5	54.7	49.0	55.1	80.9	89.0	16.0	94.1	29.7	74.2	0.0	83.1	25.8	82.5	56.6	49.0	68.5	72.1	59.5
<b>Panoptic-FusionNet</b>	L+C	94.0	61.8	63.9	78.4	61.0	85.1	92.5	30.5	95.4	28.5	75.8	0.0	86.0	31.4	83.2	59.0	49.8	72.2	73.1	64.3

表3

nuScenes验证集上的定量全视分割结果。

Method	Input	PQ	PQ <sup>†</sup>	SQ	RQ	PQ <sup>Th</sup>	SQ <sup>Th</sup>	RQ <sup>Th</sup>	PQ <sup>St</sup>	SQ <sup>St</sup>	RQ <sup>St</sup>	mIoU
DS-Net	L	42.5	51.0	83.6	50.3	32.5	83.1	38.3	59.2	84.4	70.3	70.7
EfficientLPS	L	59.2	62.8	82.9	70.7	51.8	80.6	62.7	71.5	84.3	84.1	69.4
Panoptic-PolarNet	L	67.7	71.0	86.0	78.1	65.2	87.2	74.0	71.9	83.9	84.9	69.3
Panoptic-PCSCNet	L	72.7	75.4	86.4	84.8	71.2	86.6	82.9	75.1	84.2	84.2	69.8
Panoptic-PHNet	L	74.7	77.7	88.2	84.2	74.0	89.0	82.5	75.9	86.8	86.9	79.7
GP-S3Net	L	61.0	67.5	84.1	72.0	56.0	85.3	65.2	66.0	82.9	78.7	75.8
PUPS	L	74.7	77.3	89.4	83.3	75.4	91.8	81.9	73.6	85.3	85.6	-
<b>Panoptic-FusionNet (Ours)</b>	<b>L+C</b>	<b>77.2</b>	<b>79.3</b>	<b>87.8</b>	<b>87.2</b>	<b>77.5</b>	<b>88.2</b>	<b>87.7</b>	<b>76.2</b>	<b>86.0</b>	<b>85.9</b>	<b>73.4</b>

两个对象的错误结果。相比之下，在Panoptic-FusionNet中，行人和自行车在语义预测中被很好地分类为点级别。这导致在panoptic预测中，两个对象都被很好地识别，每个对象都有唯一的实例id。

第二个场景呈现了前面有两个骑自行车的人的情况。仅使用激光雷达，左边的骑自行车的人被正确分类，但右边的骑自行车的人被错误地分类为行人。这是单个激光雷达常见的分类错误，因为骑自行车的人的点云形状几乎是正面的，与行人相似。泛视预测看起来分类很好，每个物体都正确聚类，但它包含分类错误，不可靠。然而，当在语义预测中使用摄像头进行融合时，这两个骑自行车的人都能被正确识别出来。使用Panoptic-FusionNet的全景预测，以及语义结果，非常类似于所有事物对象的地面真相。

在第三个场景中，一辆摩托车停在道路的左侧。在使用Panoptic-PCSCNet时，摩托车被错误地分类为围栏，由于摩托车属于物品类别，因此无法在panoptic预测中分配实例ID。然而，有了Panoptic-FusionNet，不仅摩托车，而且附近的围栏也能被正确分类。panoptic的结果显示，摩托车被很好地聚类，并被分配了一个实例ID。

在所有三个场景示例中，融合摄像头和激光雷达的模型比使用单个激光雷达的模型做出了更好的预测。虽然对物体的预测是相似的，但对物体的预测却有所改善。似乎该模型受益于图像中额外的密集像素纹理信息，特别是对于仅基于点云形状难以识别的物体。

#### 4.3.4. 消融研究

我们提出了一项消融研究的结果，以分析每个成分的影响，如表4所示。此表有助于评估与相机数据的融合效果，应用类别自适应处理并使用带有实例增强的预训练权重。

表4

消融实验。CA处理表明类别自适应处理，和预训练的IA

表示使用实例增强预训练模型。

Method	Fusion w/ Camera	CA Processing	Pre- trained IA	PQ	PQ <sup>Th</sup>	mIoU
Panoptic- PCSCNet <b>(L)</b>	-	-	-	55.7	50.6	59.2
	-	✓	-	57.7	55.5	59.2
	-	✓	✓	59.5	61.9	62.6
Panoptic- FusionNet <b>(L+C)</b>	✓	-	-	60.0	59.9	63.6
	✓	✓	-	62.8	67.3	64.9
	✓	✓	✓	64.2	70.9	65.5

我们最初开始训练Panoptic-PCSCNet和Panoptic-FusionNet，没有任何额外的详细实施。这两种方法的区别在于对相机数据的利用，使我们能够评估通过与相机融合获得的独家性能改进，这导致PQ提高了4.3%。其次，我们将类别自适应处理模块集成到每个网络中，导致Panoptic-PCSCNet和Panoptic-FusionNet分别增加2%和2.8%。由于类自适应处理模块的目的是增强事物的性能，我们通过检查PQ<sup>Th</sup>来验证其有效性，观察到每种方法的性能提升分别为4.9%和7.4%。第三，我们使用了带有实例增强的预训练权重，这使得Panoptic-PCSCNet的PQ上升1.8%，Panoptic-FusionNet的PQ上升1.4%。该模块也对things的性能提升做出了显著贡献，PQ<sup>Th</sup>分别提升了6.4%和3.6%。因此，通过这些消融研究，我们实验证明了每个模块都提高了全光分割性能。

## 5. 结论

在本文中，我们提出了一种基于相机-激光雷达融合的点云全光分割网络(**panoptic - fusionnet**)，首次融合

据我们所知,为这项任务设计的模型。我们提出了一种带有特征融合模块的网络,可以确保准确的特征对齐,从而通过相机-激光雷达融合增强对周围环境的理解。使用增强的融合特征,我们生成语义、实例中心和偏移量预测。最后,具有类自适应处理的panoptic处理模块,无论事物大小如何,都可以实现可靠的聚类。在SemanticKITTI和nuScenes数据集中,我们实现了比使用单个激光雷达高4.8%和4.5%的全光质量。

总之,

(1)基于融合的点云全分割网络:首次提出了基于相机-激光雷达融合的全分割网络。与使用单个激光雷达相比,我们在SemanticKITTI验证集和nuScenes验证集上实现了性能提升。特别是,对于像行人和自行车这样的小尺寸物体,性能的增强表明,通过增加摄像头,有可能减轻单个激光雷达的稀疏性缺点。

(2)精确特征对齐的特征融合模块:我们提出了一种特征融合模块,可以精确对齐从相机和激光雷达中提取的特征。通过提前准备一个点-体素对应表,并在训练和推理过程中对其进行查询,我们可以高效地生成精确的对齐。这使得高精度融合特征的生成成为可能。

在未来的工作中,首先,我们的目标是提高相机-激光雷达融合能力的性能。虽然目前提出的模块允许精确对准和融合,但我们计划通过实施额外的模块(如注意力网络)来进一步提高融合特征的质量。其次,我们计划进行网络优化以减少其计算负载,利用两个传感器确实比单个传感器需要更多的计算量。我们提出的panoptic-fusionnet实现的每秒帧数(fps)为3.5,使实时应用具有挑战性。因此,我们打算通过优化和简化网络来开发实时使用的模型,以减少计算需求。

## 6. 缩略语和符号列表

### 激光雷达

二维	光探测与测距
3 d	二维
ASPP	三维
贝芙	Atrous空间金字塔池
MLP多层感知器	鸟瞰图
NMS非最大抑制	
视场	
魅人党	视野
平方	展示全景的质量
中移动	SQ语义质量
mIoU mIoU表示并集交点	识别质量
帧/秒	
F	帧/秒
C	焦距
R	主点
T	旋转矩阵
S	转换矩阵
$S_{BEV}$	语义的预测
BEV投影语义预测	
V	Voxel-wise特性

$B$	BEV feature
$V_n$	The number of non-empty voxels
$C_n$	The number of feature channels
$I_c$	Instance center heatmap
$p$	Pixel in BEV image
$c_i$	Mass center of instance
$I_o$	Instance offset
$Th_{size}$	List of proportional value of things
$C_{cor}$	Class-adaptive searched centers

### CRediT author contribution statement(作者贡献声明)

宋海民: 概念化, 方法论, 软件。赵洁云: 验证、数据策展。Jinsu Ha: 可视化, 资源。Jaehyun Park: 写作-评论&编辑。Kichun Jo: 监督, 项目管理。

### 竞利申报

作者声明,他们没有已知的竞争经济利益或个人关系,可能会影响本文中报道的工作。

### 数据可用性

已使用的数据是保密的。

### 致谢

本工作由韩国政府(MSIT)资助的韩国国家研究基金会(NRF)资助。(no.rs - 00209252)

### 参考文献

- Bai, X., Hu, Z., Zhu, X., Huang, Q., Chen, Y., Fu, H., et al. (2022). Transfusion: Robust lidar-camera fusion for 3d object detection with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 1090–1099).
- Behley, J., Garbade, M., Milioto, A., Quenzel, J., Behnke, S., Stachniak, C., et al. (2019). SemanticKITTI: A dataset for semantic scene understanding of lidar sequences. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 9297–9307).
- Caesar, H., Bankiti, V., Lang, A. H., Vora, S., Lioung, V. E., Xu, Q., et al. (2020). Nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 11621–11631).
- Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2017). DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4), 834–848.
- Cheng, B., Collins, M. D., Zhu, Y., Liu, T., Huang, T. S., Adam, H., et al. (2020). Panoptic-deeplab: A simple, strong, and fast baseline for bottom-up panoptic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 12475–12485).
- Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T., & Ronneberger, O. (2016). 3D U-net: learning dense volumetric segmentation from sparse annotation. In *Medical image computing and computer-assisted intervention-MICCAI 2016: 19th international conference, athens, Greece, October 17-21, 2016, proceedings, part II* 19 (pp. 424–432). Springer.
- El Madawi, K., Rashed, H., El Sallab, A., Nasr, O., Kamel, H., & Yogamani, S. (2019). Rgb and lidar fusion based 3d semantic segmentation for autonomous driving. In *2019 IEEE intelligent transportation systems conference* (pp. 7–12). IEEE.
- Gasperini, S., Mahani, M.-A. N., Marcos-Ramiro, A., Navab, N., & Tombari, F. (2021). Panoster: End-to-end panoptic segmentation of lidar point clouds. *IEEE Robotics and Automation Letters*, 6(2), 3216–3223.
- He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision* (pp. 2961–2969).
- He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778).
- Hong, F., Zhou, H., Zhu, X., Li, H., & Liu, Z. (2021). Lidar-based panoptic segmentation via dynamic shifting network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 13090–13099).

- Hu, Q., Yang, B., Xie, L., Rosa, S., Guo, Y., Wang, Z., et al. (2020). Randla-net: Efficient semantic segmentation of large-scale point clouds. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 11108–11117).
- Kirillov, A., He, K., Girshick, R., Rother, C., & Dollár, P. (2019). Panoptic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 9404–9413).
- Li, J., He, X., Wen, Y., Gao, Y., Cheng, X., & Zhang, D. (2022). Panoptic-PHNet: Towards real-time and high-precision LiDAR panoptic segmentation via clustering pseudo heatmap. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 11809–11818).
- Li, Y., Yu, A. W., Meng, T., Caine, B., Ngiam, J., Peng, D., et al. (2022). Deepfusion: Lidar-camera deep fusion for multi-modal 3d object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 17182–17191).
- Lin, Z., & Wang, Y. (2022). BEV-MAE: Bird's eye view masked autoencoders for outdoor point cloud pre-training. arXiv preprint arXiv:2212.05758.
- Milioto, A., Vizzo, I., Behley, J., & Stachniss, C. (2019). Rangenet++: Fast and accurate lidar semantic segmentation. In *2019 IEEE/RSJ international conference on intelligent robots and systems* (pp. 4213–4220). IEEE.
- Mohan, R., & Valada, A. (2021). Efficientps: Efficient panoptic segmentation. *International Journal of Computer Vision*, 129(5), 1551–1579.
- Mohapatra, S., Yogamani, S., Gotzig, H., Milz, S., & Mader, P. (2021). BEVDetNet: bird's eye view LiDAR point cloud based real-time 3D object detection for autonomous driving. In *2021 IEEE international intelligent transportation systems conference* (pp. 2809–2815). IEEE.
- Pang, S., Morris, D., & Radha, H. (2020). CLOCs: Camera-LiDAR object candidates fusion for 3D object detection. In *2020 IEEE/RSJ international conference on intelligent robots and systems* (pp. 10386–10393). IEEE.
- Park, J., Kim, C., Kim, S., & Jo, K. (2023). PCSCNet: Fast 3D semantic segmentation of LiDAR point cloud for autonomous car using point convolution and sparse convolution network. *Expert Systems with Applications*, 212, Article 118815.
- Porzi, L., Bulo, S. R., Colovic, A., & Kortschieder, P. (2019). Seamless scene segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 8277–8286).
- Qi, C. R., Yi, L., Su, H., & Guibas, L. J. (2017). Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in Neural Information Processing Systems*, 30.
- Razani, R., Cheng, R., Li, E., Taghavi, E., Ren, Y., & Bingbing, L. (2021). GP-S3Net: Graph-based panoptic sparse semantic segmentation network. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 16076–16085).
- Sirohi, K., Mohan, R., Büscher, D., Burgard, W., & Valada, A. (2021). Efficientps: Efficient lidar panoptic segmentation. *IEEE Transactions on Robotics*, 38(3), 1894–1914.
- Su, S., Xu, J., Wang, H., Miao, Z., Zhan, X., Hao, D., et al. (2023). PUPS: Point cloud unified panoptic segmentation. arXiv preprint arXiv:2302.06185.
- Tan, M., & Le, Q. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning* (pp. 6105–6114). PMLR.
- Thomas, H., Qi, C. R., Deschaud, J.-E., Marcotegui, B., Goulette, F., & Guibas, L. J. (2019). Kpconv: Flexible and deformable convolution for point clouds. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 6411–6420).
- Wang, Y., Shi, T., Yun, P., Tai, L., & Liu, M. (2018). Pointseg: Real-time semantic segmentation based on 3d lidar point cloud. arXiv preprint arXiv:1807.06288.
- Wang, H., Zhu, Y., Green, B., Adam, H., Yuille, A., & Chen, L.-C. (2020). Axial-deeplab: Stand-alone axial-attention for panoptic segmentation. In *Computer vision–ECCV 2020: 16th European conference, Glasgow, UK, August 23–28, 2020, proceedings, part IV* (pp. 108–126). Springer.
- Xiong, Y., Liao, R., Zhao, H., Hu, R., Bai, M., Yumer, E., et al. (2019). Upsnet: A unified panoptic segmentation network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 8818–8826).
- Yan, X., Gao, J., Zheng, C., Zheng, C., Zhang, R., Cui, S., et al. (2022). 2Dpass: 2d priors assisted semantic segmentation on lidar point clouds. In *European conference on computer vision* (pp. 677–695). Springer.
- Yoo, J. H., Kim, Y., Kim, J., & Choi, J. W. (2020). 3D-cvf: Generating joint camera and lidar features using cross-view spatial feature fusion for 3d object detection. In *Computer vision–ECCV 2020: 16th European conference, Glasgow, UK, August 23–28, 2020, proceedings, part XXVII 16* (pp. 720–736). Springer.
- Zhang, Y., Zhou, Z., David, P., Yue, X., Xi, Z., Gong, B., et al. (2020). Polarnet: An improved grid representation for online lidar point clouds semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 9601–9610).
- Zhao, L., Zhou, H., Zhu, X., Song, X., Li, H., & Tao, W. (2021). Lif-seg: Lidar and camera image fusion for 3d lidar semantic segmentation. arXiv preprint arXiv:2108.07511.
- Zhou, Y., & Tuzel, O. (2018). Voxelenet: End-to-end learning for point cloud based 3d object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4490–4499).
- Zhou, Z., Zhang, Y., & Foroosh, H. (2021). Panoptic-polarnet: Proposal-free lidar point cloud panoptic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 13194–13203).
- Zhou, H., Zhu, X., Song, X., Ma, Y., Wang, Z., Li, H., et al. (2020). Cylinder3d: An effective 3d framework for driving-scene lidar semantic segmentation. arXiv preprint arXiv:2008.01550.
- Zhuang, Z., Li, R., Jia, K., Wang, Q., Li, Y., & Tan, M. (2021). Perception-aware multi-sensor fusion for 3d lidar semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 16280–16290).