BEVFusion:具有统一鸟瞰图表示的多任务多传感器融合

刘志健 麻省理工学院
昊天堂 麻省理工学院
亚历山大·阿米尼 麻省理工学院
杨新宇 上海交通大学
毛惠子 OmniML
丹妮拉·罗斯 麻省理工学院
宋涵 麻省理工学院

https://bevfusion.mit.edu

抽象的

多传感器融合对于准确可靠的自动驾驶系统至关重要。最近的方法是基于点级融合:用相机特征增强 Li DAR 点云。然而,相机到激光雷达的投影丢弃了相机特征的语义密度,阻碍了这种方法的有效性,特别是对于面向语义的任务(如 3D 场景分割)。在本文中,我们使用 BEVFusion 打破了这种根深蒂固的惯例,这是一种高效且通用的多任务多传感器融合框架。它统一了共享鸟瞰图(BEV)表示空间中的多模态特征,很好地保留了几何和语义信息。为了实现这一点,我们通过优化的 BEV 池来诊断和提升视图转换中的关键效率瓶颈,将延迟减少40 倍以上。 BEVFusion 基本上与任务无关,并且无缝支持不同的 3D 感知任务,几乎没有架构变化。它建立了 nuScenes 的最新技术,在 3D 对象检测上实现了1.3%的 mAP 和 NDS,在BEV 地图分割上实现了 13.6%的 mIoU ,计算成本降低了1.9 倍。重现我们结果的代码可在https://github.com/mit-han-lab/bevfusion 获得。

1 简介

自动驾驶系统配备了多种传感器。例如,Waymo 的自动驾驶汽车有 29 个摄像头、6 个雷达和 5 个激光雷达。不同的传感器提供互补的信号:例如,相机捕捉丰富的语义信息,激光雷达提供准确的空间信息,而雷达提供即时速度估计。因此,多传感器融合对于准确可靠的感知具有重要意义。

来自不同传感器的数据以根本不同的方式表示:例如,摄像机以透视图捕获数据,而 LiDAR 以 3D 视图捕获数据。为了解决这种视图差异,我们必须找到一个适合多任务多模态特征融合的统一表示。由于2D 感知的巨大成功,自然的想法是将 LiDAR 点云投影到相机上,并使用 2D CNN 处理 RGB-D 数据。然而,这种 LiDAR 到相机的投影引入了严重的几何失真(参见图1a) ,这使得它对于面向几何的任务(例如 3D 对象识别)效果较差。

最近的传感器融合方法遵循另一个方向。他们使用语义标签[54]、 CNN 特征[55、23]或 2D 图像[68] 中的虚拟点来增强 LiDAR 点云,然后应用

表示贡献相等。前两位作者按字母顺序列出。

预印本。正在审查中。

(a) 到相机:几何损失(b) 到激光雷达:语义损失 (c) 共享BEVFusion空间

图 1:BEVFusion 在共享的 BEV 空间中统一了摄像头和 LiDAR 功能,而不是将一种模态映射到另一种模态。它保留了相机的语义密度和激光雷达的几何结构。

一个现有的基于 LiDAR 的检测器来预测 3D 边界框。尽管它们在大规模检测基准上表现出了卓越的性能,但这些点级融合方法几乎不能用于面向语义的任务,例如 BEV 地图分割[37、39、22、70] 。这是因为相机到 LiDAR 的投影在语义上是有损的（见图1b） :对于典型的 32 束 LiDAR扫描仪,只有 5% 的相机特征会与 LiDAR 点匹配,而其他所有特征都会被丢弃。

对于更稀疏的激光雷达（或成像雷达） ,这种密度差异将变得更加剧烈。

在本文中,我们提出 BEVFusion 来统一共享鸟瞰图 (BEV)表示空间中的多模态特征,以用于任务不可知的学习。我们保持几何结构和语义密度（见图1c)并自然支持大多数 3D 感知任务（因为它们的输出空间可以自然地在 BEV 中捕获） 。在将所有特征转换为 BEV 时,我们确定了视图转换中主要的效率瓶颈:即,仅 BEV 池操作就占用了模型运行时间的 80% 以上。然后,我们提出了一个具有预计算和间隔缩减的专用内核来消除这个瓶颈,实现超过40 倍的加速。最后,我们应用全卷积 BEV 编码器来融合统一的 BEV 特征,并附加一些特定于任务的头来支持不同的目标任务。

BEVFusion 在 nuScenes 基准测试中设置了新的最先进的性能。在 3D 对象检测方面,它在所有解决方案中排名第一。 BEVFusion 展示了 [illegible] 本提供所有这些结果。

BEVFusion打破了长期以来认为点级融合是多传感器融合最佳解决方案的信念。简单也是它的关键优势。我们希望这项工作能够为未来的传感器融合研究提供一个简单而强大的基线,并激发研究人员重新思考通用多任务多传感器融合的设计和范式。

2 相关工作

基于 LiDAR 的 3D 感知。研究人员设计了单级 3D 对象检测器[72、21、65、73、66、71] ,它们使用PointNets [41]或 SparseConvNet [17]提取扁平点云特征,并在 BEV 空间中执行检测。后来,尹等人。 [67]和其他人[15, 5, 42, 14, 6, 60]探索了无锚定 3D 对象检测。另一个研究方向[49, 10, 50, 47, 48, 24]侧重于两阶段目标检测,它将 RCNN 网络添加到现有的单阶段目标检测器中。

还有专门用于 3D 语义分割的 U-Net 模型[17、13、52、33、75] ,这是离线高清地图构建的一项重要任务。

基于相机的 3D 感知。由于激光雷达传感器的高成本,研究人员在仅相机的 3D 感知上投入了大量精力。 FCOS3D [57]使用额外的 3D 回归分支扩展了图像检测器[53] ,后来在深度建模方面进行了改进[58, 4]。

DETR3D [59]、 PETR [30]和 Graph DETR3D [11]设计了基于 DETR [74, 61] 的检测头,而不是在透视图中执行对象检测,并且在 3D空间中具有可学习的对象查询。受基于 LiDAR 的检测器设计的启发,另一种仅使用相机的 3D 感知模型使用视图转换器[37、46、45、39]将相机特征从透视图显式转换为鸟瞰图。 BEVDet [20]和M2BEV [63]有效地将 LSS [39]和OFT [46]扩展到 3D 对象检测,在发布时实现了最先进的性能。 CaDDN [43]向视图转换器添加了显式深度估计监督。 BEVDet4D [19]、 BEVFormer [25]和 PETRv2 [31]在多相机 3D 对象检测中利用时间线索,取得了显着

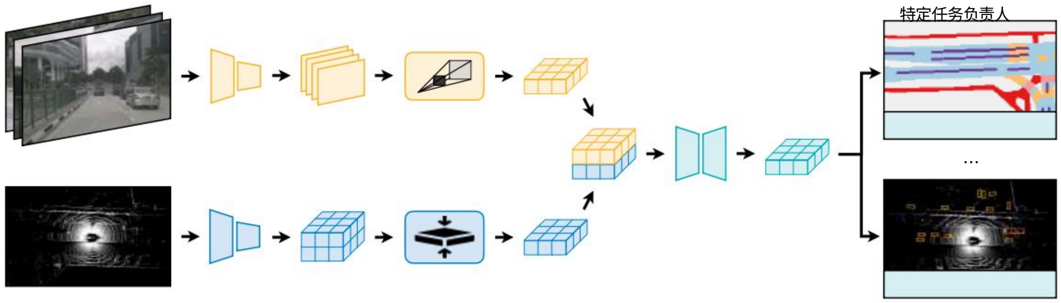


图 2： BEVFusion从多模式输入中提取特征,并使用视图转换有效地将它们转换为共享鸟瞰图 (BEV) 空间。它将统一的 BEV 特征与全卷积 BEV 编码器融合在一起,并支持具有特定任务头的不同任务。

对单帧方法的改进。 BEVFormer [25]、 CVT [70]和 Ego3RT [35]也研究使用多头注意力来执行视图转换。

多传感器融合。最近,多传感器融合引起了 3D 检测社区越来越多的兴趣。现有的方法可以分为提案级和点级融合方法。

MV3D [7]在 3D 中创建对象建议并将建议投影到图像以提取 RoI 特征。
F-PointNet [40]、 F-ConvNet [62]和 CenterFusion [36]都将图像提议提升为 3D 平截头体。
最近,FUTR3D [8]和 T[illegible]并将图像特征融合到这些建议中。提案级融合方法是以对象为中心的,不能简单地推广到其他任务,例如 BEV 地图分割。另一方面,点级融合方法通常将图像语义特征绘制到前景 LiDAR 点上,并对修饰的点云输入执行基于 LiDAR 的检测。因此,它们既以对象为中心,又以几何为中心。

在所有这些方法中,PointPainting [54]、 PointAugmenting [55]、 MVP [68]、 FusionPainting [64]、 AutoAlign [12]和 FocalSparseCNN [9]是（LiDAR)输入级装饰,而 Deep Continuous Fusion [27]和 DeepFusion [23]是特征级装饰。

多任务学习。多任务学习在计算机视觉社区中得到了很好的研究。研究人员研究了联合执行对象检测和实例分割[44, 3] ,并已扩展到姿势估计和人与对象交互[18, 51, 56, 16]。包括M2BEV [63]、 BEVFormer [25]和 BEVerse [69]在内的一些并行工作在 3D 中联合执行对象检测和 BEV 分割。以上方法均未考虑多传感器融合。 MMF [26]同时使用相机和激光雷达输入进行深度补全和对象检测,但仍然以对象为中心,不适用于 BEV 地图分割。

与所有现有方法相比,BEVFusion 在共享的 BEV 空间中执行传感器融合,并平等地对待前景和背景、几何和语义信息。 BEVFusion 是一个通用的多任务多传感器感知框架。

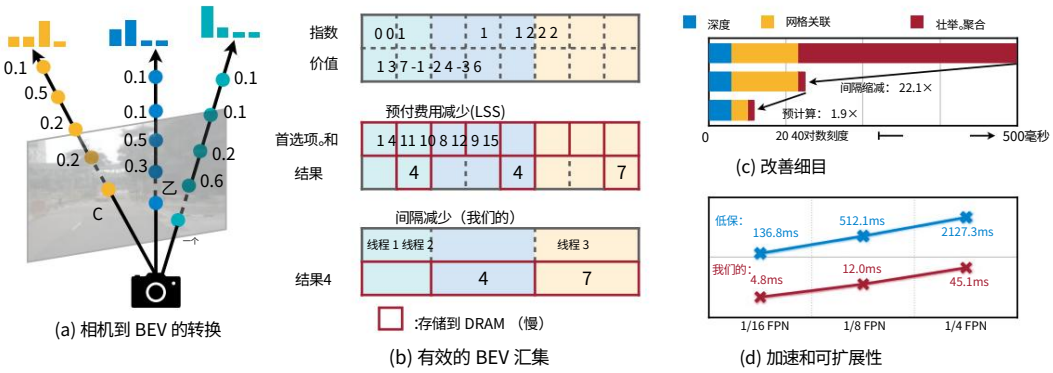
3 方法

BEVFusion专注于多传感器融合（即多视角相机和LiDAR）,用于多任务3D感知（即检测和分割）。我们在图2中概述了我们的框架。
给定不同的感官输入,我们首先应用特定于模态的编码器来提取它们的特征。我们将多模态特征转换为统一的 BEV 表示,同时保留几何和语义信息。我们确定了视图转换的效率瓶颈,并通过预计算和间隔缩减来加速 BEV 池化。然后,我们将基于卷积的 BEV编码器应用于统一的 BEV 特征,以减轻不同特征之间的局部错位。

最后,我们附加了一些特定于任务的头来支持不同的 3D 任务。

3.1 统一表示

不同的特征可以存在于不同的视图中。例如,相机特征位于透视图中,而 LiDAR/雷达特征通常位于 3D/鸟瞰图中。即使对于相机功能,它们中的每一个都具有不同的视角（即,前、后、左、右）。此观点不符



(a) 相机到 BEV 的转换 (b) 有效的 BEV 汇集 (c) 改善细目 (d) 加速和可扩展性

图 3:Camera-to-BEV 转换（a)是在统一中执行传感器融合的关键步骤
纯电动车空间。但是,现有的实现速度非常慢,一次可能需要 2 秒
场景。我们提出了有效的 BEV 池 (b),使用间隔缩减和快速网格关联
预计算,为视图转换模块（c,d)带来约40 倍的加速。

使特征融合变得困难,因为不同特征张量中的相同元素可能对应
到完全不同的空间位置（并且天真的元素特征融合在
这个案例）。因此,找到一个共享的表示是至关重要的,这样（1)所有传感器特征都可以
可以很容易地转换为它而不会丢失信息,并且（2)它适用于不同类型的任务。

到相机。在 RGB-D 数据的推动下,一种选择是将 LiDAR 点云投影到
相机平面并渲染 2.5D 稀疏深度。但是,这种转换在几何上是有损的。二
深度图上的邻居在 3D 空间中可以彼此远离。这使得相机
view 对于专注于对象/场景几何的任务（例如 3D 对象检测)效果较差。

到激光雷达。大多数最先进的传感器融合方法[54, 68, 23]用
它们对应的相机特征（例如,语义标签、CNN 特征或虚拟点）。然而,
这种相机到激光雷达的投影在语义上是有损的。摄像头和 LiDAR 功能已大幅提升
不同的密度,导致只有不到 5% 的相机特征与 LiDAR 匹配
点（用于 32 通道 LiDAR 扫描仪）。严重放弃相机特征的语义密度
损害模型在面向语义的任务（例如 BEV 地图分割)上的性能。相似的
缺点也适用于潜在空间中更新的融合方法（例如,对象查询） [8, 1]。

到鸟瞰图。我们采用鸟瞰图（BEV)作为融合的统一表示。
这种观点对几乎所有感知任务都很友好,因为输出空间也在 BEV 中。更多的
重要的是,向 BEV 的转换保留了几何结构（来自 LiDAR 特征)和
语义密度（来自相机特征）。一方面,LiDAR 到 BEV 的投影变平
沿高度维度的稀疏 LiDAR 特征,因此不会在
图1a。另一方面,相机到 BEV 投影将每个相机特征像素投射回
3D 空间中的射线（在下一节中详细介绍）,这可能会导致密集的 BEV 特征图
图1c保留了来自摄像机的完整语义信息。

3.2 高效的Camera-to-BEV转换

相机到 BEV 的转换并不简单,因为与每个相机特征相关的深度
像素本质上是模棱两可的。在 LSS [39]和 BEVDet [20, 19] 之后,我们明确预测
每个像素的离散深度分布。然后我们将每个特征像素分散到D个离散点中
沿着相机光线并通过其相应的深度概率重新缩放相关特征
（图3a）。这会生成一个大小为NHW D 的相机特征点云,其中N是
相机和(H, W)是相机特征图大小。这样的 3D 特征点云被量化
x, y轴,步长为r （例如,0.4m）。我们使用 BEV 池化操作来聚合所有
每个 r × r BEV 网格内的特征,并沿 z 轴展平特征。

虽然简单,但 BEV 池出人意料地低效且缓慢,在 RTX 上花费超过 500 毫秒
3090 GPU（而我们模型的其余部分只需要大约 100 毫秒）。这是因为相机功能

点云非常大:对于典型的工作负载*,每帧可能会生成大约 200 万个点,比 LiDAR 特征点云密集两个数量级。为了消除这个效率瓶颈,我们建议通过预计算和间隔减少来优化 BEV 池。

预计算。 BEV 池化的第一步是将相机特征点云中的每个点与一个 BEV 网格相关联。与 LiDAR 点云不同的是,相机特征点云的坐标是固定的(只要相机的内参和外参保持不变,通常是经过适当校准后的情况)。受此启发,我们预先计算了每个点的 3D 坐标和 BEV 网格索引。我们还根据网格索引对所有点进行排序并记录每个点的排名。在推理过程中,我们只需要根据预先计算的等级对所有特征点进行重新排序。这种缓存机制可以将网格关联的延迟从 17ms 降低到 4ms。

间隔减少。网格关联后,同一 BEV 网格内的所有点在张量表示中将是连续的。 BEV 池化的下一步是通过一些对称函数(例如,平均值、最大值和总和)聚合每个 BEV 网格内的特征。如图3b 所示,现有实现[39]首先计算所有点的前缀和,然后减去索引变化的边界处的值。然而,前缀和运算需要在GPU 上进行树缩减,并产生许多未使用的部分和(因为我们只需要边界上的那些值),这两者都是低效的。为了加速特征聚合,我们实现了一个直接在 BEV 网格上并行化的专用 GPU 内核:我们为每个网格分配一个 GPU 线程,计算其间隔总和并将结果写回。该内核消除了输出之间的依赖关系(因此不需要多级树缩减)并避免将部分和写入 DRAM,将特征聚合的延迟从 500 毫秒减少到 2 毫秒(图3c)。

外卖。使用我们优化的 BEV 池,相机到 BEV 的转换速度提高了40 倍:延迟从超过 500 毫秒减少到 12 毫秒(仅占我们模型端到端运行时间的 10%),并且可以很好地跨不同的特征分辨率进行扩展(图3d)。这是在共享 BEV 表示中统一多模态感官特征的关键推动因素。我们的两个同时进行的工作也确定了仅相机 3D 检测中的效率瓶颈。他们通过假设均匀的深度分布[63]或截断每个 BEV网格[20]内的点来近似视图变换器。相比之下,我们的技术是精确的,没有任何近似,同时仍然更快。

3.3 全卷积融合

将所有感官特征转换为共享的 BEV 表示后,我们可以轻松地将它们与元素运算符(例如连接)融合在一起。尽管在同一个空间中,由于视图转换器中的深度不准确,LiDAR BEV 特征和相机 BEV 特征仍然可能在一定程度上在空间上错位。为此,我们应用了基于卷积的 BEV 编码器(带有一些残差块)来补偿这种局部失准。我们的方法可能会受益于更准确的深度估计(例如,监督具有真实深度的视图变换器[43, 38]),我们将其留给未来的工作。

3.4 多任务头

我们将多个特定于任务的头应用于融合的 BEV 特征图。我们的方法适用于大多数3D 感知任务。我们展示了两个示例:3D 对象检测和 BEV 地图分割。

检测。我们使用特定类别的中心热图头来预测所有对象的中心位置,并使用一些回归头来估计对象大小、旋转和速度。我们向读者推荐以前的 3D 检测论文[1, 67, 68]以了解更多详细信息。

分割。不同的地图类别可能重叠(例如,人行横道是可驾驶空间的子集)。
因此,我们将此问题表述为多个二元语义分割,每个类别一个。
我们按照 CVT [70]用标准焦点损失[29] 训练分割头。

4个实验

我们评估 BEVFusion 在 3D 对象检测和 BEV 地图分割上的相机-LiDAR 融合,涵盖面向几何和语义的任务。我们的框架可以很容易地扩展*N = 6, (H, W) = (32, 88)和D = (60 − 1)/0.5 = 118。这对应于六个多视图相机,每个相机与

一个32 ×88相机特征图(从256×704图像下采样8×)。按照 BEVDet [20] ,深度离散为 [1, 60] 米,步长为 0.5 米。

表 1:BEVFusion 在 nuScenes （val和
测试)没有花里胡哨。它打破了将相机功能装饰到
LiDAR 点云,提供至少 1.3% 的 mAP 和 NDS,计算量降低1.5-2 倍
成本。（ ：我们的重新实现； † :带测试时间增强（TTA）； ‡ ：带模型合奏和 TTA)

模态	mAP (test)	NDS (test)	mAP (val)	NDS (val)	MACs (G)	Latency (ms)
贝维德[20]	C	42.2 †	48.2 †	–	–	–	–
M2BEV [63]	C	42.9	47.4	41.7	47.0	–	–
BEV前[25]	C	44.5	53.5	41.6	51.7	–	–
BEVDet4D [19]	C	45.1 †	56.9 †	–	–	–	–
点柱[21]	大号	–	–	52.3	61.3	65.5	34.4
第二[65]	大号	52.8	63.3	52.6	63.0	85.0	69.8
中心点[67]	大号	60.3	67.3	59.6	66.8	153.5	80.7
点画[54]	C+L	–	–	65.8*	69.6*	370.0	185.8
PointAugmenting [55]	C+L	66.8 †	71.0 †	–	–	408.5	234.4
MVP [68]	C+L	66.4	70.5	66.1*	70.0*	371.7	187.1
融合绘画[64]	C+L	68.1	71.6	66.5	70.7	–	–
自动对齐[12]	C+L	–	–	66.6	71.1	–	–
FUTR3D [8]	C+L	–	–	64.5	68.3	1069.0	321.4
转融合[1]	C+L	68.9	71.6	67.5	71.3	485.8	156.6
BEVFusion （我们的）	C+L	70.2	72.9	68.5	71.4	253.2	119.2
CenterPoint-Fusion	C+R+L	72.4 ‡	74.9 ‡	–	–	–	–
FusionVPE	C+L	73.3 ‡	75.5 ‡	–	–	–	–
BEVFusion （我们的）	C+L	75.9 ‡	76.1 ‡	73.7 ‡	74.9 ‡	–	–

表 2:BEVFusion在 BEV 上的性能优于最先进的多传感器融合方法13.6%
在 nuScenes (val) 上进行地图分割,并在不同类别之间进行一致的改进。

模态	驾驶	Ped。	叉。	人行道	停车线	停车场	分隔线	平均值
常[46]	C	74.0	35.3	45.9	27.5	35.9	33.9	42.1
LSS [39]	C	75.4	38.8	46.3	30.3	39.1	36.5	44.4
无级变速器[70]	C	74.3	36.8	39.9	25.8	35.0	29.4	40.2
M2BEV [63]	C	77.2	–	–	–	–	40.5	–
BEVFusion （我们的）	C	81.7	54.8	58.4	47.4	50.7	46.4	56.6
点柱[21]	大号	72.0	43.1	53.1	29.7	27.7	37.5	43.8
中心点[67]	大号	75.6	48.4	57.5	36.5	31.7	41.9	48.6
点画[54]	C+L	75.9	48.5	57.1	36.9	34.5	41.9	49.1
MVP [68]	C+L	76.1	48.7	57.0	36.9	33.0	42.2	49.0
BEVFusion （我们的）	C+L	85.5	60.5	67.6	52.0	57.0	53.7	62.7

支持其他类型的传感器（如雷达和基于事件的相机)和其他 3D 感知
任务（例如 3D 对象跟踪和运动预测）。

模型。我们使用 Swin-T [32]作为我们的图像主干,并使用 VoxelNet [65]作为我们的 LiDAR 主干。
我们应用 FPN [28]来融合多尺度相机特征,以生成 1/8 输入大小的特征图。
我们将相机图像下采样到256×704 ,并将 LiDAR 点云体素化为 0.075m（对于
检测)和0.1m（用于分割）。由于检测和分割任务需要 BEV 特征
具有不同空间范围和大小的地图,我们之前应用了双线性插值的网格采样
每个特定任务的头在不同的 BEV 特征图之间显式转换。

训练。与冻结相机编码器的现有方法[54, 55, 1] 不同,我们训练整个
以端到端的方式进行建模。我们同时应用图像和 LiDAR 数据增强来防止
过拟合。使用权重衰减为10－2的 AdamW [34]进行优化。

数据集。我们在 nuScenes [2]上评估我们的方法,这是一个发布的大型户外数据集
在CC BY-NC-SA 4.0下执照。它具有多种注释以支持各种任务（例如
3D 对象检测/跟踪和 BEV 地图分割）。 40,157 个带注释的样本中的每一个
包含六个具有 360 度 FoV 和 32 光束 LiDAR 扫描的单目相机图像。

图 4:BEVFusion 在 3D 对象检测和 BEV 地图分割上的定性结果。它
准确识别远处和小物体（上)并解析拥挤的夜间场景（下）。

4.1 3D物体检测

我们首先在以几何为中心的 3D 对象检测基准上进行实验,其中 BEVFusion
以更低的计算成本和测量的延迟实现卓越的性能。

环境。我们使用 10 个前景类和 nuScenes 的平均精度 (mAP)
检测分数（NDS)作为我们的检测指标。我们还测量了单推理#MACs 和
所有开源方法在 RTX3090 GPU 上的延迟。我们使用单一模型,没有任何
验证和测试结果的测试时间增加。

结果。如表1 所示,BEVFusion 在 nuScenes 检测上取得了最先进的结果
基准测试,在桌面 GPU 上具有接近实时(8.4 FPS)的推理速度。和....相比
TransFusion [1]、BEVFusion 在测试集分 mAP 和 NDS方面实现了 1.3% 的改进,同时显着降低了1.9 倍的 MAC
和1.3 倍的测量延迟。 BEVFusion 还比较
优于具有代表性的点级融合方法 PointPainting [54]和MVP [68]
测试集上1.6 倍加速， 1.5 倍MAC 减少和3.8%更高的 mAP 。我们认为，
BEVFusion 的效率增益来自我们选择 BEV 空间作为共享融合的事实
空间,它充分利用了所有相机功能,而不仅仅是 5% 的稀疏集。因此,BEVFusion
可以使用更小的 MAC 实现相同的性能。结合高效的 BEV
第3.2节中的池化算子， BEVFusion 将 MAC 的减少转化为测量的加速比。

4.2 BEV地图分割

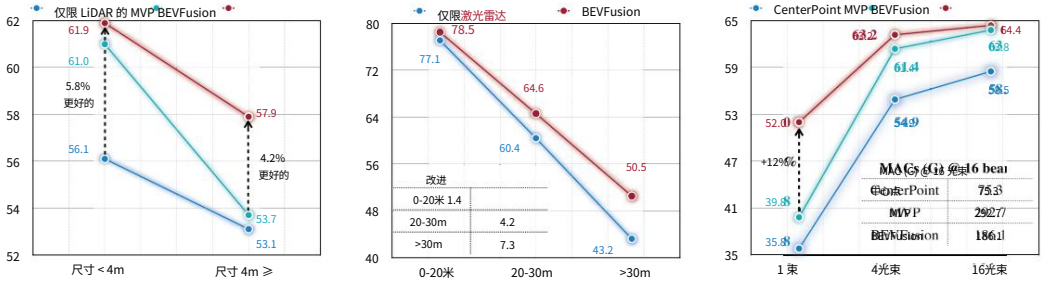
我们进一步将 BEVFusion 与以语义为中心的最先进的 3D 感知模型进行比较
BEV 地图分割任务,其中 BEVFusion 实现了更大的性能提升。

环境。我们报告了 6 个背景类（可驱动空间、
人行横道、人行道、停车线、停车场和车道分隔线)和班级平均
平均 IoU 作为我们的评估指标。由于不同的类别可能有重叠（例如停车场
区域也是可驱动的） ,我们分别评估每个类的二进制分割性能
并选择跨不同阈值的最高 IoU [70]。对于每一帧,我们只执行
在[39, 70, 63, 25]之后的自我汽车周围的[-50m, 50m]×[-50m, 50m] 区域中进行评估。在
BEVFusion,我们使用单个模型联合对所有类执行二进制分割
遵循传统方法为每个类训练一个单独的模型。这导致6×
更快的推理和训练。我们复制了所有开源竞争方法的结果。

结果。我们在表 2 中报告了 BEV 图分割结果。与 3D 对象检测相比
这是一个面向几何的任务,地图分割是面向语义的。因此,我们仅使用相机的 BEVFusion 模型比仅使用 LiDAR 的
基线高8-13%。这个观察准确
与表1 中的结果相反,其中最先进的仅使用相机的 3D 检测器表现出色
仅使用 LiDAR 的探测器几乎 20 mAP。我们的仅相机型号提高了

表 3:BEVFusion 在不同的光照和天气条件下都很稳健,显着提升
具有挑战性的雨天和夜间条件下单模态基线（以灰色标记)的性能
情景。 （* :BEVDet-Tiny 和 LSS 的变体,具有更大的主干和视图转换器）

		晴天		下雨		天		夜晚	
模态		mAP	mIoU	mAP	mIoU	mAP	mIoU	mAP	mIoU
中心点	大号	62.9	50.7	59.2	42.3	62.8	48.9	35.4	37.0
BEVDet/LSS	C	32.9	59.0	33.7	50.5	33.7	57.4	13.5	30.8
MVP	C+L	65.9 (+3.0)	51.0 (-8.0)	66.3 (+7.1)	42.9 (-7.6)	66.3 (+3.5)	49.2 (-8.2)	38.4 (+3.0)	37.5 (+6.7)
BEVFusion	C+L	68.2 (+5.3)	65.6 (+6.6)	69.9 (+10.7)	55.9 (+5.4)	68.5 (+5.7)	63.1 (+5.7)	42.8 (+7.4)	43.6 (+12.8)



(a) 不同的物体尺寸 (b) 不同的物距 (c) 不同的激光雷达稀疏度

图 5:BEVFusion 始终优于最先进的单模态和多模态检测器
在不同的 LiDAR 稀疏度、物体尺寸和距 ego car 的物体距离下,尤其是在
更具挑战性的设置（即稀疏的点云、小/远的物体）。

现有的单目 BEV 图分割方法至少减少了12%。在多模态设置中，
我们进一步将单目 BEVFusion 的性能提高了6 mIoU 并达到了>13%
对最先进的传感器融合方法的改进[54, 68]。这是因为两个基线
方法是以对象为中心和面向几何的。 PointPainting [54]只装饰前景
LiDAR 点和 MVP 仅对前景 3D 对象进行增密。这两种方法都没有帮助
分割地图组件。更糟糕的是,这两种方法都假设 LiDAR 应该是
传感器融合的有效方式,根据我们在表2 中的观察,这是不正确的。

5 分析

我们对 BEVFusion 在不同情况下对单模态模型和最先进的多模态模型进行了深入分析。

天气和照明。我们系统地分析了 BEVFusion 在不同情况下的性能
表 3 中的天气和照明条件。在雨天检测物体是一项挑战
由于明显的传感器噪声,仅限 LiDAR 模型。得益于相机传感器的稳健性
在不同的天气下,BEVFusion 将 CenterPoint 提升了10.7 mAP,性能接近
晴天和雨天之间的差距。恶劣的照明条件对两种检测都具有挑战性
和分割模型。对于检测,与 MVP 相比,MVP 的改进要小得多
BEVFusion,因为它需要精确的 2D 实例分割来生成多模态虚拟
积分（MVP）。这在黑暗或曝光过度的场景中可能非常具有挑战性（例如,第二个场景）
图4）。对于分割,即使仅使用相机的 BEVFusion 在
表 2 中的整个数据集,其在夜间的性能要差得多。我们的 BEVFusion 显着
将其性能提升了12.8 mIoU,这甚至比白天的提升还要大，
当相机传感器发生故障时,证明几何线索的重要性。

尺寸和距离。我们还分析了不同物体大小和距离下的性能。
从图5a 中可以看出， BEVFusion 相对于仅 LiDAR 的对应物实现了持续改进
小型和大型对象,而 MVP 仅对大于
4米。这是因为较大的物体通常密度更大,从那些增强的物体中受益较少
多模态虚拟点（MVP）。此外,BEVFusion 为 LiDAR 带来了更大的改进

表 4:用于验证我们的设计选择的消融实验。默认设置以灰色标记。

(a) 形式

地图	NDS	mIoU
L 57.6 C 33.3	64.9	48.6
L+C 66.4	40.2	56.6
	69.5	62.7

(b) 体素大小

地图	NDS	mIoU
0.075	67.1	70.2	60.6
0.1	66.4	69.5	62.7
0.125	65.1	68.6	63.7

(c) 图像尺寸

地图	NDS	mIoU
128×352 64.0	68.2	60.5
256×704 66.4	69.5	62.7
384×1056 67.2	70.0	59.9

(d) FPN 尺寸

地图	NDS	mIoU
1/16	66.2	69.4	62.7
1/8	66.4	69.5	62.7
1/4	66.0	69.2	58.8

(e) 数据扩充

地图	NDS	mIoU
图像 63.8 68.1 激光雷达 65.7		62.3
69.1 66.4 69.5		61.3
两个都		62.7

(f) 图像主干

地图	NDS	mIoU
ResNet50 65.3 68.9 59.2		
SwinT（冻结） 66.1 69.3 52.8		
斯威特	66.4 69.5 62.7	

仅适用于较小物体（图5a)和较远物体（图5b)的模型,两者都是
LiDAR 覆盖不佳,因此可以从密集的相机信息中受益更多。

更稀疏的激光雷达。我们展示了仅 LiDAR 检测器 CenterPoint [67] 的性能，
图5c中不同 LiDAR 稀疏度下的多模态检测器 MVP [68]和我们的 BEVFusion。
BEVFusion 在所有稀疏级别下始终优于 MVP， MACs 减少1.6 倍,并且
在 1-beam LiDAR 场景中实现了12%的改进。 MVP 装饰输入点云
并直接将 CenterPoint 应用于绘制和增密的 LiDAR 输入。所以自然需要
仅 LiDAR 的 CenterPoint 检测器表现良好,这在稀疏 LiDAR 设置下无效
（图5c 中 35.8 NDS,1 光束输入）。相比之下,BEVFusion 融合了多感官信息
在共享的 BEV 空间中,因此不假设一个强大的仅 LiDAR 检测器。

多任务学习。本文着重于设置
分别训练不同的任务。在这里,我们介绍
联合 3D 检测和分割训练的试点研究。
我们将不同任务的损失重新缩放到相同的大小
并为每个任务应用单独的 BEV 编码器以提供
学习更多特定于任务的功能的能力。从
表5,联合训练不同的任务一起有负面影响
影响每个单独任务的性能,即
俗称“负迁移”。分离 BEV 编码器部分缓解了这个问题。一个
更复杂的培训计划可以进一步缩小这一差距,我们将其留给未来的工作。

表 5:联合检测和分割训练（训练 10 个 epoch）。

	NDS	mIoU
仅检测	70.4	–
仅细分	–	58.5
关节（共享 BEV 编码器）	69.7	54.0
关节（单独的 BEV 编码器）	69.9	58.4

消融研究。我们在表4中展示了消融研究,以证明我们的设计选择是合理的,其中我们
对检测器使用更短的训练计划。在表4a 中,我们观察到 BEVFusion 带来了巨大的
仅 LiDAR 检测(+8.8%)和仅相机分割(+6.1%) 的改进。这个
表明共享 BEV 空间中的传感器融合对于面向几何和语义的任务都是有益的。表4b、表4c和表4d表明 BEVFusion
的检测变体
在体素和图像分辨率上都可以很好地扩展,而 BEV 分割性能稳定
当图像分辨率增加到256×704 以上时。我们还在表4d中注意到使用 FPN 特征
从 1/8 输入分辨率为检测和分割提供最佳性能
进一步增加计算没有帮助。表4f表明我们的 BEVFusion 是通用的,并且
适用于不同的骨干。还值得注意的是,通常的做法是冻结
现有多传感器 3D 对象检测研究[54, 55, 1] 中的图像主干没有利用
即使在检测方面也能充分发挥相机特征提取器的潜力,并导致出色的性能
在 BEV 细分中下降(10%) 。我们在表4e中进一步证明了两者的增强
图像和 LiDAR 输入有助于提高 BEVFusion 的性能。

六,结论

我们提出了 BEVFusion,这是一种用于多任务多传感器 3D 感知的高效通用框架。
BEVFusion 将相机和 LiDAR 功能统一在一个共享的 BEV 空间中,完全保留了两者
几何和语义信息。为了实现这一点,我们将慢速相机到 BEV 的转换加速了40 倍以上。 BEVFusion 打破
了长期存在的点级惯例
融合是多传感器感知系统的黄金选择。 BEVFusion 达到最先进的水平
3D 检测和 BEV 地图分割任务的性能,计算量减少1.5-1.9 倍

和现有解决方案的 1.3-1.6 倍测量加速。我们希望 BEVFusion 可以作为一个简单但强大的基线，以激发未来对多任务多传感器融合的研究。

限制。目前，BEVFusion 在联合多任务训练中仍然存在性能下降，这还没有释放出在多任务设置中更大的推理加速潜力。更准确的深度估计[43, 38]也是本文未充分探索的方向，它可能会进一步提高 BEVFusion 的性能。

社会影响。高效准确的多传感器感知对于自动驾驶汽车的安全至关重要。 BEVFusion 将最先进的多传感器融合模型的计算成本降低了一半，并在小而远的物体以及雨天和夜间条件下实现了大幅精度提升。它为安全可靠的自动驾驶铺平了道路。

确认。我们要感谢陈玄耀和周布雷迪对检测和分割评估的指导，以及刘英飞和王天才的有益讨论。

这项工作得到了美国国家科学基金会、现代汽车、高通、英伟达和苹果公司的支持。刘志坚获得了高通创新奖学金的部分支持。

参考

[1]白旭阳, 胡泽宇, 朱新格, 黄清秋, 陈奕伦, 付洪波, 和Chiew-Lan Tai。 TransFusion：使用 Transformers 进行 3D 对象检测的稳健 LiDAR-Camera Fusion。在CVPR， 2022年。3、4、5、6、7、9 _

[2] Holger Caesar、Varun Bankiti、Alex H. Lang、Sourabh Vora、Venice Erin Liong、Qiang Xu、Anush Krishnan、 Yu Pan、Giancarlo Baldan 和 Oscar Beijbom。 nuScenes：自动驾驶的多模式数据集。在 CVPR，2020 年。6

[3]蔡兆伟和努诺·瓦斯康塞洛斯。 Cascade R-CNN：深入研究高质量目标检测。在 CVPR, 2018. 3

[4]陈汉生、王碧超、王凡、田伟、鲁雄、郝丽。 EPro-PnP：用于单目物体姿态估计的广义端到端概率透视 n 点。 CVPR, 2022. 2 [5] Qi Chen, Lin Sun, Zhixin Wang, Kui Jia, and Alan Yuille。对象作为热点：无锚的 3D 对象

通过热点触发的检测方法。在 ECCV，2020 年。2

[6] Qi Chen、Sourabh Vora 和 Oscar Beijbom。 PolarStream：流式激光雷达目标检测和用极柱分割。在 NeurIPS，2021 年。2

[7]陈晓志，马慧敏，季万，李伯，夏天下。多视图 3D 对象检测网络自动驾驶。在 CVPR，2017 年。3

[8]陈轩尧，张天元，王悦，王一伦，赵航。 FUTR3D：统一传感器用于 3D 检测的融合框架。 arXiv, 2022. 3, 4, 6

[9]陈玉康，李彦伟，张翔宇，孙健，贾家亚。用于 3D 对象检测的焦点稀疏卷积网络。在 CVPR，2022. 3

[10] 陈一伦，刘舒，沉晓勇，贾家亚。快速点 R-CNN。在 ICCV，2019. 2 [11]陈泽辉，李振宇，张世全，方良基，姜庆红，赵峰。Graph-DETR3D：重新思考多视图 3D 对象检测的重叠区域。在 ACM-MM, 2022. 2 [12]陈泽辉、李振宇、张世全、方良基、姜庆红、赵峰、周伯雷和赵航。 AutoAlign：用于多模式 3D 对象检测的像素实例特征聚合。 arXiv, 2022. 3, 6

[13] Christopher Choy、JunYoung Gwak 和 Silvio Savarese。 4D 时空卷积网络：Minkowski卷积神经网络。在 CVPR，2019 年。2

[14]吕凡，熊轩，王峰，王乃燕，张兆祥。 RangeDet：捍卫范围查看基于 LiDAR 的 3D 对象检测。在ICCV，2021. 2

[15]葛润洲，丁壮壮，胡一涵，邵文新，黄丽，李坤，刘强。 2021 年Waymo开放数据集挑战赛实时 3D 检测和最有效模型的第一名。在CVPRW，2021 年。2

[16] Georgia Gkioxari、Ross Girshick、Piotr Dollár 和 Kaiming He。检测和识别人与物体的交互。在 CVPR，2018年。3

[17]本杰明·格雷厄姆、马丁·恩格尔克和劳伦斯·范德马腾。使用子流形稀疏卷积网络进行3D 语义分割。在 CVPR，2018 年。2

[18] Kaiming He、Georgia Gkioxari、Piotr Dollár 和 Ross Girshick。掩码 R-CNN。在ICCV，2017. 3 [19]黄俊杰和黄冠。BEVDet4D：利用多相机 3D 对象检测中的时间线索。arXiv, 2022. 2, 4, 6

[20]黄俊杰，黄冠，郑朱，叶云，杜大龙。 BEVDet：高性能多鸟瞰视图中的相机 3D 对象检测。 arXiv, 2021. 2, 4, 5, 6

[21] Alex H. Lang、Sourabh Vora、Holger Caesar、Lubing Zhou 和 Jiong Yang。 PointPillars：快速编码器
用于点云的目标检测。在 CVPR, 2019. 2, 6
[22]齐立，王悦，王一伦，赵航。 HDMapNet：在线高清地图构建和评估框架。在 ICRA, 2022. 2 [23]李英伟, Adams Wei Yu, Tianjian
Meng, Ben Caine, Jiquan Ngiam, Daiyi Peng, Junyang Shen, Bo Wu, Yifeng Lu, Denny Zhou, et al. DeepFusion：用于多
模态 3D 对象检测的激光雷达-相机深度融合。在 CVPR, 2022. 1, 3, 4

[24]李志超，王峰，王乃燕。 LiDAR R-CNN：一种高效且通用的 3D 对象
探测器。简历，2021. 2
[25]李志奇，王文海，李洪阳，谢恩泽，司马崇浩，童璐，余乔，戴继峰。
BEVFormer：通过时空变换器从多相机图像中学习鸟瞰图表示。 arXiv, 2022. 2, 3, 6, 7

[26]梁明、杨斌、陈云、胡瑞、拉奎尔·乌尔塔松。用于 3D 的多任务多传感器融合
物体检测。在 CVPR, 2019. 3
[27]梁明，杨斌，王神龙，拉奎尔·乌尔塔松。多传感器深度连续融合
3D 物体检测。在 ECCV，2018. 3
[28] Tsung-Yi Lin、Piotr Dollár、Ross Girshick、Kaiming He、Bharath Hariharan 和 Serge Belongie。特征
用于对象检测的金字塔网络。在 CVPR，2017 年。6
[29] Tsung-Yi Lin、Priya Goyal、Ross Girshick、Kaiming He 和 Piotr Dollár。密集物体检测的焦点损失。在 ICCV，2017. 5 [30]刘英飞，
王天才，张翔宇，孙健。 PETR：用于多视图 3D 对象检测的位置嵌入转换。 arXiv, 2022. 2 [31]刘英飞、闫俊杰、范佳、李帅林、高启、
王天才、张翔宇、孙健。

PETRv2：多摄像头图像 3D 感知的统一框架。 arXiv，2022. 2
[32]刘泽，林宇彤，曹悦，韩虎，魏以轩，张正，林史蒂芬，郭白宁。 Swin Transformer：使用 Shifted Windows 的分层视觉转换器。载于
ICCV，2021. 6 [33]刘志健，唐昊天，赵胜宇，邵凯文，宋涵。 PVNAS：3D 神经架构

使用点体素卷积进行搜索。 TPAMI, 2021. 2
[34] 伊利亚·洛希洛夫和弗兰克·哈特。解耦权重衰减正则化。载于 ICLR，2019. 6 [35]陆家辰，周哲元，朱夏天，徐航，张丽。学习
Ego 3D 表示作为光线追踪。 arXiv, 2022. 3 [36] Ramin Nabati 和 Hairong Qi。 CenterFusion：用于 3D 对象检测的基于中心的雷
达和相机融合。在 WACV, 2021. 3 [37]潘博文、孙建凯、何寅 Tiga Leung、Alex Andonian 和 Bolei Zhou。跨视图语义

感知环境的分割。 RA-L, 2020. 2
[38] Dennis Park、Rares Ambrus、Vitor Guizilini、Jie Li 和 Adrien Gaidon。是否需要伪激光雷达
单目 3D 物体检测？在ICCV, 2021. 5, 10
[39] Jonah Philion 和 Sanja Fidler。 Lift、Splat、Shoot：对来自任意摄像机装置的图像进行编码
隐式不投影到 3D。在ECCV， 2020年。2、4、5、6、7
[40] Charles R Qi, Wei Liu, Chenxia Wu, Hao Su, and Leonidas J Guibas.用于从 RGB-D 数据进行 3D 对象检测的
Frustum PointNets 。在 CVPR，2018年。3
[41]查瑞中泰齐、李毅、苏浩、列奥尼达斯·J·吉巴斯。 PointNet++：深度层次特征
学习度量空间中的点集。在 NeurIPS，2017 年。2
[42] Charles R Qi、Yin Zhou、Mahyar Najibi、Pei Sun、Khoa Vo、Boyang Deng 和 Dragomir Anguelov。
点云序列的外接 3D 对象检测。在 CVPR，2021 年。2
[43] Cody Reading、Ali Harakeh、Julia Chae 和 Steven L. Waslander。分类深度分布网络
用于单目 3d 物体检测。在 CVPR, 2021. 2, 5, 10
[44]任少卿、何凯明、罗斯·吉尔希克、孙健。 Faster R-CNN：迈向实时对象
使用区域建议网络进行检测。在 NeurIPS，2015 年。3
[45]托马斯·罗迪克和罗伯托·西波拉。使用Pyramid Occupancy Networks从图像中预测语义地图表示。在 CVPR，2020 年。2 [46]
Thomas Roddick、Alex Kendall 和 Roberto Cipolla。单目 3D 的正交特征变换

物体检测。在 BMVC, 2019. 2, 6
[47]施绍帅、郭朝旭、李江、王哲、施建平、王小刚、李洪生。
PV-RCNN：用于 3D 对象检测的点体素特征集抽象。 In CVPR, 2020. 2 [48]施绍帅、姜丽、邓家军、王哲、郭朝旭、施
佳平、王晓刚、李洪生。 PV-RCNN++：用于3D 对象检测的具有局部向量表示的点体素特征集抽象。 arXiv, 2021. 2

[49]施少帅，王晓刚，李宏生。 PointRCNN：来自点云的 3D 对象建议生成和检测。在 CVPR，2019 年。2

[50]施绍帅、王哲、施建平、王小刚、李洪生。从点到部分：使用部分感知和部分聚合网络从点云进行3D对象检测。 TPAMI，2020 年。2

[51]孙科、肖斌、刘东、王京东。用于人体姿势估计的深度高分辨率表示学习。 CVPR, 2019. 3 [52]唐昊天, 刘志健, 赵胜宇, 林玉军, 吉林, 王涵瑞, 宋涵。使用稀疏点体素卷积搜索高效的 3D 架构。载于 ECCV,2020. 2 [53]田志,沉春华,陈昊,何佟。 FCOS:完全卷积的单阶段目标检测。

在ICCV,2019. 2

[54] Sourabh Vora、Alex H Lang、Bassam Helou 和 Oscar Beijbom。 PointPainting:3D 的顺序融合物体检测。在CVPR， 2020年。1、3、4、6、7、8、9 _ _

[55]王春伟,马超,朱明,杨小康。 PointAugmenting:用于 3D 对象检测的跨模态增强。 In CVPR, 2021. 1, 3, 6, 9 [56]王京东, 孙科, 程天恒, 姜博瑞, 邓超瑞, 赵洋, 刘东, 穆亚东,谭明奎, 王兴刚, 刘文宇, 斌肖。用于视觉识别的深度高分辨率表示学习。TPAMI, 2019. 3

[57]王太,朱新格,庞江淼,林大华。 FCOS3D:全卷积一级单目 3D 对象检测。在 ICCVW,2021. 2

[58]王太,朱新格,庞江淼,林大华。概率和几何深度:透视检测对象。在 CoRL,2021. 2

[59]王悦、Vitor Guizilini、Tianyuan Zhang、Yilun Wang、Hang Zhao 和 Justin M. Solomon。 DETR3D:通过 3D 到 2D 查询从多视图图像中检测 3D 对象。在 CoRL,2021 年。2 [60]王悦和贾斯汀 M. 所罗门。对象 DGCNN:使用动态图的 3D 对象检测。在

NeurIPS, 2021. 2

[61]王英明,张翔宇,杨童,孙健。 Anchor DETR:Transformer 的查询设计基于检测器。在 AAAI,2022. 2

[62]王志新、奎佳． Frustum ConvNet:滑动 Frustums 以聚合Amodal 3D 对象检测的局部逐点特征。在 IROS,2019 年。3

[63]谢恩泽、余志丁、周大全、乔纳·菲利昂、阿尼玛·阿南德库玛、桑贾·菲德勒、罗平和何塞·M·阿尔瓦雷斯。 M2BEV:具有统一鸟瞰图表示的多相机联合 3D 检测和分割。 arXiv, 2022. 2, 3, 5, 6, 7 [64]徐绍庆,周定富,方金,尹俊波,周斌,张良军。 FusionPainting:具有自适应注意的多模态融合,用于 3D 对象检测。 In ITSC, 2021. 3, 6 [65] Yan Yan, Yuxing Mao, and Bo Li.第二:稀疏嵌入卷积检测。传感器,2018 年。

2.6 _

[66]杨泽彤,孙亚楠,刘树,贾家亚。 3DSSD:基于点的 3D 单级物体检测器。CVPR, 2020. 2

[67]尹天维、周兴一和菲利普·克雷恩布尔。基于中心的 3D 对象检测和跟踪。 CVPR , 2021. 2, 5, 6, 9 [68] Tianwei Yin, Xingyi Zhou, and Philipp Krähenbühl。多模态虚拟点 3D 检测。在 NeurIPS 中，

2021. 1, 3, 4, 5, 6, 7, 8, 9

[69]张云鹏,朱正,郑文昭,黄俊杰,黄冠,周杰,陆继文。BEVerse:用于以视觉为中心的自动驾驶的鸟瞰视图中的统一感知和预测。 arXiv, 2022. 3

[70] Brady Zhou 和 Philipp Krähenbühl。用于实时地图视图语义分割的跨视图转换器。 In CVPR, 2022. 2, 3, 5, 6, 7 [71] Yin Zhou, Pei Sun, Yu Zhang, Dragomir Anguelov, Jiyang Gao, Tom Ouyang, James Guo, Jiquan Ngiam, and Vijay Vasudevan。用于激光雷达点云中 3D 对象检测的端到端多视图融合。

科尔,2019. 2

[72]殷周和奥塞尔图泽尔。 VoxelNet:基于点云的 3D 对象检测的端到端学习。在 CVPR,2018 年。2

[73]朱本进,姜正凯,周祥新,李泽明,于刚。类平衡分组和点云 3D 对象检测的采样。 arXiv, 2019. 2

[74]朱喜洲,苏伟杰,卢乐伟,李斌,王小刚,戴继峰。 Deformable DETR:用于端到端对象检测的可变形变压器。载于 ICLR,2021. 2

[75]朱新格、周惠、王泰、洪方舟、马跃新、李伟、李宏生和林大华。

用于 LiDAR 分割的圆柱形和非对称 3D 卷积网络。在 CVPR,2021 年。2