BEV 学术界和工业界方案、优化方法与tricks综述

极市平台 2022-09-19 22:01:01 发表于广东 手机阅读

以下文章来源于自动驾驶之心 ，作者汽车人

作者丨汽车人

来源丨自动驾驶之心

编辑丨极市平台

极市导读

本文回顾了关于BEV感知的最新工作，并对不同解决方案进行了深入分析。此外，还描述了行业中BEV方法的几个系统设计，介绍了一整套实用指南，以提高BEV感知任务的性能，包括相机、激光雷达和融合输入。最后，论文指出了该领域未来的研究方向，希望本报告能为领域提供一些信息，并鼓励更多关于BEV感知的研究工作 >>视觉AI工程项目实训周第2期开启招募——手把手教你实现模型开发到落地

学习感知任务的鸟瞰图（BEV）中的强大表示法是一种趋势，并引起了工业界和学术界的广泛关注。大多数自动驾驶常规方法是在前视图或透视图中执行检测、分割、跟踪等。随着传感器配置变得越来越复杂，集成来自不同传感器的多源信息并在统一视图中表示特征变得至关重要。BEV perception继承了几个优势，如在BEV中表示周围场景直观且融合友好；并且在BEV中表示对象对于后续模块最为理想，如在规划和/或控制中。BEV感知的核心问题在于： （a） 如何通过从透视图到BEV的视图转换来重建丢失的3D信息； （b） 如何在BEV网格中获取GT； （c） 如何制定pipelines，以纳入来自不同来源和view的特征； （d） 如何适应和推广算法，因为传感器配置在不同场景中有所不同；

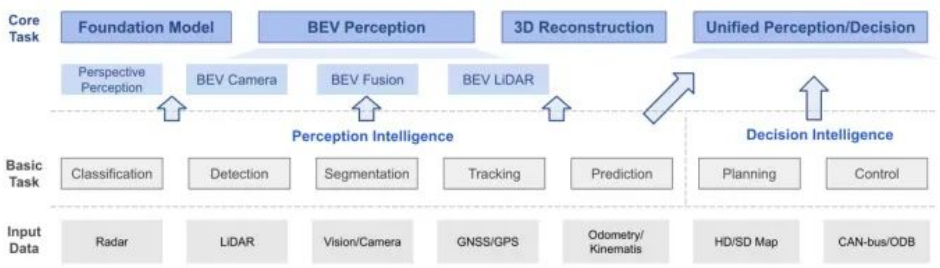
本调查回顾了关于BEV感知的最新工作，并对不同解决方案进行了深入分析。此外，还描述了行业中BEV方法的几个系统设计，介绍了一整套实用指南，以提高BEV感知任务的性能，包括相机、激光雷达和融合输入。最后，论文指出了该领域未来的研究方向，希望本报告能为社区提供一些信息，并鼓励更多关于BEV感知的研究工作。

领域介绍

自动驾驶中的感知识别任务本质上是对物理世界的三维几何重建。随着传感器的多样性和数量越来越复杂，自动驾驶系统的装备也越来越复杂，以统一的视角表示不同视图中的特征至关重要。众所周知的鸟瞰图（BEV）是一种自然而直接的候选视图，可作为统一表示。与二维视觉

领域中广泛研究的前视图或透视图相比，BEV表示具有若干固有优点。首先，它没有2D任务中普遍存在的遮挡或缩放问题。可以更好地解决具有遮挡或交叉交通的车辆识别问题。此外，以这种形式表示对象或道路元素将有利于方便后续模块（如规划、控制）的开发和部署。

基于输入数据，论文将BEV感知研究主要分为三个部分：BEV camera、BEV激光雷达和BEV fusion，下图描述了BEV感知家族的总体图，具体地，BEV camera指示用于从多个环绕相机检测或分割3D目标的视觉或以视觉为中心的算法；BEV激光雷达描述了从点云输入的检测或分割任务；BEV fusion描述了来自多个传感器输入的融合机制，如相机、激光雷达、GNSS、里程计、高清地图、CAN总线等；



当谈到BEV感知研究的动机时，需要检查三个重要方面。

1.意义

BEV感知是否会对学术界和/或社会产生真正和有意义的影响？众所周知，与基于激光雷达或融合的解决方案相比，基于视觉的解决方案存在巨大的性能差距，例如，截至2022年8月提交时，仅视觉与激光雷达之间的第一排名方法差距超过了nuScenes数据集上NDS的20%，Waymo基准的差距甚至超过30%。这自然促使我们研究视觉解决方案是否能够超越或等同于激光雷达方法。从学术角度来看，设计基于camera的pipelines以使其优于激光雷达的本质在于更好地理解从2D外观输入到3D几何输出的视图转换过程。如何像在点云中那样将相机特征转换为几何表示，对学术界产生了有意义的影响。从工业角度考虑，将一套激光雷达设备纳入SDV的成本很高。此外基于camera的pipelines可以识别长距离物体和基于颜色的道路元素（如交通灯），这两种激光雷达方法都无法实现。

2.空间

BEV感知中是否存在需要大量创新的开放性问题？BEV感知背后的要点是从camera和激光雷达输入中学习鲁棒和可概括的特征表示，这在激光雷达分支中很容易，因为输入（点云）具有这样的3D特性。在相机分支中，这是非常重要的，因为从单目或多视图设置中学习3D空间信息是困难的。虽然看到有人试图通过姿势估计[9]或时间运动[10]来学习更好的2D-3D对应关系，但BEV感知背后的核心问题需要从原始传感器输入进行深度估计的实质性创新，特别是对于相机分支。另一个关键问题是如何在pipelines的早期或中期融合特征，大多数传感器融合算法将该问题视为简单的对象级融合或沿blob channel的朴素特征连接。这可能解释了为什么由于相机和激光雷达之间的未对准或不准确的深度预测，某些融合算法表现不如仅使用激光雷达的解决方案。如何对齐和集成多模态输入的特征起着至关重要的作用，从而为创新留下了广阔的空间。

论文主要回顾了近年来BEV感知研究的全貌，详细阐述了BEV感知文献的综合分析，涵盖了深度估计、视图转换、传感器融合、域自适应等核心问题。介绍并讨论了几种重要的BEV感知工业系统级设计。除了理论贡献外，我们还提供了一本实用的操作指南，用于提高各种BEV感知任务的性能。

数据集和Metrics

1.数据集

论文介绍了一些流行的自动驾驶数据集和常用的评估指标。下表总结了BEV感知的主要基准统计数据。通常，数据集由各种场景组成，每个场景在不同的数据集中具有不同的长度。总持续时间从几十分钟到几百小时不等。对于BEV感知任务，3D边界框标注和3D分割标注至关重要，高清地图配置已成为主流趋势，其中大部分可以用于不同的任务。

Dataset	Year	City	Sensor Data				Annotation				HD-Map	Other Data	# Subm.
			Scenes	Hours	Scans	Images	Frames	3D bbox	3D lane	3D seg.			
KITTI [11]	2012	EU	22	1.5	15k	15k	15k	80k	-	-	✗	-	380+
Waymo [8]	2019	NA	1150	6.4	230k	12M	230k	12M	-	50k	✗	-	200+
nuScenes [7]	2019	NA/AS	1000	5.5	390k	1.4M	40k	1.4M	-	40k	✓	CAN-bus	350+
Argo v1 [24]	2019	NA	113	0.6	22k	490k	22k	993k	-	-	✓	-	100+
Argo v2 [12]	2022	NA	1000	4	150k	2.7M	1.5M	†	-	-	✓	-	10+
ApolloScape [25]	2018	AS	103	2.5	29k	144k	144k	70k	†	-	✓	-	200+
OpenLane [26]	2022	NA	1000	6.4	1000	200k	2.2M	-	8.8M	-	✗	-	-
Lyft L5 [27]	2019	AF	366	2.5	46k	240k	46k	1.3M	-	-	✗	-	500+
A* 3D [28]	2019	AS	39k	55	39k	39k	39k	230k	-	-	✗	-	-
H3D [29]	2019	NA	160	0.8	27k	83k	27k	1.1M	-	-	✗	-	-
SemanticKITTI [30]	2019	EU	22	1.2	43k	-	-	-	-	43k	✗	-	30+
A2D2 [31]	2020	EU	†	†	12.5k	41k	12.5k	43k	-	41k	✗	-	-
Cityscapes 3D [32]	2020	-	†	2.5	-	5k	5k	40k	-	-	✗	IMU/GPS	400+
PandaSet [33]	2020	NA	179	†	16k	41k	14k	†	-	60k	✗	-	-
KITTI-360 [34]	2020	EU	11	†	80k	320k	80k	68k	-	80k	✗	-	30+
Cirrus [35]	2020	-	12	†	6285	6285	6285	†	-	-	✗	-	-
ONCE [36]	2021	AS	1M	144	1M	7M	15k	417k	-	-	✗	-	-
AIODrive [37]	2021	Sim	100	2.8	100k	1M	100k	26M	-	100k	✓	-	-
DeepAccident [38]	2022	Sim	464	†	131k	786k	131k	1.8M	-	131k	✓	-	-

2.Metrics

LET-3D-APL：在仅camera的3D检测中，使用LET-3D-APL代替3D-AP作为度量。与三维联合交集（IoU）相比，LET-3D-APL允许预测边界框的纵向定位误差达到给定公差。LET-3D-APL通过使用定位亲和力缩放精度来惩罚纵向定位误差。LET-3D-APL的定义在数学上定义为：

$$\text{LET-3D-APL} = \int_0^1 p_L(r)dr = \int_0^1 \overline{a}_l \cdot p(r)dr,$$

mAP：类似于2D目标检测中的AP度量，但匹配策略被从IoU替换为BEV平面上的2D中心距离。AP是在不同的距离阈值下计算的：0.5米、1米、2米和4米。通过平均上述阈值中的AP来计算mAP： NDS： nuScenes检测分数（NDS）是几个指标的组合，mAP、mATE（平均平移误差）、mASE（平均标度误差）、mAOE（平均方位误差）、mAVE（平均速度误差）和mAAE（平均属性误差）。通过使用上述度量的加权和来计算NDS。mAP的权重为5，其余为1：

$$\begin{aligned}
\text{TP}_{\text{score}} &= max(1 - \text{TP}_{\text{error}}, 0.0), \\
\text{NDS} &= \frac{5 \cdot \text{mAP} + \sum_{i=1}^{5} \text{TP}^{\text{i}}_{\text{score}}}{10}.
\end{aligned}$$

BEV感知方法

如下表所示，近年来BEV感知文献汇总。在输入模式下，“L”表示激光雷达，“SC”表示单相机，“MC”表示多相机，“T”表示时间信息。在任务下，“ODet”用于3D对象检测，“LDet”用于三维车道检测，“MapSeg”用于地图分割，“Plan”用于运动规划，“MOT”用于多对象跟踪。深度监督意味着仅camera模型使用稀疏/密集深度图来监督模型。在数据集下，“nuS”代表nuScenes数据集，“WOD”代表Waymo开放数据集，“KITTI”代表KITTI数据集，“Lyft”代表Lyft 5级数据

集，“OpenLane”代表OpenLane数据集，“AV”代表Argosse数据集，“Carla”代表Carla模拟器，“SUN”代表SUN RGB-D数据集，“ScanNet”代表ScanNet室内场景数据集。

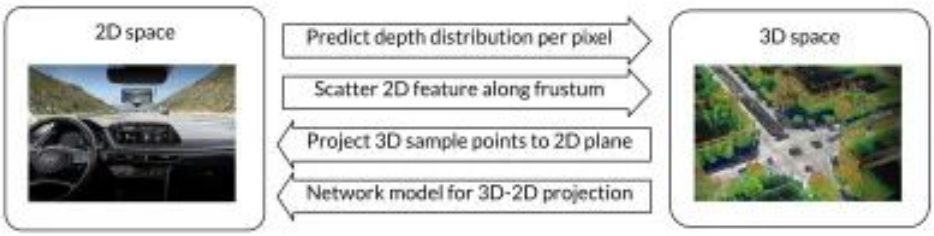
Method	Venue	Input Modality	Task	Depth Supervision	Dataset	Contribution
OFT [42]	*BMVC 2019*	SC	ODet	✗	KITTI	Feature Projection to BEV
VoxelNet [43]	*CVPR 2018*	L	ODet	-	KITTI	Implicit voxel grids transformed to BEV
PointPillars [44]	*CVPR 2019*	L	ODet	-	KITTI	Voxelization with pillars as BEV
CaDDN [45]	*CPVR 2021*	SC	ODet	✓	KITTI/WOD	Depth Distribution with Supervision
DfM [10]	*ECCV 2022*	SC	ODet	✓	KITTI	Motion to Depth to Voxel to BEV
BEVDet [46]	*arXiv 2022*	MC	ODet	✗	nuS	BEV space data augmentation
PETR [47]	*arXiv 2022*	MC	ODet	✗	nuS	Implicit BEV Pos Embed
BEVDepth [48]	*arXiv 2022*	MC/T	ODet	✓	nuS	Depth Correction
ImVoxelNet [49]	*WACV 2022*	SC/MC	ODet	✗	nuS/KITTI SUN/ScanNet	Camera Pose with Grid Sampler to BEV
M^2BEV [3]	*arXiv 2022*	MC	ODet/MapSeg	✗	nuS	BEV Representation without Depth
PolarFormer [50]	*arXiv 2022*	MC	ODet/MapSeg	✗	nuS	Polar-ray Representation in BEV
BEVFormer [4]	*ECCV 2022*	MC/T	ODet/MapSeg	✗	nuS/WOD	Transformer for BEV feature
BEVFusion [5]	*arXiv 2022*	MC/L	ODet/MapSeg	-	nuS	Fusion on BEV from Camera and LiDAR
Cam2BEV [51]	*ITSC 2020*	MC	MapSeg	✗	Synthetic	Homo-graphic Projection to BEV
FIERY [52]	*ICCV 2021*	MC	MapSeg	✗	nuS/Lyft	Future Prediction in BEV space
CVT [53]	*CVPR 2022*	MC	MapSeg	✗	nuS	Camera Intrinsics BEV Projection
HDMapNet [54]	*ICRA 2022*	MC/L/T	MapSeg	-	nuS	Feature Fusion under BEV
Image2Map [55]	*ICRA 2022*	SC/T	MapSeg	✗	nuS/Lyft/AV	Polar-ray Transformer in BEV
LSS [56]	*ECCV 2020*	MC	MapSeg/Plan	✗	nuS/Lfyt	First Depth Distribution
ST-P3 [57]	*ECCV 2022*	MC/T	MapSeg/Plan	✓	nuS/Carla	End to End P3 with Temporal Info
3D LaneNet [58]	*ICCV 2019*	SC	LDet	✗	OpenLane	IPM Projection to BEV space
STSU [59]	*ICCV 2021*	SC	LDet	✗	nuS	Query-based Centerline in BEV
PersFormer [26]	*ECCV 2022*	SC	LDet	✗	OpenLane	Perspective Transformer for BEV

1.基于Camera的BEV

只有camera的3D感知吸引了学术界的大量关注，因为与基于激光雷达的3D感知相比，这是一个未解决的问题，因此值得探索。核心问题是2D图像自然不保留3D信息，因此当从2D图像中不准确地提取深度信息时，难以获得对象的精确3D定位。仅camera的3D感知可分为三个领域：单相机设置、stereo设置和多camera设置，它们有不同的技能来解决深度问题。

视图转换

最近的研究集中于视图转换模块[3、4、10、26、46、47、48、50、55、58]，其中3D信息是根据2D特征或3D先验假设构建的。从二维特征构造三维信息通常表示为深度估计或cost volume。从3D先验假设构造3D信息通常被表示为采样2D特征以通过3D-2D投影映射构造3D特征，视图变换在仅camera 3D感知中起着至关重要的作用，因为它是构建3D信息和编码3D先验假设的主要模块。大体上，它可以分为两个方面，一是利用2D特征构造深度信息并将2D特征“提升”到3D空间，另一个是通过3D到2D投影映射将2D特征编码到3D空间。我们将第一种方法命名为2D-3D，第二种方法称为3D-2D。下图给出了通过这两种方法执行视图转换的概要路线图：



从2D到3D，基于LSS的方法[5、45、46、48、56、63、95]根据2D特征预测每个像素的深度分布，而立体视觉方法[64、96]沿着由成本体积构建的平截头体散布2D特征。

从3D到2D，基于单应矩阵的方法[4，26，47，55，85，112]假定稀疏的3D采样点，并通过摄像机参数将其投影到2D平面。基于纯网络的方法[106、107、108、109、110]使用MLP或transformer隐式建模从3D空间到2D平面的投影矩阵。

LSS[56]引入了2D-3D方法，其中预测2D特征上每个网格的深度分布，然后通过相应的体素空间深度"提升"每个网格的2D特征，并执行基于激光雷达的下游任务方法。这一过程可以表述为：

$$\mathcal{F}_{3D}(x,y,z) = \left[\mathcal{F}^{*}_{2D}(\hat{u},\hat{v}) \otimes \mathcal{D}^{*}(\hat{u},\hat{v})\right]_{xyz},$$

请注意，这与伪激光雷达方法[92、93、94]非常不同，伪激光雷达的深度信息是从预训练的深度估计模型中提取的，过程发生在2D特征提取之前。在LSS[56]之后，还有另一项工作遵循了将深度公式化为按bin-wise分布的相同思想，即CaDDN。CaDDN使用类似的网络来预测深度分布（分类深度分布），将体素空间特征压缩到BEV空间，并在最后执行3D检测。LSS[56]和CaDDN之间的主要区别在于，CaDDN使用深度地面真相来监督其分类深度分布预测，因此，由于具有从2D空间提取3D信息的优越深度网络。

当我们声称"更好的深度网络"时，它实际上是在学习路面和透视图之间在特征级别的隐式投影。这一轨迹来自后续工作，如BEVDet及其时间版本BEVDet4D、BEVDepth、BEVFusion和其它。请注意，在stereo设置中，通过强先验更容易获得深度值/分布，其中一对摄像机（即系统的基线）应该是恒定的。这可以公式化为：

$$\mathcal{D}(u,v) = f \times \frac{b}{d(u,v)},$$

LIGA Stereo和DSGN等立体方法利用了这种强大的先验，并与KITTI排行榜上基于激光雷达的替代方案不相上下。

第二个分支（3D到2D）可以追溯到三十年前，当时逆透视映射（IPM）通过有条件地假设3D空间中的对应点位于水平面上，制定了从3D空间到2D空间的投影。这种变换矩阵可以从相机的内外参数中数学推导。一系列工作[99、100、101、102、103、104、105]应用IPM以预处理或后处理的方式将元素从透视图变换为鸟瞰图。

在视图变换的背景下，OFTNet[42]首先引入了3D-2D方法，即从3D到2D的特征投影，其中将2D特征投影到体素空间（3D空间）。它基于这样的假设：从相机原点到3D空间中的特定点，深度分布沿光线是均匀的。这种假设适用于自动驾驶中的大多数场景，但当涉及起伏道路时，有时会中断。同时，许多BEV地图分割工作[106、107、108、109、110]利用多层感知器或transformer架构[111]来隐式地建模3D-2D投影，而无需摄像机参数。最近，3D-2D几何投影和神经网络的组合变得流行[4，26，47，55，85，112]，受特斯拉发布其感知系统技术路线图[6]的启发。请注意，transformer架构中的交叉注意）

$$\mathcal{F}_{3D}(x,y,z) = CrossAttn(q: P_{xyz}, k\ v: \mathcal{F}^{*}_{2D}(\hat{u},\hat{v})),$$

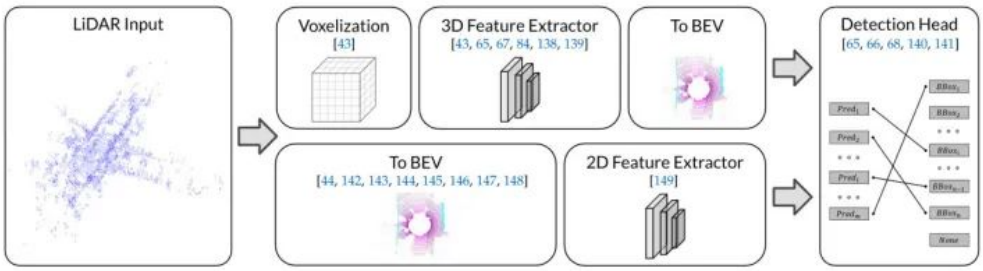
为了获得稳健的检测结果，BEVFormer[4]利用transformer中的交叉关注机制来增强3D-2D视图转换的建模。

BEV和透视法的讨论

在仅camera3D感知的开始，主要焦点是如何从透视图（即2D空间）预测3D对象定位。这是因为2D感知在该阶段得到了很好的发展，如何为2D检测器配备感知3D场景的能力成为主流方法[61、81、82、117、118、119、120、121、122、123、124、125、126、127、128、129]。后来，一些研究达到了BEV表示，因为在这种观点下，很容易解决3D空间中具有相同尺寸的对象由于与相机的距离而在图像平面上具有非常不同的尺寸的问题。这一系列工作[42、45、64、92、96]要么预测深度信息，要么利用3D先验假设来补偿相机输入中3D信息的损失。虽然最近的基于BEV的方法[3、4、5、46、48、95、130]已经风靡了3D感知世界，但值得注意的是，这一成功主要得益于三个方面。第一个原因是nuScenes数据集[7]，它具有多摄像机设置，非常适合在BEV下应用多视图特征聚合。第二个原因是，大多数仅使用相机的BEV感知方法从基于激光雷达的方法[43、44、66、83、84、131、132、133、134、135]中获得了大量帮助，其形式为检测头和相应的损失设计。第三个原因是，单目方法的长期发展[81、82、117、120、121、136、137]使基于BEV的方法蓬勃发展，成为处理透视图中特征表示形式的良好起点。核心问题是如何从2D图像中重建丢失的3D信息。为此，基于BEV的方法和透视方法是解决同一问题的两种不同方法，它们并不相互排斥。

2.基于LiDAR的BEV

在特征提取部分，主要有两个分支将点云数据转换为BEV表示。根据pipilines顺序，将这两个选项分别称为前BEV和后BEV，指示主干网络的输入是来自3D表示还是来自BEV表示。如下图所示，BEV激光雷达感知的一般流程。主要有两个分支将点云数据转换为BEV表示。上分支提取3D空间中的点云特征，提供更准确的检测结果。下分支提取2D空间中的BEV特征（原始点云转换），提供更高效的网络。



BEV前特征提取

除了对原始点云进行基于点的方法处理之外，基于体素的方法将点体素化为离散网格，这通过离散化连续三维坐标提供了更有效的表示。基于离散体素表示、3D卷积或3D稀疏卷积可用于提取点云特征。VoxelNet[43]堆叠多个体素特征编码（VFE）层以编码体素中的点云分布作为体素特征，

PV-RCNN将点和体素分支结合起来，以学习更具辨别力的点云特征。具体而言，高质量的3D提案由体素分支生成，而点分支为提案细化提供额外信息。SA-SSD设计了一个辅助网络，将主干网络中的体素特征转换回点级表示，以明确利用3D点云的结构信息，并减少下采样中的损失。Voxel R-CNN采用3D卷积主干提取点云特征。然后在BEV上应用2D网络以提供目标proposal，这些proposal通过提取的特征进行细化。它实现了与基于点的方法相当的性能。object DGCNN[141]将3D目标检测任务建模为BEV中动态图上的消息传递。在将点云转换为BEV特征图之后，预测查询点迭代地从关键点收集BEV特征。VoTr[139]引入了局部注意力、扩展注意力和快速体素查询，以使大量体素上的注意力机制能够用于大上下文信息。SST[67]将提取的体素特征视为标记，然后在非重叠区域中应用稀疏区域注意和区域移位，以避免对基于体素的网络进行下采样。AFDetV2[68]通过引入关键点辅助监控和多任务头，形成了单级无锚网络。

BEV后特征提取

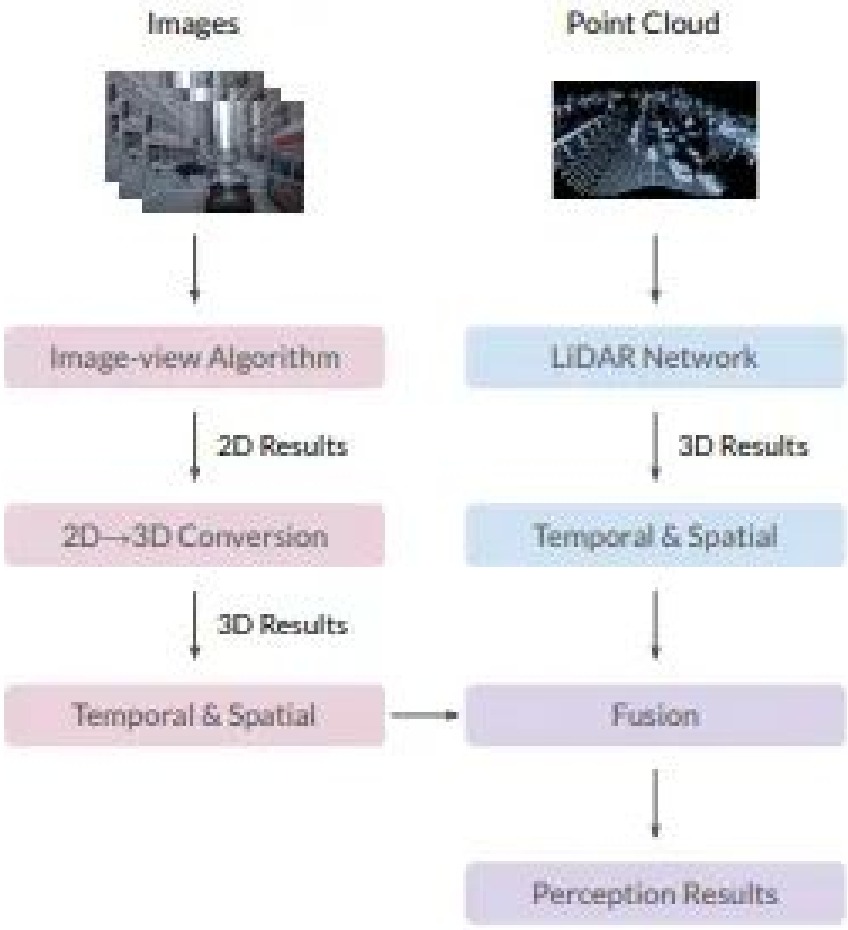
由于三维空间中的体素稀疏且不规则，应用三维卷积是低效的。对于工业应用，可能不支持3D卷积等算子，期望合适和有效的3D检测网络。MV3D[142]是将点云数据转换为BEV表示的第一种方法。在将点离散到BEV网格中之后，根据网格中的点获得高度、强度和密度的特征，以表示网格特征。由于BEV网格中有许多点，因此在此过程中，信息损失相当大。其它工作[143、144、145、146、147、148]遵循类似模式，使用BEV网格中的统计数据表示点云，例如最大高度和强度平均值。PointPillars[44]首先介绍了柱的概念，这是一种具有无限高度的特殊类型的体素。它利用PointNet[131]的简化版本来学习柱中点的表示。然后，编码特征可以由标准2D卷积网络和检测头处理。尽管点柱的性能不如其他3D主干网令人满意，但其及其变体具有高效率，因此适合于工业应用。

一些讨论

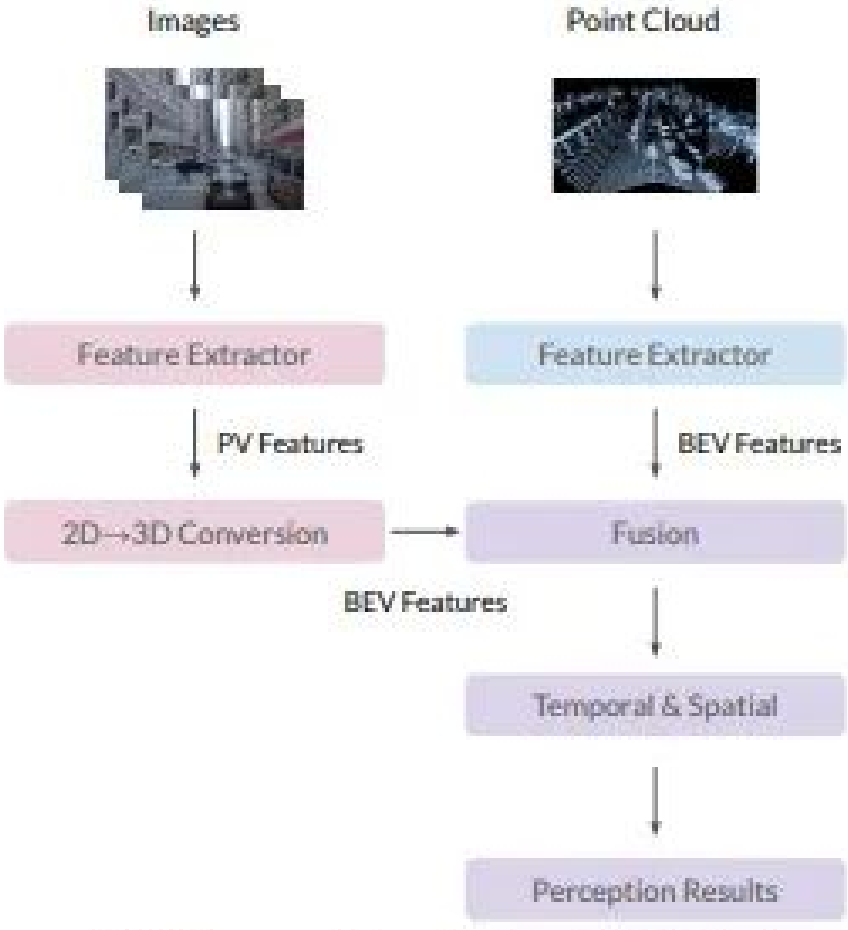
点云数据由神经网络直接处理，在连续3D空间中计算点之间的邻域关系，这带来了额外的时间消耗并限制了神经网络的感受域。最近的工作[43，84]利用离散网格来表示点云数据；采用卷积运算提取特征。然而，将点云数据转换为任何形式的表示不可避免地会导致信息丢失。BEV前特征提取中的现有技术方法利用具有细粒度大小的体素，保留了点云数据中的大部分3D信息，因此有利于3D检测，作为一种权衡，它需要高内存消耗和计算成本。将点云数据直接转换为BEV表示避免了3D空间中的复杂操作。当高度维度被压缩时，信息的巨大损失变得不可避免。最有效的方法是使用统计数据表示BEV特征图，但其结果较差。基于pillar的方法[44]平衡了性能和成本，成为工业应用的流行选择。如何处理性能和效率之间的权衡成为基于激光雷达应用的关键挑战。

3.BEV Fusion

逆透视映射（IPM）[157]利用摄像机内外矩阵的几何约束将像素映射到BEV平面。尽管由于平地假设而不准确，但它提供了在BEV中统一图像和点云的可能性。Lift splat Shot（LSS）[56]是第一种预测图像特征深度分布的方法，引入神经网络来学习不适定相机到激光雷达转换问题。其它工作[41，58]开发了不同的方法来进行视图转换。考虑到从透视图到BEV的视图转换方法，下图显示了融合图像和点云数据的一般管道。模态特定特征提取器用于分别提取透视图和BEV中的特征。在转换为BEV中的表示之后，融合来自不同传感器的特征图。也可以在BEV表示中引入时间和自我运动信息。



(a) Perspective view (PV) perception pipeline



(b) BEV perception pipeline (BEV Fusion)

激光雷达相机融合

两部同名的作品BEVFusion[5，95]从不同方向探索了BEV中的融合。由于摄像机到激光雷达投影[72，159]抛弃了相机特征的语义密度，BEVFusion[5]设计 一种有效的相机到BEV变换方法，将相机特征有效地投影到BEV中，然后使用卷积层将其与激光雷达BEV特征融合。BEVFusion[95]将BEV融合视为保持感知系统稳定性的鲁棒性主题，它将摄像机和激光雷达特征编码到同一BEV中，以确保相机和激光激光雷达流的独立性。这种设计使感知系统能够在传感器故障时保持稳定性。除了BEVFusion[5，95]，UVTR[158]表示模态特定体素空间中的不同输入模式，无需高度压缩，以避免语义歧义，并实现进一步交互。图像体素空间是通过将每个视图的图像特征变换为预定义空间来构建的，其中为每个图像生成深度分布。使用常见的3D卷积网络构建点体素空间。然后在两个体素空间之间进行跨模态交互，以增强模态特定信息。

时间融合

时间信息在推断对象的运动状态和识别遮挡方面起着重要作用。BEV为连接不同时间戳中的场景表示提供了一个理想的桥梁，因为BEV特征地图的中心位置对ego-car来说是永久的。MVFuseNet[160]利用BEV和range视图进行时间特征提取，其它工作[52、62、63]使用ego运动将先前的BEV特征与当前坐标对齐，然后融合当前BEV特征以获得时间特征。BEVDet4D[63]使用空间对齐操作，然后连接多个要素图，将先前的要素图与当前帧融合。BEVFormer[4]和UniFormer[161]采用软方式融合时间信息，注意模块用于分别融合来自先前BEV特征图和先前帧的时间信息。关于ego car的运动，注意模块在不同时间戳表征中的位置也会被自我运动信息所修正。

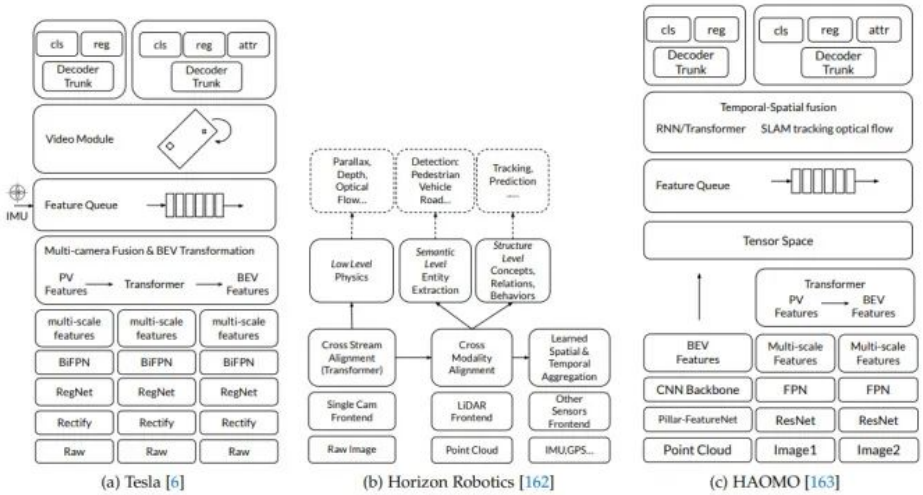
一些讨论

由于图像在透视坐标中，点云在3D坐标中，两种模式之间的空间对齐成为一个重要问题。尽管使用几何投影关系很容易将点云数据投影到图像坐标上，但点云数据的稀疏特性使得提取信息特征变得困难。相反，由于透视图中缺乏深度信息，将透视图中的图像转换为3D空间将是一个不适定问题。基于现有知识，以前的工作，如IPM[157]和LSS[56]可以将透视图中的信息转换为BEV，为多传感器和时间融合提供统一表示。激光雷达和摄像机数据在BEV空间的融合为3D检测任务提供了令人满意的性能。这种方法还保持了不同模式的独立性，这为构建更强大的感知系统提供了机会。对于时间融合，通过考虑自我运动信息，可以在BEV空间中直接融合不同时间戳中的表示。由于BEV坐标与3D坐标一致，通过监控控制和运动信息很容易获得自我运动补偿。考虑到鲁棒性和一致性，BEV是多传感器和时间融合的理想表示。

工业界中的BEV感知设计

近年来，BEV感知在行业中的流行趋势。上图描述了工业应用中传感器融合的两个典型范例，在BEV感知研究之前，大多数自动驾驶公司基于perspective view输入构建感知系统。图a基于几何先验，将来自图像的3D结果从2D结果转换。然后，我们融合图像和激光雷达的预测，利用一些手工制作的方法，这些方法在现实场景中并不总是表现良好。相反，图b基于BEV的方法使用神经网络执行2D到3D转换，并集成特征，而不是来自不同模态的直接检测输出，从而减少手工设计，提高鲁棒性。

下图总结了全球公司提出的各种BEV感知架构：



(a) Tesla [6] (b) Horizon Robotics [162] (c) HAOMO [163]

下表描述了详细的模型/输入选项，请注意，本调查中提供的所有信息均来自公共资源；不同计划之间的比较和分析基于事实：

TABLE 4: Detailed input and network options for BEV architectures as shown in Fig. 7. As we can observe, modality and feature extractor are different; Transformer and ViDAR are the most common choice for BEV transformation in industry. "-" indicates unknown information.

Company	Modality			Feature Extractor	BEV Transformation	Temporal & Spatial Fusion
	Camera	LiDAR	IMU/GPS			
Tesla [6]	✓	✗	✓	RegNet+BiFPN	Transformer ViDAR	✓
Mobileye (SuperVison) [164]	✓	✓	-	-	ViDAR	-
Hrizon Robotics [162]	✓	✓	✓	-	Transformer	✓
HAOMO.AI [163]	✓	✓	✓	ResNet+FPN Pillar-Feature Ne	Transformer	✓
PhiGent Robotics [165]	✓	✗	✗	-	ViDAR	-

1.输入数据

基于BEV的感知算法支持不同的数据模式，包括相机、激光雷达、雷达、IMU和GPS。摄像机和激光雷达是自动驾驶的主要感知传感器，一些产品仅使用摄像机作为输入传感器，例如特斯拉[6]、PhiGent[166]、Mobileye[164]。其他采用一套相机和激光雷达组合，例如Horizon[162]，HAOMO[163]。请注意，IMU和GPS信号通常用于传感器融合计划[6、162、163]，特斯拉和Horizon等的情况也是如此。

2. Feature Extractor

特征提取器用于将原始数据转换为适当的特征表示，该模块通常由主干和neck组成。特征提取器有不同的组合，例如，HAOMO[163]中的ResNet[149]和Tesla[6]中的RegNet[167]可以用作图像主干，neck可以是HAOMO[163]的FPN[79]，Tesla[6]的BiFPN[168]等。对于点云输入，HAOMO[163]的基于pilliar的选项或Mobileye的基于体素的选项是主干的理想候选。

3.PV到BEV转换

在行业中执行视图转换主要有四种方法：

（a）固定IPM。基于平坦地面假设，固定变换可以将PV特征投影到BEV空间，固定IPM投影也处理地平面，然而，它对车辆颠簸和路面平整度敏感。

（b） 自适应IPM利用通过一些姿态估计方法获得的SDV的外部参数，并相应地将特征投影到BEV。尽管自适应IPM对车辆姿态具有鲁棒性，但它仍然假设地面平坦。

（c） 基于transformer的BEV变换采用密集transformer将PV特征投影到BEV空间。这种数据驱动的转换在没有事先假设的情况下运行良好，因此被特斯拉、Horizon和HAOMO广泛采用[61，62，163]。

（d） ViDAR于2018年初由Waymo和Mobileye在不同地点并行提出[13，164]，以表明基于相机或视觉输入使用像素级深度将PV特征投影到BEV空间的实践，类似于激光雷达中的表示形式。

术语ViDAR相当于大多数学术文献中提出的伪激光雷达概念。配备ViDAR，可以将图像和后续特征直接转换为点云，然后，可以应用基于点云的方法来获得BEV特征。最近已经看到许多ViDAR应用，特斯拉、Mobileye、Waymo、丰田[6、13、164、169、170]等。总体而言，transformer和ViDAR的选择在行业中最为普遍。

4.Fusion模块

在先前的BEV变换模块中完成了不同摄像机源之间的对准。在融合单元中，进一步整合了摄像机和激光雷达的BEV特征。通过这样做，不同形式的特征最终被整合成一种统一的形式。

5.时空模块

通过在时间和空间上堆叠BEV特征，可以构建特征队列。时间堆栈每固定时间推送和弹出一个特征点，而空间堆栈每固定距离推送一个。在将这些堆栈中的特征融合为一种形式后，可以获得对遮挡具有鲁棒性的时空BEV特征[61，63]。聚合模块可以是3D卷积、RNN或transformer的形式。基于时间模块和车辆运动学，可以维护围绕ego车辆的大型BEV特征图，并局部更新特征图，就像特斯拉的空间RNN模块[6]中那样。

6.预测头

在BEV感知中，多头设计被广泛采用。由于BEV特征聚集了来自所有传感器的信息，所有3D检测结果都从BEV特征空间解码。同时，PV结果（对于自动驾驶仍然有价值）也从一些设计中的相应PV特征中解码。预测结果可分为三类：（a）低水平结果与物理约束有关，如光流、深度等。（b） 实体级结果包括对象的概念，即车辆检测、车道线检测等。（c）结构级结果表示对象之间的关系，包括对象跟踪、运动预测等。

经验和trick

数据增强

用于2D识别任务的图像上的通用数据增强适用于基于相机的BEV感知任务。一般来说，可以将增强分为静态增强和空间变换，静态增强仅涉及颜色变化，基于颜色变化的增强是直接适用的。对于涉及空间变换的增强，除了相应变换的地面真相外，还需要摄像机参数的校准。最近的工作中采用的常见增强是颜色抖动、翻转、多尺度调整大小、旋转、裁剪和网格遮罩。在BEVFormer++中，采用了颜色抖动、翻转、多尺度调整大小和网格掩码。输入图像按0.5和1.2之间的因子缩放，以0.5的比率翻转；总面积的最大30%被正方形掩模随机掩模。值得注意的是，在BEV感知中有两种翻转图像的方法。第一种方法是简单地相应地翻转图像、GT和相机参数。第二种方法还翻转图像顺序，以保持图像之间重叠区域的一致性，这类似于对称翻转整个3D空间。下图为BEV下的一些trick和消融实验：

TABLE 5: **BEV camera detection track.** Ablation studies on `val` set with improvements over BEVFormer [4], i.e., BEVFormer++. Some results are only reported as *L1/mAPH* on *car* category with *iou*≥*0.5*. DeD (Deformable DETR head). FrA (FreeAnchor head). CeP (Centerpoint head). ConvO (Conv offsets in TSA). DE (Deformable view Encoder). CoP (Corner Pooling). DA (2D Auxiliary loss). GR (Global location regression). MS (Multi-Scale), FL (filp), SL (Smoooth L1 Loss), EMA (Exponential Moving Average), SB (Sync BN), 2BS (2x BEV Scale), LLW (Leanable Loss Weight), LS (Label Smoothing) LE (LET-IoU based Assignment). DS (Dataset). The `mini` dataset contains 1/5 training data. *denotes the model is trained with 24 epochs, otherwise with 12 epochs.

ID	DeD	FrA	CeP	ConvO	DE	CoP	DA	GR	MS	FL	SL	EMA	SB	2BS	LLW	LS	LE	TTA	Backbone	DS	LET-mAPL	LET-mAPH	L1/mAPH
0	✓																		R101	mini	34.6	46.1	25.5
1	✓			✓															R101	mini	35.9	48.1	25.6
2	✓				✓														R101	mini	36.1	48.1	25.9
3	✓					✓													R101	mini	35.6	46.9	26.0
4	✓						✓								✓				R101	mini	36.2	48.1	25.4
5	✓							✓											R101	mini	35.4	47.2	27.2
6	✓								✓	✓									R101	mini	-	-	26.8
7	✓										✓								R101	mini	-	-	27.3
8	✓											✓							R101	mini	-	-	26.2
9	✓												✓						R101	mini	-	-	25.6
9	✓													✓					R101	mini	-	-	25.5
10	✓														✓				R101	mini	-	-	26.5
11	✓															✓			R101	mini	36.0	46.7	-
12	✓																✓		R101	mini	34.7	44.2	-
13	✓			✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓				✓	R101	mini	-	-	37.5
14*	✓																		SwinL	mini	40.0	55.6	51.9
15*	✓			✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓					SwinL	mini	44.7	60.8	55.5
16		✓																	R101	mini	35.9	49.9	45.9
17		✓															✓		R101	mini	36.3	51.1	46.6
18			✓																R101	mini	34.0	47.9	43.5
19	✓			✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓					SwinL	full	48.4	64.8	60.4
20		✓		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓					SwinL	full	47.2	61.2	56.8
21		✓		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓			✓		SwinL	full	47.6	61.4	57.0
22			✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓					SwinL	full	41.9	54.6	48.2

在lidar分割任务中，与检测任务不同，重数据增强可以应用于分割任务，包括随机旋转、缩放、翻转和点平移。对于随机旋转，从[0，2π）范围内选取一个角度，旋转应用于x-y平面上的每个点。从[0.9,1.1]范围中选择比例因子，然后乘以点云坐标，沿X轴、Y轴或X轴和Y轴进行随机翻转。对于随机平移，每个轴的偏移分别从均值为0和标准偏差为0.1的正态分布中采样。除了坐标和反射率，还可以利用额外的信息来提高模型性能。对于未标记的图像数据，通过将点云标签投影到相应的图像上并加密稀疏注释，从注释的点云数据中获得图像上的语义标签。训练图像模型以提供2D语义分割结果，然后，将预测的语义标签绘制为点云数据的一个热矢量，作为表示图像语义信息的附加通道。此外，还可以使用时间信息，因为自动驾驶中的数据集通常是按顺序收集的，过去的连续帧与当前帧连接。

TABLE 6: **BEV LiDAR segmentation track.** Ablation studies on `val` set with improvements over SPVCNN [83], i.e., Voxel-SPVCNN. Aug (heavy data augmentation). Arch (adjustments on model architecture). TTA (test-time augmentation). Painting (one-hot painting from image semantic segmentation). Temporal (multi-frame input). V-SPV (Voxel-SPVCNN). Expert (ensemble with more expert models). Post (post-processing techniques including object-level refinement and segmentation with tracking).

ID	Aug	Loss	Arch	TTA	Painting	Temporal	V-SPV	Ensemble	Expert	Post	mIoU
0											67.4
1	✓										67.8
2	✓	✓									68.4
3	✓	✓	✓								69.6
4	✓	✓	✓	✓							71.1
5	✓	✓	✓	✓	✓						71.6
6	✓	✓	✓	✓	✓	✓					72.4
7	✓	✓	✓	✓	✓	✓	✓				73.5
8	✓	✓	✓	✓	✓	✓	✓	✓			74.2
9	✓	✓	✓	✓	✓	✓	✓	✓	✓		74.5
10	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	75.4

Test-time Augmentation

2D任务的常见测试时间增加，包括多尺度和翻转测试，以提高3D情况下的精度。在BEVFormer++中，这一部分以使用标准数据增强（如多尺度和翻转）的形式进行了简单探索。多尺度增强的程度与训练相同，从0.75到1.25不等。

点云数据在推理过程中，使用了多个TTA，包括旋转、缩放和翻转。对于缩放，所有模型的缩放因子都设置为｛0.90、0.95、1.00、1.05、1.10｝，因为缩放因子越大或越小对模型性能有害。翻转与训练阶段相同，即沿X轴、Y轴以及X轴和Y轴。

后处理

虽然BEV检测消除了多摄像机对象级融合的负担，但也观察到了可从进一步后处理中获益的显著事实，利用2D检测结果对3D检测结果进行重复移除是有益的，其中2D box和3D box是二分匹配的。

参考

[1] Delving into the Devils of Bird's-eye-view Perception: A Review, Evaluation and Recipe.2022

极市干货

算法竞赛：国际赛事证书，220G数据集开放下载！ACCV2022国际细粒度图像分析挑战赛开赛！

技术综述：浅聊对比学习（Contrastive Learning）｜深度学习图像分类任务中那些不得不看的11个tricks总结

极视角动态：极视角与华为联合发布基于昇腾AI的「AICE赋能行业解决方案」｜算法误报怎么办？自训练工具使得算法迭代效率提升50%！