arXiv:2303.01212v2 [cs.CV] 9 Jun 2024

面向自动驾驶的以栅格为中心的交通场景感知：全面综述

Yining Shi[1,3], Kun Jiang[1*]、Jiusi Li[1], Zelin Qian[1],Junze Wen[1],Mengmeng Yang[1],Ke Wang[2] 和 Diange Yang[1*]

摘要——**以网格为中心的感知是移动机器人感知与导航的关键领域。然而，与以目标为中心的感知相比，以网格为中心的感知并不常见，因为自动驾驶车辆需要准确感知高度动态、大规模的交通场景，而以网格为中心的感知复杂度高、计算成本大。近年来，深度学习技术和硬件的快速发展为以网格为中心的感知演进带来了新的思路。以网格为中心的流程与以目标为中心的流程之间的根本区别在于，以网格为中心的感知遵循几何优先范式，因此在开放世界驾驶场景中更具鲁棒性，能够应对无穷无尽、长尾且语义未知的障碍物。近期研究表明，以网格为中心的感知具有显著优势，例如能够提供全面、精细的环境表示，对遮挡和形状不规则的物体具有更强的鲁棒性，能够更好地进行地面估计，并制定更安全的规划策略。此外，占用网络的能力不断增强，已扩展至四维场景感知与预测；最新技术也与自动驾驶领域的四维占用预测、生成式人工智能和世界模型等新兴研究主题密切相关。鉴于目前针对这一快速发展的领域尚缺乏综述，本文对自动驾驶车辆的以网格为中心的感知进行了层次化综述。我们沿着从二维鸟瞰视角栅格到三维占用，再到四维占用预测这一主线，梳理了占用栅格技术的既有与最新研究成果。此外，我们还总结了标签高效占用学习，以及以网格为中心的感知在驾驶系统中的作用。最后，我们总结当前的研究趋势，并展望未来的发展方向。**

索引词—**自动驾驶、以网格为中心的感知、占用流、场景重建、预测式世界模型、高效学习、感知与规划**

一、引言

SAFE自动驾驶车辆的运行需要准确、全面地表征周围环境。自动驾驶车辆的感知流程主要有两种：以对象为中心的流程和以网格为中心的流程。以对象为中心的流程应用广泛，由三维目标检测、跟踪和轨迹预测组成。然而，在开放世界交通场景中，如果网络无法明确定义物体的外观或模式，以对象为中心的流程通常会失效。这些障碍物也称为长尾障碍物，包括可变形障碍物（如双节拖车）、特殊形状障碍物（如翻覆车辆）、类别未知的障碍物（如路面上的碎石、垃圾）、部分被遮挡的物体等。因此，亟需针对这些长尾问题采用更稳健的表征方式。

作为应对无尽开放世界感知挑战的一种新兴方案，网格中心化流程被认为前景广阔，因为它能够在无需知道物体具体类别的情况下，提供三维周围空间中任意位置的占用和运动。基于网格的表示相较于基于物体的表示具有以下显著优势：不受障碍物几何形状或语义类别影响，且抗遮挡能力更强；采用统一的空间坐标来对齐多模态传感器数据；不确定性估计稳健。然而，以网格为中心的感知的主要缺点是计算复杂度和标签生成复杂度较高，因此，标签高效且计算高效的占用网络更有利于大规模训练和车载部署。

从技术发展路径来看，相关技术呈现出从二维建模到三维建模，再到四维建模的明显趋势，旨在更精细地描绘时空变化的三维环境。长期以来，占据栅格地图构建（OGM）一直被广泛视为移动机器人和自动驾驶车辆实现安全导航与碰撞规避的重要前提。随着深度神经网络的发展，以网格为中心的方法迅速发展，如今相比传统OGM，能够更全面地理解语义和运动。此外，基于BEV的网络已扩展为完整的三维占用网络，并进一步借助时间预测模块扩展至四维占用预测。

近期进展表明，以网格为中心的感知是自动驾驶车辆领域新兴、前景广阔且颇具挑战性的研究课题之一。为此，我们旨在首次对以网格为中心的感知技术进行全面综述。

与现有综述的比较。 现有关于汽车感知的综述涵盖了三维目标检测 [1], 3图像三维目标检测 [2],神经辐射场（NeRF） [3], 视觉中心化鸟瞰图感知 [4], [5],等内容，涉及以网格为中心的感知的部分技术。然而，这些评述并未全面讨论以网格表示为核心的感知任务、算法和应用。最新的占用网络综述预印本 [6], [7] 仅涵盖以网格为中心的感知中的有限领域：

[1]Yining Shi、Kun Jiang、Jiusi Li、Zelin Qian、Junze Wen、Mengmeng Yang 和 Diange Yang 任职于中国北京清华大学车辆与运载学院。{syn21,ljs23,qzl22, wjz22}@mails.tsinghua.edu.cn, {jiangkun,yangmm_qh,ydg}@mail.tsinghua.edu.cn [2]Ke Wang 任职于 KargobotInc. kewang@kargobot.ai部分工作是在 Yining Shi 于滴滴出行实习期间完成的。通讯作者：Kun Jiang、Diange Yang。

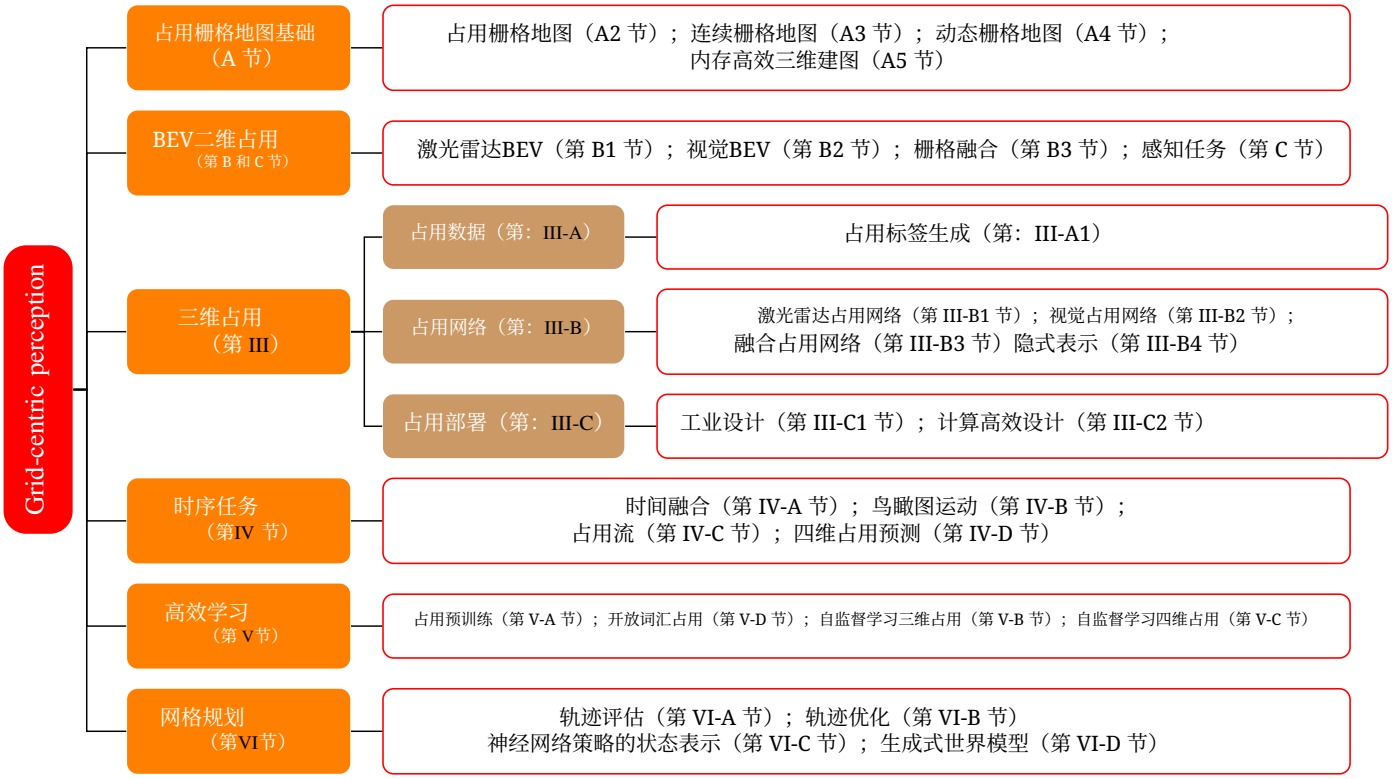

图 1. 自动驾驶以网格为中心的感知的层次化分类体系。

分别是基于图像的三维占用预测和信息融合。

与此不同，我们全面概述了面向自动驾驶车辆应用的以网格为中心的感知方法，并对以网格为中心的感知进行了深入研究和系统比较。在特征表示方面，我们涵盖鸟瞰视图二维表示、三维占用和四维预测。在数据效用方面，我们还探讨了面向占用的标签高效学习。在驾驶系统方面，我们研究了以网格为中心的感知如何助益决策和规划。

总而言之，本文综述的贡献总结如下：

1）据我们所知，我们首次从多个角度全面综述了自动驾驶中的以网格为中心的感知方法。2）我们对以网格为中心的感知技术进行了结构化、层次化的概述，并分析了学术界和工业界在自动驾驶实践中对以网格为中心的感知的看法。3）我们总结了当前趋势的观察结果，并展望了以网格为中心的感知的未来。

如图 I所示，本文采用层次化分类体系进行组织。我们聚焦于该分类体系中的四个核心问题：特征的空间表示、特征的时间表达、高效学习，以及以网格为中心的感知中的规划。第 II 节介绍以网格为中心的感知的背景，包括任务介绍以及与以对象为中心的流水线的比较，常用数据集、热门基准和评估指标。第 III节讨论三维占用网络的数据准备、网络设计和部署技术。第 IV 节介绍时间融合、运动预测和四维占用预测。第 V节介绍标签高效范式，包括利用自监督学习学习开放词汇占用，以及将占用作为下游任务的统一预训练。第 VI节介绍规划方法如何受益于以网格为中心的感知输入。第 VII 节讨论当前趋势的局限性，并展望以网格为中心的感知技术的未来发展。第 VIII 节对本文进行总结。

尽管过去五年中以网格为中心的感知领域取得了进展，附录仍收录了对成熟的非深度学习占用栅格地图及其变体的综述。此外，附录还讨论了通用的鸟瞰图感知流水线。

II. 背景

本节介绍以网格为中心的感知的一般任务形式化、常用数据集、热门基准和评估指标。

A. 任务介绍

以网格为中心的感知是指这样一种概念：给定车载传感器的多模态输入，算法需要将原始信息转换为鸟瞰视图或体素网格，并执行

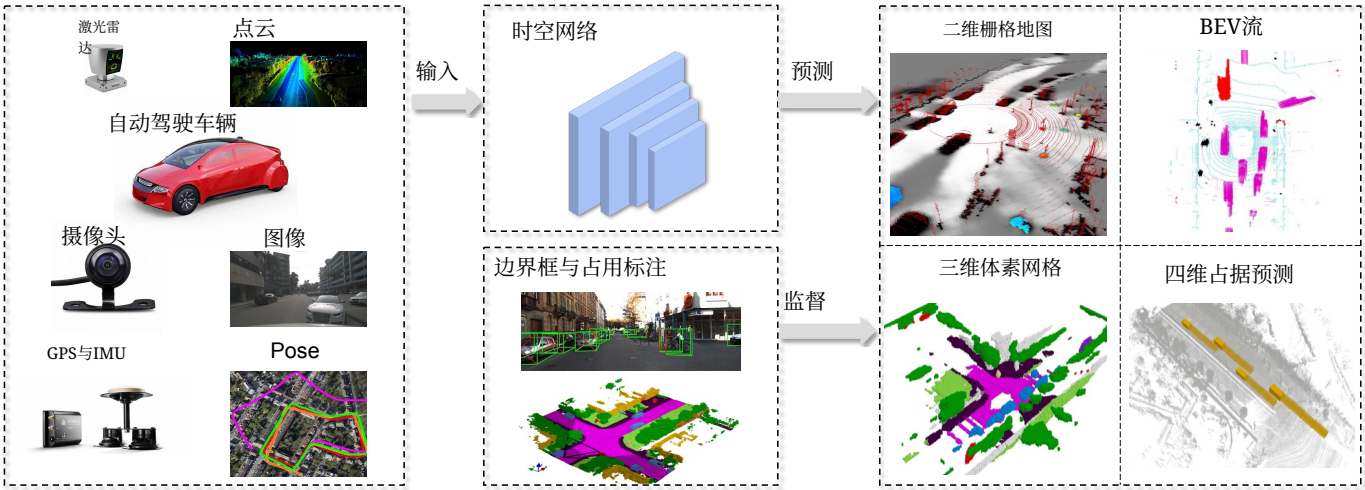

图2. 自动驾驶场景中以网格为中心的感知示意图。（左）：自动驾驶车辆配备了激光雷达、雷达和摄像头等不同传感器，以及用于定位的全球定位系统和惯性测量单元。（中）：在标注提供的边界框或占用监督下，原始数据、点云和图像由时空网络进行处理。（右）：以网格为中心的感知所产生的不同感知输出。图像来源：[8]–[11]。

在每个网格单元上执行各种感知任务。以网格为中心的感知的一般公式可表示为：

$$\mathcal{G} = f(I_{sensor}), \tag{1}$$

其中，$\mathcal{G}$ 是一组过去和未来的网格级表征，I_{sensor} 表示一个或多个传感器输入。如何表示网格属性和网格特征是这项任务中的两个关键问题。以网格为中心的感知流程示意图如图2所示。根据任务和标注的不同，占用表达形式多种多样，包括二维、三维和四维语义占据以及流预测。

分析：与其他感知流程的比较。以对象为中心的流水线，包括三维目标检测、跟踪和预测，专注于使用三维边界框表示尺寸通常相近且定义明确的障碍物。然而，在开放道路上，新类别和新实例不可避免。针对这一长尾问题，已有多种解决方案。在图像域中，开放世界 [12]–[14]检测方法学习对象性学习器，以在不了解具体类别的情况下学习物体所在位置。在点云域中，Openset3D [15] 采用了类似的开放世界范式。这种方法存在一定局限性：未知物体必须呈现出物体的外观，不能存在严重遮挡，并且需要与已知物体具有某种相似性。

基于分割的流水线，包括图像分割和点云分割，可从原始传感器数据中分割出未知物体。开放世界图像分割使用大型视觉模型（例如 Segment-Anything [16]）。开放词汇点云分割依赖于对标定点云模型和图像模型进行蒸馏。基于分割的方法面临的问题是难以克服传感器数据本身的噪声，例如高强度激光雷达点云中的“重影”和“膨胀”问题，以及视觉在低照度和逆光条件下拍摄的问题。此外，基于分割的流水线通常采用复杂的多传感器融合策略。

网格中心化流程对道路障碍物的低层占用和语义线索进行分割。以网格为中心的感知具有若干优势：它放宽了对障碍物形状的限制，能够描述形状可变的可动关节物体；同时也降低了对障碍物典型性的要求。它能够准确描述新类别和新实例的占用与运动线索，从而增强系统的鲁棒性。此外，通过模型设计或多模态融合，可以消除原始传感器数据中的噪声。

从本质上说，占据感知遵循先进行空间感知、再进行语义感知的理念。系统首先理解空间占据情况，聚类前景对象，然后识别语义。由于自动驾驶对空间信息较为敏感，但并不要求对障碍物的确切语义类别具有很高的准确性。一个典型例子是，MotionNet [11]无需学习具体类别，也无需将残障人士分类为行人，就能成功检测出残障人士。占据感知非常适合实现稳健的开放世界三维感知。

B. 数据集

以网格为中心的方法大多基于现有的大规模自动驾驶数据集开展，这些数据集包含三维物体边界框、激光雷达分割标签、二维和三维车道线标注以及高清地图。此外，人们还基于现有数据集构建了新的基准和三维占用标签。这些数据集包括 KITTI [17],nuScenes [18],Argoverse [19],Lyft L5 [20], SemanticKITTI [21], KITTI-360[22], Waymo 开放数据集（WOD） [23], Argoverse2 [24] 和 ONCE数据集 [25]。表 II-C3 总结了这些基准的相关信息。

C. 基准与评估指标

本节介绍基于公共数据集构建的面向占用的基准及其评估指标。

*1）BEV*分割基准：BEV分割定义为对鸟瞰视图网格进行语义分割或实例分割。常见类别包括动态物体（车辆、卡车、行人和骑行者）、静态道路布局和地图元素（车道、人行横道、可行驶区域、步道）。

BEV分割的评估指标。对于将网格分类为占用或空闲的空闲空间分割，大多数研究使用准确度这一简单指标。对于语义分割，主要评估指标是每个类别的**交并比 (I**oU)，以及所有类别的平均交并比(mIoU)。预测 A 与标签 B 之间的IoU表示为：

$$\mathrm{IoU}(A,B)=\frac{A\cap B}{A\cup B}, \tag{2}$$

2) 三维占用预测基准：**语义场景补全**。SemanticKITTI [21]数据集首次定义了室外语义场景补全(SSC)任务。SSC任务最初旨在根据单帧激光雷达点云预测特定空间范围内的完整场景，之后进一步扩展到摄像头。在自车周围的场景中，该空间范围由均匀的体素网格表示，每个体素都具有占用属性（空闲或占用）及其语义标签。SSCBench [26] 在KITTI-360 [22], nuScenes [18] 和Waymo开放数据集 [23]上标注了SSC标签。

三维占用预测（Occ3D）。在特斯拉发布其占用网络后，三维占用基准 [27],[28] 建立在公开数据集之上。三维占用预测被形式化为联合估计场景中每个体素的占据状态和语义标签。占据状态分为“空闲”“占用”或“未观测”，语义标签则包括预定义类别和通用物体。

常用的占用基准包括 Occ3D-nuScenes [27], OpenOccupancy-nuScenes [28] 以及 Occ3D-Waymo [27]。近期举办了挑战赛，以推动以视觉为中心的三维占用预测研究，其中包括基于 Occ3D-nuScenes [27]的CVPR2023研讨会三维占用预测挑战赛[1] 。

分析：语义场景补全与 Occ3D 任务定义的差异。语义场景补全与 Occ3D 之间主要存在两点差异。其一是遮挡处理：语义场景补全基于可见区域补全被遮挡区域，而 Occ3D 仅在可见区域上进行评估。其二是动态物体：语义场景补全在真值生成时假设场景是静态的，而 Occ3D 则预测动态物体。

三维占用预测（Occ3D）和语义场景补全（SSC）的评估指标。语义场景补全的主要评估指标是所有语义类别的**平均交并比**。**交并比**、精确率和召回率用于评估场景补全的几何重建质量。三维占用预测挑战赛使用F分数作为完整性 P_c 和准确度 P_a的调和平均数，F分数的计算公式如下：

$$F-score=(\frac{P_c^{-1}+P_a^{-1}}{2})^{-1} \tag{3}$$

[1]https://opendrivelab.com/AD23Challenge.html#Track3

其中， P_a 表示预测体素中与真实值体素的距离在距离阈值内的比例，而P_c 表示真实值体素中与预测体素的距离在距离阈值内的比例。由于大多数真实世界数据集的真实标注是半稠密的，因此所有评估指标仅在标注空间内计算。

3) 占用与流预测基准：除了占用之外，速度信息对于动态场景的安全性也至关重要。传统的以对象为中心的流程在跟踪过程中获取速度，这会降低可靠性。相比之下，直接在时空网格上获取占用速度，可以绕过训练中的匹配过程，并获得更稳健的运动估计。

BEV占用运动与流。BEV占用任务有不同的形式。MotionNet将BEV运动任务定义为预测每个网格单元在短期未来的运动位移，即每个网格单元在接下来一秒内可能移动多远。FIERY [29] 将流任务定义为未来两秒内的未来视频实例分割任务。Waymo引入了长期（未来最多8秒）的占用预测，同时将其作为轨迹集预测的有用补充，并在CVPR 2022研讨会上举办了占用与流挑战赛[2] 。该任务规定，给定场景中真实主体的一秒历史信息，任务必须预测所有主体在未来八秒内的流场。

三维占用与流。 OpenOCC [30] 考虑前景物体运动，并利用物体体素的流（速度）标注，基于 nuScenes [18] 构建新的三维占用基准。CVPR 2024研讨会举办占用与流挑战赛[3], , 要求根据来自多个摄像头的图像，预测场景中每个体素的语义和流。

分析：与场景流的比较。 光流和场景流 [31] 已被广泛用于估计相邻帧中图像像素或激光雷达点的运动。相比之下，BEV流是对动态场景更具结构化的表示，其运行速度远快于无序点云和稠密图像像素。此外，场景流关注瞬时运动，而BEV流则延伸至更远的未来占用。

BEV占用与流预测的评估指标。Motion-Net [11] 通过在BEV地图中为每个网格单元关联一个位移向量来编码运动信息，并提出了运动预测评估指标：将非空网格单元划分为三个速度范围：静态、慢速（$\leq$ 5m/s）和快速（$>$ 5m/s）。在每个速度范围内，均值和中位数 L_2被用于衡量预测位移与真实位移之间的距离。

FIERY [29]使用**视频全景质量 (VPQ)** [32]指标来评估 BEV 地图中的未来实例分割和运动预测。该指标定义如下：

$$\mathrm{VPQ}=\sum_{t=0}^{H}\frac{\sum_{(x_t,y_t)\in TP_t}\mathrm{IoU}(x_t,y_t)}{|TP_t|+\frac{1}{2}|FP_t|+\frac{1}{2}|FN_t|}, \tag{4}$$

[2]https://waymo.com/open/challenges/2022/occupancy-flow-prediction-挑战赛/

[3]https://opendrivelab.com/challenge2024/#占用 and 流

数据集	Year	时长（小时）	场景	激光雷达扫描	RGB图像	标注帧	标注	类别	夜间/雨天	视角	立体视觉	地点	辅助
KITTI [17]	2012	1.5	22	15k	15k	15k	200k	8	无/无	1	Yes	德国	-
Lyft L5 [20]	2019	2.5	366	46k	323k	46k	1.3M	9	否/否	6	No	USA	Maps
Argoverse [19]	2019	0.6	113	44k	490k	22k	993k	15	是/是	7	Yes	USA	Maps
SemanticKITTI* [21]	2019	-	22	43k	-	-	4549M*	28	-/-	-	-	德国	-
nuScenes [18]	2020	5.5	1000	400k	1.4M	40k	1,4M	23	是/是	6	No	新加坡,美国	地图,雷达数据
Waymo Open [23]	2020	6.4	1150	230k	1M	230k	12M	23(4)**	是/是	5	No	USA	-
KITTI-360 [22]	2021	-	-	80k	300k	-	68k	37	-/-	3	Yes	德国	-
Argoverse 2 [24]	2021	-	1000	-	-	-	-	30	是/是	9	Yes	USA	Maps
ONCE [25]	2021	144	-	1M	7M	-	417k	5	是/是	7	No	中国	-

表 I 以网格为中心的感知领域重要数据集的详细信息。(*) 基于 KITTI 的 SemanticKITTI 数据集包含带有密集逐点标注的激光雷达扫描。对于其他数据集标注，我们列出边界框的数量。(**) Waymo 开放数据集为每个激光雷达点提供包含 23 个类别的密集标签，以及包含 4 个类别的三维边界框标签。(-) 表示未提供相关信息，或该数据集不包含此项目。SG：新加坡。

其中， TP_t、 FP_t、 FN_t 分别表示时间步 t上的真正例、假正例和假负例实例预测， H 表示未来预测时域。请注意，真正例预测与真实标注的交并比超过 0.5，且在整个时间范围内具有与真实标注相同的实例 ID。

Waymo 的占用流挑战赛评估算法在较长时间跨度（8 秒）内预测占用分布的能力。占用流的评估指标包括占用指标、流指标和联合指标。主要的占用指标是**曲线下面积 (AUC)**和**软交并比 (Soft-IoU)** [33]，用于二元分割。类别 $\mathcal{K}$(车辆/行人)的 AUC($O_t^{\mathcal{K}}$, $\hat{}O_t^{\mathcal{K}}$)用于估算精确率-召回率曲线下的面积。每个类别的软交并比指标 $\mathcal{K}$ 表示为：

$$\text{Soft-IoU}(O_t^{\mathcal{K}}, \hat{O}_t^{\mathcal{K}}) = \frac{\sum_{x,y} O_t^{\mathcal{K}} \cdot \hat{O}_t^{\mathcal{K}}}{\sum_{x,y} O_t^{\mathcal{K}} + \hat{O}_t^{\mathcal{K}} - O_t^{\mathcal{K}} \cdot \hat{O}_t^{\mathcal{K}}}, \quad (5)$$

$\hat{}$其中 O_t, O_t 分别表示时间步 t 的真实占用和预测占用。

对于流指标，**端点误差(EPE)**用于衡量真实流场 $F_t^{\mathcal{K}}(x, y)$ 与预测流场 $\hat{}F_t^{\mathcal{K}}(x, y)$ 之间的平均 L2 距离：

$$\left\| F_t^{\mathcal{K}}(x,y) - \hat{F}_t^{\mathcal{K}}(x,y) \right\|_2, where\ F_t^{\mathcal{K}}(x,y) \neq (0,0), \quad (6)$$

其中，时间 t 的流场 F_t 为每个像素提供一个运动向量（dx, dy）。

联合指标用于衡量每个时间步 t的流和占用预测准确度，因此使用 F_t 对真实占用（O_{t-1}）进行变换：

$$\hat{W}_t = \hat{F}_t \circ O_{t-1}, \quad (7)$$

其中， $\circ$ 将流场作为函数来变换占用。如果联合预测足够准确，$\hat{}\hat{}W_t O_t$ 应接近真实值 O_t。因此，采用了**流场引导曲线下面积**($O_t^{\mathcal{K}}$, $\hat{}W_t \hat{}O_t^{\mathcal{K}}$) 和 **流场引导 Soft-交并比**($O_t^{\mathcal{K}}$, $\hat{}W_t$ $\hat{}O_t^{\mathcal{K}}$)。

三维占用与流挑战赛的评估指标。CVPR 2024 占用与流挑战赛将占用分数设为主要评估指标，其中包括基于射线的平均交并比，以及占用流的绝对速度误差 (AVE)。与逐体素交并比相比，基于射线的平均交并比具有优势 [34]，因为逐体素交并比对深度估计的惩罚不一致，而且它无需使用可见性掩码即可避免评估激光雷达点未扫描到的区域。

对于基于射线的平均交并比，将查询射线投射到预测占用和真实占用中。每条射线与任意表面相交时，都会获得其深度、类别标签和流。如果类别标签一致，且真实深度与预测深度之间的 L1 误差小于某个阈值，则将该查询射线判定为真正例。同样，可以得到假正例和假负例的数量，从而计算各类别的平均交并比。最后，基于射线的平均交并比是距离阈值为 1，2，4 米时的平均值。对于绝对速度误差（AVE），挑战赛会计算阈值为 $2m$的真正例的速度误差。最终占用分数表示如下：

$$Score_{occ} = 0.9 \times \text{mIoU} + 0.1 \times max(1 - \text{mAVE}, 0.0) \quad (8)$$

4）四维占用预测基准：四维占用预测将三维占用预测和流预测扩展为 4D 时空占用预测。给定过去和当前的连续点云 [35] 或相机图像 [10]，4，四维占用预测旨在输出当前占用以及当前坐标系下未来一段时间内的占用。四维占用预测既可采用监督式或无监督式方法，也可使用真实标签或原始点云序列进行评估。OCFBench [35] 和 Cam4DOcc [10] 利用目标跟踪标签和三维占用标签构建四维真实标注，而 Argoverse 2 4D 占用挑战赛[4] 和 CVPR 2024 预测世界模型挑战赛[5] 则通过从未来激光雷达原点向未来激光雷达点云进行射线投射来评估四维占用。

四维占用预测的评估指标。如果评估使用四维占用标签，这些基准会采用广泛使用的评估指标，例如平均交并比、平均精度均值、精确率、召回率和 F1 分数，以评估模型在整个序列中的性能。

如果评估时没有四维占用标签，则每个激光雷达点都会沿特定方向，从激光雷达原点向投影表面形成一条射线。预测深度是

[4]https://eval.ai/web/challenges/challenge-page/1977/overview [5] https://opendrivelab.com/challenge2024/#predictive world model

方法	模态	交并比	平均交并比	Road	Sidewalk	Parking	Other-grnd	Building	Car	Truck	Bicycle	Motorcycle	Other-veh.	Vegetation	Trunk	Terrain	Person	Bicyclist	Motorcyclist.	Fence	Pole	Traf.-sign
S3CNet [38]	L	45.60	**29.53**	42.00	22.50	17.00	7.90	**52.20**	31.20	6.70	**41.50**	**45.00**	**16.10**	39.50	**34.00**	21.20	**45.90**	**35.80**	**16.00**	**31.30**	**31.00**	**24.30**
LMSCNet [39]	L	56.72	17.62	64.80	34.68	29.02	4.62	38.08	30.89	1.47	0.00	0.00	0.81	41.31	19.89	32.05	0.00	0.00	0.00	21.32	15.01	0.84
JS3C-Net [40]	L	56.60	23.75	64.70	39.90	34.90	**14.10**	39.40	33.30	**7.20**	14.40	8.80	12.70	43.10	19.60	40.50	8.00	5.10	0.40	30.40	18.90	15.90
Local-DIFs [41]	L	**58.90**	23.56	**69.60**	**44.50**	**41.80**	12.70	41.30	**35.40**	4.70	3.60	2.70	4.70	**43.80**	27.40	**40.90**	2.40	1.00	0.00	30.50	22.10	18.50
OpenOccupancy [28]	C+L	-	20.42	60.60	36.10	29.00	**13.00**	**38.40**	33.80	4.70	3.00	2.20	5.90	**41.50**	20.50	35.10	0.80	2.30	**0.60**	26.00	18.70	15.70
Co-Occ [42]	C+L	-	**24.44**	**72.00**	**43.50**	**42.50**	10.20	35.10	**40.00**	**6.40**	**4.40**	**3.30**	**8.80**	41.20	**30.80**	**40.80**	**1.60**	**3.30**	0.40	**32.70**	**26.60**	**20.70**
MonoScene [8]	C	34.16	11.08	54.70	27.10	24.80	5.70	14.40	18.80	3.30	0.50	0.70	4.40	14.90	2.40	19.50	1.00	1.40	0.40	11.10	3.30	2.10
TPVFormer [43]	C	34.25	11.26	55.10	27.20	27.40	6.50	14.80	19.20	3.70	1.00	0.50	2.30	13.90	2.60	20.40	1.10	2.40	0.30	11.00	2.90	1.50
OccFormer [44]	C	34.53	12.32	55.90	30.30	**31.50**	6.50	15.70	21.60	1.20	1.50	1.70	3.20	16.80	3.90	21.30	2.20	1.10	0.20	11.90	3.80	3.70
SurroundOcc [45]	C	34.72	11.86	56.90	28.30	30.20	6.80	15.20	20年.	601.40	1.60	1.20	4.40	14.90	3.40	19.30	1.40	2.00	0.10	11.30	3.90	2.40
NDC-Scene [46]	C	36.19	12.58	58.12	28.05	25.31	6.53	14.90	19.13	4.77	1.93	2.07	6.69	17.94	3.49	25.01	**3.44**	2.77	1.64	12.85	4.43	2.96
RenderOcc [47]	C	-	8.24	43.64	19.10	12.54	0.00	11.59	14.83	2.47	0.42	0.17	1.78	17.61	1.48	20.01	0.94	3.20	0.00	4.71	1.17	0.88
Symphonies [48]	C	42.19	15.04	58.40	29.30	26.90	**11.70**	**24.70**	23.60	3.20	3.60	**2.60**	5.60	24.20	**10.00**	23.10	3.20	1.90	**2.00**	16.10	7.70	8.00
HASSC [49]	C	42.87	14.38	55.30	29.60	25.90	11.30	23.10	23.00	**9.80**	1.90	1.50	4.90	24.80	9.80	26.50	1.40	3.00	0.00	14.30	7.00	7.10
BRGScene [50]	C	43.34	15.36	**61.90**	**31.20**	30.70	10.70	24.20	22.80	8.40	3.40	2.40	6.10	23.80	8.40	27.00	2.90	2.20	0.50	**16.50**	7.00	7.20
VoxFormer [51]	C	**44.15**	13.35	53.57	26.52	19.69	0.42	19.54	**26.54**	7.26	1.28	0.56	**7.81**	**26.10**	6.10	**33.06**	1.93	1.97	0.00	7.31	**9.15**	4.94
MonoOcc [52]	C	-	**15.63**	59.10	30.90	27.10	9.80	22.90	23.90	7.20	**4.50**	2.40	7.70	25.00	9.80	26.10	2.80	**4.70**	0.60	16.90	7.30	**8.40**

表 II SemanticKITTI 测试集上的三维占用预测对比 [21]。MOD.：模态。C：摄像头。L：激光雷达。

沿射线投射至该方向上最近的占用表面的距离。随后，预测深度会转换为预测点云。预测点云与真实点云之间的倒角距离（CD）是主要评估指标。评估时也会使用 L1 误差和相对绝对L1误差。4d-occ-forecasting [36] 会对所有未来帧进行联合评估，而 ViDAR [37]则报告每帧的倒角距离。这些评估通常会设定一个位于四维占用网格范围内的感兴趣区域（RoI）。

III. 三维占用映射

机器人学领域长期以来一直采用二维网格进行导航。然而，在一些复杂的交通场景中，使用鸟瞰视图网格进行驾驶容易失败，例如地面不平坦，或障碍物在高度维度上的形状不一致（悬空物体）。相比之下，三维网格能够表示驾驶场景的完整几何信息，包括路面和障碍物的形状，但代价是更高的计算成本。因此，高计算效率的占用网络是以网格为中心的感知中的一个重要问题。

本节首先回顾占用网络的完整流程，包括数据和标签如何准备、不同模态的网络如何设计，以及占用网络如何部署到车载平台。我们汇总了大多数方法，这些方法的性能均在三个常用数据集及其基准上进行了验证（SSC-SemanticKITTI、Occ3D-nuScenes、OpenOccupancy-nuScenes）。典型方法的比较结果分别见表。III、表III和表III。

A. 占用数据流水线

1）真实标签生成：与直接在原始传感器数据上标注标签的其他感知任务不同，占用标签通常需要进行第二阶段标注该阶段依赖于用于背景内容的激光雷达分割标签，以及用于前景对象的边界框。根据自车位姿坐标，将时间上连续的多帧点云转换到当前位置，然后在固定范围和分辨率下对点云进行体素化，以获取体素占用和语义标签。SurroundOcc 仓库中提供了一个标签生成工具包示例[6]。

Occ3D 任务的另一个重要标签是每个传感器的可见性掩码，因为在占用预测中预测不可见区域并不合理。使用可见性掩码进行训练时，大多数网络的性能都会略有提升。在 Occ3D 设置中，激光雷达可见性掩码通过计算从原点到点的每条射线得到：如果体素反射激光雷达点并被标记为占用，或被射线穿过并被标记为空闲，则该体素被标记为可见；否则，该体素被标记为不可见或未观测区域。对于摄像头掩码，则通过构造虚拟射线来推断遮挡：将最近的占用体素设为占用，将射线经过的区域设为空闲，并将最近观测点后方的体素设为不可见。

就具体操作而言，Occ3D任务与SSC任务的真实效果略有不同，各自也存在局限。

分析：SSC标签的局限性。现有户外SSC基准 [21], [22] 通过聚合所有帧的语义点云来生成标签。标签虽然稠密，但动态物体的痕迹不可避免地会对标签造成干扰，MotionSC [65]将其称为时空管。由于SemanticKITTI中存在大量停放车辆，所有现有SSC方法都会将动态物体预测为静态物体，并因此受到基准评估指标的惩罚，这也是SSC任务主要用于静态场景重建的原因。

分析：Occ3D标签的局限性。Occ3D基准[27], [28] 通过聚合当前帧前后各 20 个扫描帧来生成标签

[6]https://github.com/weiyithu/SurroundOcc

方法	图像骨干网络	mIoU	Others	Barrier	Bicycle	Bus	Car	Const. veh.	Motorcycle	Pedestrian	Traffic cone	Trailer	Truck	Drive. suf.	Flat	Sidewalk	Terrain	Manmade	Vegetation
验证集上的性能																			
MonoScene [8]	R101-DCN	6.06	1.75	7.23	4.26	4.93	9.38	5.67	3.98	3.01	5.90	4.45	7.17	14.91	6.32	7.92	7.43	1.01	7.65
CTF-Occ [27]	R101-DCN	28.53	8.09	39.33	20.56	38.29	42.24	16.93	24.52	22.72	21.05	22.98	31.11	53.33	33.84	37.98	33.23	20.79	18.00
BEVFormer [53]	R101-DCN	39.24	10.13	47.91	24.90	47.57	54.52	20.23	28.85	28.02	25.73	33.03	38.56	81.98	40.65	50.93	53.02	43.86	37.15
PanoOcc [54]	R101-DCN	42.13	11.67	50.48	29.64	49.44	55.52	23.29	**33.26**	30.55	30.99	34.43	42.57	**83.31**	44.23	54.40	56.04	45.94	40.40
BEVDet [55]	Swin-B	42.02	12.15	49.63	25.10	52.02	54.46	27.87	27.99	28.94	27.23	36.43	42.22	82.31	43.29	54.62	57.90	48.61	43.55
RadOcc [56]	Swin-B	**46.06**	9.78	**54.93**	20.44	**55.24**	**59.62**	30.48	28.94	**44.66**	28.04	**45.69**	**48.05**	81.41	39.80	52.78	56.16	**64.45**	**62.64**
三维占用预测挑战赛上的表现																			
BEVFormer [53]	R101-DCN	23.70	10.24	36.77	11.70	29.87	38.92	10.29	22.05	16.21	14.69	27.44	33.13	48.19	33.10	29.80	17.64	19.01	13.75
SurroundOcc [45]	R101-DCN	42.26	11.70	50.55	32.09	41.59	57.38	27.93	38.08	30.56	29.32	48.29	38.72	80.21	48.56	53.20	47.56	46.55	36.14
BEVDet [55]	Swin-B	42.83	18.66	49.82	31.79	41.90	56.52	26.74	37.31	30.01	31.33	48.18	38.59	80.95	50.59	53.87	49.67	46.62	35.62
PanoOcc [54]	Intern-XL	47.16	23.37	50.28	36.02	47.32	59.61	31.58	39.59	34.58	33.83	52.25	43.29	83.82	55.81	59.41	53.81	53.48	43.61
RadOcc [56]	Swin-B	49.98	21.13	55.17	39.31	48.99	59.92	33.99	46.31	**43.26**	39.29	52.88	44.85	83.72	53.93	59.17	55.62	60.53	51.55
UniOcc [57]	ConvNeXt-Base	51.27	26.94	56.17	39.55	49.40	60.42	35.51	44.77	42.96	38.45	59.33	45.90	83.90	53.53	59.45	56.58	63.82	**54.98**
MiLO [58]	Intern-XL&Swin-L	52.45	27.80	56.28	42.62	50.27	61.01	35.41	47.97	38.90	40.29	56.66	47.03	86.96	57.48	63.64	62.53	63.00	53.74
FB-Occ [59]	Intern-H	**54.19**	**28.95**	**57.98**	**46.40**	**52.36**	**63.07**	**35.68**	**48.81**	42.98	**41.75**	**60.82**	**49.56**	**87.29**	**58.29**	**65.93**	**63.30**	**64.28**	53.76

表 III Occ3D-nuScenes 上的三维占用预测性能。所有方法仅使用视觉作为输入。CVPR 2022年三维占用预测挑战赛采用 Occ3D-nuScenes [27] 测试集。

方法	模态	交并比	平均交并比	Barrier	Bicycle	Bus	Car	Const. veh.	Motorcycle	Pedestrian	Traffic cone	Trailer	Truck	Drive. suf.	Other flat	Sidewalk	Terrain	Manmade	Vegetation
MonoScene [8]	C	18.4	6.9	7.1	3.9	9.3	7.2	5.6	3.0	5.9	4.4	4.9	4.2	14.9	6.3	7.9	7.4	10.0	7.6
TPVFormer [43]	C	15.3	7.8	9.3	4.1	11.3	10.1	5.2	4.3	5.9	5.3	6.8	6.5	13.6	9.0	8.3	8.0	9.2	8.2
3DSketch [60]	C&D	25.6	10.7	12.0	5.1	10.7	12.4	6.5	4.0	5.0	6.3	8.0	7.2	21.8	14.8	13.0	11.8	12.0	21.2
AICNet [61]	C&D	23.8	10.6	11.5	4.0	11.8	12.3	5.1	3.8	6.2	6.0	8.2	7.5	24.1	13.0	12.8	11.5	11.6	20.2
LMSCNet [39]	L	27.3	11.5	12.4	4.2	12.8	12.1	6.2	4.7	6.2	6.3	8.8	7.2	24.2	12.3	16.6	14.1	13.9	22.2
JS3C-Net [40]	L	30.2	12.5	14.2	3.4	13.6	12.0	7.2	4.3	7.3	6.8	9.2	9.1	27.9	15.3	14.9	16.2	14.0	24.9
Point-Occ [62]	L	**34.1**	**23.9**	24.9	**19.0**	20.9	25.7	13.4	**25.6**	**30.6**	**17.9**	16.7	21.2	36.5	**25.6**	**25.7**	**24.9**	**24.8**	**29.0**
OccFusion [63]	C&L	31.1	17.0	15.9	15.1	15.8	18.2	15.0	17.8	17.0	10.4	10.5	15.7	26.0	19.4	19.3	18.2	17.0	21.2
M-baseline [28]	C&L	29.1	15.1	14.3	12.0	15.2	14.9	13.7	15.0	13.1	9.0	10.0	14.5	23.2	17.5	16.1	17.2	15.3	19.5
M-CONet [28]	C&L	29.5	20.1	23.3	13.3	21.2	24.3	**15.3**	15.9	18.0	13.3	15.3	20.7	33.2	21.0	22.5	21.5	19.6	23.2
Co-Occ [42]	C&L	30.6	21.9	**26.5**	16.8	22.3	**27.0**	10.1	20.9	20.7	14.5	16.4	21.6	**36.9**	23.5	25.5	23.7	20.5	23.5
OccGen [64]	C&L	30.3	22.0	24.9	16.4	**22.5**	26.1	14.0	20.1	21.6	14.6	**17.4**	**21.9**	35.8	24.5	24.7	24.0	20.5	23.5

表 IV nuScenes-OpenOccupancy 验证集上的三维语义占用预测结果。我们报告几何指标交并比、语义指标平均交并比，以及各语义类别的交并比。C、D 和 L 分别表示摄像头、深度和激光雷达。

。随后，为每一帧提供摄像头可见掩码和激光雷达可见掩码。这种方法可以准确标注动态物体，但其主要缺点是占用标签稀疏。例如，在不连续的远处地面，只有被激光雷达点占据的体素会被识别为地面，其余则被识别为空。

分析：标签生成的未来发展。标注质量对训练效果有很大影响。然而，依赖激光雷达分割的现有二次标注方法成本极高，严重限制了占用网络的可扩展性；另一方面，点云作为真值系统也存在固有的稀疏性缺陷。未来，一种很有前景的解决方案是采用神经渲染，利用 NeRF [66] 或高斯泼溅 [67]重建四维场景中的交通场景，然后将神经场体素化以生成真值。Emernerf [68]已进行了有价值的尝试：在四维 NeRF 重建质量与 Occ3D-Waymo 标签进行比较时，以完全自监督的方式获得令人满意的真值占用结果。

B. 占用网络

1）基于激光雷达的占用网络：由于激光雷达传感器天然适用于三维占用网格，基于激光雷达的重建所面临的挑战是：从障碍物表面反射的点推断完整的场景几何，并根据稀疏的激光雷达输入推断未观测区域的稠密几何结构。

采用二维卷积神经网络的激光雷达三维占用。SSCNet 和 TS3D 中的稠密三维卷积神经网络模块对内存和计算资源的需求较高，并且会导致数据流形膨胀。解决这一问题的一种替代方案是利用二维卷积神经网络的高效性。LMSCNet [39]采用轻量级 U-Net 架构，结合二维骨干卷积和三维分割头。将高度维度转换为特征维度，已成为交通场景中的常见做法，因为数据主要沿纵向和横向变化。基于柱体的 LMSCNet能够以较高速度实现良好性能，并具备推断多尺度语义场景补全的能力。类似地，Local-DIFs [41]创建点云的鸟瞰图特征图，并将其输入二维 U-Net，以输出三个尺度上的特征图；这些特征图共同构成三维场景的新型表示，即连续的深度隐式函数（DIF）。通过查询所有体素角点处的函数值，可以在 SemanticKITTI基准上评估 Local-DIFs，并且其几何补全精度表现良好。

采用稀疏卷积的激光雷达三维占用。另一种颇具前景的替代方案是使用稀疏三维网络，例如 JS3C-Net [40] 中使用的 SparseConv，以及 S3CNet [38]，中使用的 Minkowski [70] ；这些网络仅处理非空体素。JS3C-Net 是一种稀疏激光雷达点云语义分割框架，将语义场景补全作为辅助任务。该框架包含一个点-体素交互（PVI）模块，以增强这种多任务学习，并促进两项任务之间的知识迁移。在语义分割方面，它采用三维稀疏卷积U-Net。级联语义场景补全模块会预测粗略补全结果，并在 PVI 模块中进行细化。实验表明，JS3C-Net 在两项任务上均取得了最先进的结果。S3CNet 从单次激光雷达扫描中构建稀疏二维和三维特征，并行地将它们输入稀疏二维和三维 U-Net 风格的网络。为避免在解码器中使用稠密卷积，S3CNet 提出了一种鸟瞰视图与三维预测的动态体素化融合方法，以进一步提高场景的稠密度，随后采用空间传播网络来细化结果。值得一提的是，它在 SemanticKITTI 的稀有类别上取得了令人瞩目的结果。

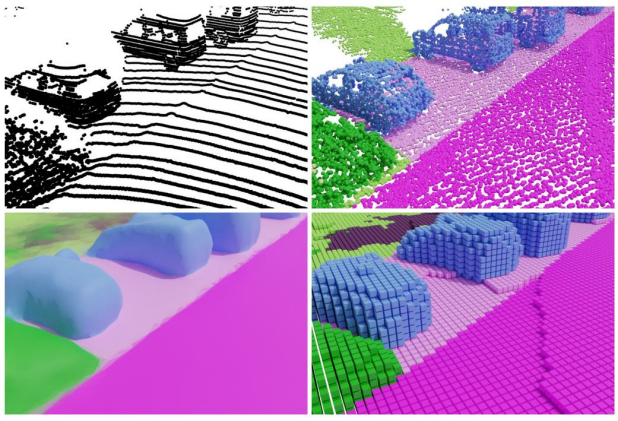

图3. Local-DIF [41]: Local-DIF 为三维语义占用生成连续表示，但在倾斜表面（例如路面）或物体之间的边缘处，会因离散化为体素而产生量化伪影。

采用三视角视图的激光雷达三维占用。PointOcc [62]是一种无需后处理的基于二维投影的模型。PointOcc 提出了一种新颖且有效的点云表示方法——圆柱三视角视图（Cylindrical TPV），并且无需使用三维算子，即可采用任意二维骨干网络进行特征提取。每个点都会被空间池化到三个平面上，其最终特征由这三个平面的特征相加得到，进而生成最终的分割结果。TPV 表示能够通过两个简单的分割头预测点云分割和占用分割，无需复杂的后处理。

使用非理想标签进行学习。为了解决 SSC 标签的局限性，并聚焦于输入瞬间的 SSC，Local-DIFs [41] 提出了一种基于 SemanticKITTI 的数据集变体：仅保留动态物体在单个时刻的扫描，并移除动态物体阴影范围内的自由空间点。此外，如图3所示，Local-DIF 可以连续表示场景，从而避免离散化造成的伪影。不过，Local-DIFs 在细微几何细节方面的分割性能仍有待提升，这归因于补全任务占据主导地位，以及小物体类别的真实标注存在歧义。Wilson 等人 [65] 开发了一个无遮挡的合成室外数据集 CarlaSC，并在 CARLA 中跟踪自车周围的场景 [71]。他们提出了一种实时稠密局部语义建图方法 MotionSC [65] ，该方法结合了 MotionNet 的时空骨干网络 [11] 和 LMSCNet 的分割头 [39]。值得注意的是，MotionSC 即使忽略时间信息，在 SemanticKITTI 基准上也表现良好。

基于雷达点云的占用网络。与激光雷达相比，雷达点云较为稀疏，而且传统三维雷达不包含垂直方向的信息，因此难以实现三维占用。RadarOcc [72] 首次尝试使用四维成像雷达实现三维占用。雷达栅格重建的局限在于需要保留关键原始信号，其挑战包括减少庞大的数据量、减轻旁瓣测量的影响，以及实现无插值的特征编码与聚合。

2）基于摄像头的占用网络：基于摄像头的方法正逐渐应用于三维占用领域。与以运动恢复结构（SFM）为代表的离线建图方法不同，在线感知会将像素投影到三维空间，并预测一个统一的三维占用网格，其中可能包含语义类别。视觉占用网络的通用流程是将鸟瞰视图特征表示扩展为体素特征表示。例如，FB-OCC [59] 是 FB-BEV [73] 的扩展，它采用 LSS 进行前向视图变换，并采用 BEVFormer 进行后向视图变换。采用 I nternImage-XL 的 FB-OCC 最大模型 [74]赢得了 CVPR 2023 占用预测挑战赛的冠军。

基于稠密体素特征的占用学习。MonoScene [8] 是首个基于单目摄像头的户外三维体素重建框架。它使用来自 SSC任务的稠密语义体素标签进行评估指标计算。该框架包含用于二维特征视线投影（FLoSP）的模块，以连接二维和三维U-Net，还包括三维上下文关系先验（CRP）层，用于增强上下文学习

信息。OccDepth [75] 是一种基于立体视觉的方法，通过立体视觉软特征分配模块将立体视觉特征提升至三维空间。它使用立体视觉深度网络作为教师模型，对深度增强占用感知模块进行蒸馏，使其成为学生模型。不同于上述需要稠密语义体素标签的方法。

基于稀疏到稠密体素特征的占用学习。VoxFormer [51] 是一种两阶段基于Transformer的框架，首先从深度图中获取稀疏的可见与占用查询，然后通过自注意力将其传播至稠密体素。

基于三平面特征的占用学习。TPVFormer[43] 是首个仅使用稀疏激光雷达语义标签作为监督的环视三维重建框架。TPVFormer将鸟瞰视图推广至三透视视图（TPV），即通过垂直于 x、 y、 z轴的三个切片来表达三维空间中的特征。该方法查询三维空间中的点，以任意分辨率解码占用。对于监督方法，TPVFormer [43] 使用稀疏激光雷达分割标签替代稠密体素网格标签，以监督来自环视摄像头的稠密语义占用。与固定分辨率的体素标签相比，点云标签更易获取，并且可以作为具有任意感知范围和分辨率的体素网格的监督。

将语义分割扩展为面向占用的全景分割。 语义占用预测对每个体素进行语义分割，但缺乏实例提取能力。语义分割与实例分割是相关任务，实例信息可以消除同一前景物体的语义预测不一致问题，以及相邻物体的预测混杂问题。PanoOcc [54]使用检测框从占用预测中分离实例。Symphonies [48] 和 PaSco [76] 使用端到端的基于Transformer的全景分割头，从多尺度SSC特征中生成实例。PaSco还提出了结合模型集成的体素不确定性估计，并展示了其在多种天气条件下的鲁棒性。OccFormer [44] 提出了一种双路径Transformer，以不同方向对三维体素进行编码，并采用改进的Mask2Former [77] 进行语义分割和全景分割。

3) 基于融合的占用网络： OpenOccupancy 为基于融合的占用网络引入了多模态融合基线（M-baseline）。M-baseline 由 Sp- Conv4x 激光雷达编码器、基于LSS的占用池化图像编码器、体素级融合器和预测头组成。M-CONET [28] 将 M-baseline 的预测头扩展为级联结构，并提升了准确度。OccFusion [63]是一种基于Transformer的融合框架，将激光雷达体素特征设为查询，将多视图图像特征设为键和值。OccFusion [78]在多尺度三维特征体上进行全局-局部注意力体素级融合。Co-Occ [42] 增加了一个额外的体素渲染分支，将体素特征转换为密度场和颜色场，并根据深度和像素颜色计算损失。OccGen [64] 提出了一种渐进式细化解码器，用于对三维高斯噪声图进行去噪，以获得预测目标。考虑到车辆到万物（V2X）的数据采集难度，多智能体协同融合的研究仍较少。PointSSC [79] 建立了一个基于激光雷达的SSC基准，其构建于 V2X-seq 数据集之上。CoHFF [80] 在 CARLA 仿真器中展示了多车辆混合特征融合带来的显著性能提升。

4）三维占用的隐式表示： 占用的显式表示在结构上便于并行化，但缺乏灵活性。大多数网络训练后只能在固定的感知范围和分辨率下运行。隐式表达则能够灵活地构建不同分辨率的占用栅格。有符号距离场（SDF）是一种重要工具，可根据与表面的距离对三维场景进行隐式表示。SIREN [81]提出了一种弱监督感知方法，可从点云补全三维网格曲面。LODE [82]提出了一种生成对抗网络（GAN），利用占用标签监督SDF。对于不以激光雷达为输入的视觉模型隐式学习，SurroundSDF [83] 提出了用于SDF建模的Sandwich Eiikonal 公式，可在表面两侧提供正确且密集的约束。

分析：三维占用预测中计算需求与精确率之间的权衡。普遍认为，感知范围越大、网格分辨率越高，精确率和计算需求也越高，反之亦然。目前，驾驶系统通常将占用预测视为备用系统，感知范围仅为 40 米至 70 米，分辨率也仅为 0.2 米至 0.4 米。Occ3D-Waymo [27]提供了分辨率为 0.1 米的体素标签，但很少有网络采用这一分辨率。远距离占用预测仍是一个尚未得到充分探索的领域。采用隐式表示的可变分辨率网络，以及多分辨率网络（近处采用较小体素网格，远处采用较大体素网格），都是可行的方向。网格大小也与具体工程设计有关，例如，城市区域的辅助驾驶需要精细的占用栅格，而卡车在高速公路上行驶时则需要远距离的占用网格。

C. 适于部署的三维占用网络

1) 三维占用网络的工业设计：特斯拉率先研究了可在嵌入式 FSD 计算机上 $10ms$ 运行的高性能、低延迟实时占用网络。特斯拉首先在 CVPR 2022 自动驾驶研讨会（WAD）上[84] 介绍了占用网络 [85],，随后在特斯拉 AI 日上展示了以网格为中心的感知系统的完整流程 2022 [86]。占用网络的模型结构如图 4 所示。首先，模型的骨干网络使用 RegNet [87] 和 BiFPN [88] 从多个摄像头获取特征；随后，模型通过结合三维空间位置的空间查询，对二维图像特征进行基于注意力的多摄像头融合。然后，模型根据提供的自车位姿对齐并聚合三维特征空间，从而执行时间融合。在融合不同时间范围的特征并经过反卷积模块处理后，解码器对体积与表面状态进行解码。体素网格与神经隐式表示的结合也值得关注。受 NeRF 启发，模型最后采用可接受任意坐标值的隐式可查询多层感知机解码器 x， y， z来解码有关

该空间位置的信息，即占用、语义和流。通过这种方式，占用网络能够以任意分辨率进行三维占用映射。

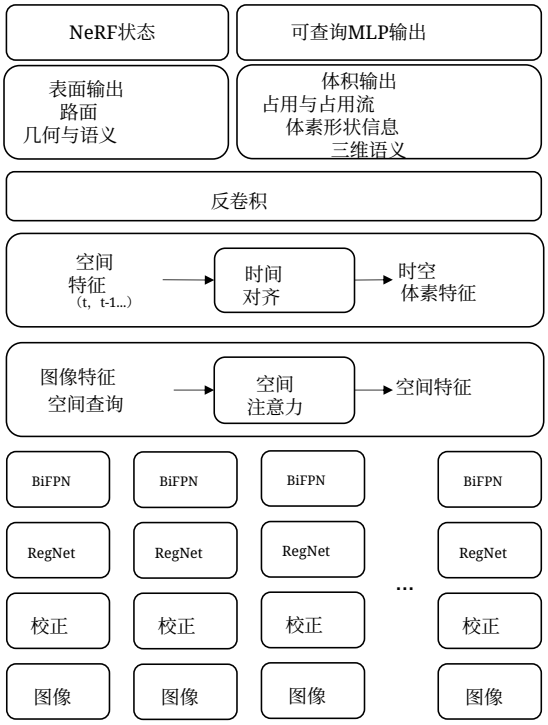

图 4. 特斯拉 FSD [86]：基于 OccupancyNetwork 构建的特斯拉 FSD Beta 感知系统。

2）计算高效的三维占用：**从透视视图到鸟瞰视图的高效视图变换**原始 LSS 需要复杂的体素累积和技巧，才能将概率深度特征投射到 BEV 空间中。后续工作 [89]–[91]主要通过设计面向体素网格的高效算子，优化原始 LSS [92] 的计算成本。LSS [92] 采用累积和技巧，将视锥特征排序到其唯一的 BEV ID 中，但在 BEV 网格上的排序过程效率较低。BEVFusion [90]通过预计算网格索引，并利用专用 GPU 内核在 BEV 网格上进行并行化的区间归约，提出了一种高效且无需近似的 BEV 池化方法。BEVDepth [91] 提出了高效的体素池化，为每个视锥特征分配一个 CUDA 线程，并将每个像素点对应到该线程。GKT [89] 利用几何先验引导 Transformer 关注具有判别性的区域，并展开核特征以生成 BEV 表示。为实现快速推理，GKT 引入了查找表索引，用于在运行时处理摄像头标定参数无关的配置。MatrixVT [93] 发现，特征传输矩阵（FTM）的稀疏性是造成低效率的根本原因。通过 Ring & Ray 分解方法，FTM 被正交分解为两个相互独立的矩阵，分别编码自动车辆在极坐标中的距离和方向。经过 Ring & Ray 分解后，VT 可以重排为数学上等价但效率更高的公式。MatrixVT 将 VT 的 GPU 内存和计算开销降低了数百倍。Fast-BEV[94] 是一种基于 M2BEV 的强大实时 BEV 算法 [95]提出了两种加速设计：预计算投影索引，以及投影到同一个体素特征上。WidthFormer [96]通过仅使用一层 Transformer 解码器来计算 BEV 表示，并结合新颖的参考位置编码（RefPE），实现了快速视图变换。

高效占用网络。为了将占用网络部署到车载设备上，研究者 [34], [97]–[99] 基于现有占用网络提出了不同的轻量化结构和加速方法。从模型角度来看，一项重要改进是用轻量级算子 [94], [99]取代稠密三维卷积。从数据角度来看，车辆周围超过 80% 的网格都是空气。PanoOcc 会在每一层预测非空掩码，在栅格上进行稀疏化，然后用稀疏卷积取代稠密卷积。SparseOcc [98]（上海交通大学）在 LSS 视图变换后移除 80% 的零值栅格，而 SparseOcc [34]（OpenDriveLab）则用于在稀疏体素上执行 Transformer。

分析：体素网格与鸟瞰视图网格上的特征表示对比。与鸟瞰视图特征相比，体素特征消耗更多内存和计算资源。对于基于三维体素网格的任务的实时应用，一种折中方案是将垂直维度视为鸟瞰视图柱状网格的潜在特征。这种“高度作为通道”的方法有时甚至优于稠密体素方法。一个可能的原因是，空体素占据了周围环境的大部分空间，巨大的不平衡会损害体素特征。

IV. 以网格为中心的时间感知

由于自动驾驶场景在时间上连续，序列信息也能自然地增强对真实世界的观测，因此，利用时空特征解码运动线索并预测未来状态，是以网格为中心的感知需要解决的重要问题。

A. 序列*BEV*特征的时间融合

激光雷达点云只需将所有点云转换到当前帧，即可轻松进行时间融合；但对于视觉模型，时间融合则需要更多技巧。对于激光雷达模型，研究人员发现，使用超过 1.5 秒的激光雷达序列会降低性能。为此，D- Align [100]设计了一种新颖的双查询协同注意力网络和时间上下文引导的可变形注意力，以实现帧间特征对齐。STEMD [101]提出了两阶段时间融合：利用卷积门控循环单元进行 BEV特征融合，并利用注意力进行目标查询融合。

对于视觉模型，一种常见做法是根据多帧相对自车位姿，将 BEV特征变换到当前帧，并通过时间模块进行处理。早期方法 [29], [102], [103] 使用简单的卷积模块进行时间聚合。BEVDet4D [104] 将变换后的空间拼接在一起，并通过定制的 ResNet 层进行处理。BEVFormer[53] 使用可变形自注意力融合变换后的 BEV 空间。UniFormer [105]认为，基于变换的方法是低效的串行方法，不支持长距离融合，并且会丢失感知

范围边缘处的宝贵信息。为此，UniFormer 提出在当前 BEV 与缓存的历史 BEV 之间的虚拟视角中进行注意力计算，从而扩大感知范围，并更好地建模长距离融合。总的来说，使用更多历史帧会显著增加训练和推理时间。为此，SOLOFusion [106]率先提出一种流式训练方法，每次仅使用当前帧更新历史状态，从而连续利用完整序列来感知当前帧。

*B. BEV*运动与流预测

直接在鸟瞰视图网格上推断速度的优势在于，速度估计处于更早的阶段，且比目标检测与跟踪流程中的估计更稳定，同时不受检测中的识别失败和跟踪中的关联失败影响。运动估计的主要挑战在于，网格之间不存在明确的对应关系。

真实标签生成。 生成网格流（场景流）标签的常见做法，是对相邻帧中带有唯一实例 ID 的三维边界框进行后处理。所有跟踪到的边界框都会变换到当前帧，分配唯一的实例标记，并映射到鸟瞰视图网格作为真实标签。

激光雷达模型。 点云天然存在于三维空间中，可在数据层面进行聚合。聚合需要精确定位，这可通过高精度全球导航卫星系统设备或点云配准方法（例如迭代最近点 [107]，正态分布变换 [108]）获得，从而将点云坐标转换到当前自车坐标系。以多帧点云为输入的特征提取骨干网络能够同时提取空间和时间维度的信息，从而降低计算负载。一种紧凑的设计是将点云体素化，将点云视为伪鸟瞰视图地图，并将垂直信息作为每个鸟瞰视图网格上的特征 [11], [109], [110]。MotionNet 提出了一种轻量高效的时空金字塔网络（STPN），用于提取时空特征。BE-STI 提出 TeSE 和 SeTE，以对特征进行双向增强。TeSE 用于理解各个单独帧的空间信息。SeTE 则利用具有空间判别性的特征提取高质量的运动线索。用于运动预测的解码器以空间—时间骨干网络的 BEV 特征作为输入。各预测头由 1 至 2 个卷积块堆叠组成。MotionNet [11] 中的预测头包括用于类别估计的网格分类头、用于速度估计的运动头，以及用于区分动态或静态网格的状态估计头。BE-STI [110] 采用类别无关运动预测头，进一步利用语义来提高运动预测的准确性。

在损失设计方面，空间回归损失以 L1 或均方误差范数的方式回归运动位移。分类则使用交叉熵损失。由于一致性天然由序列数据保证，MotionNet [11] 提出了针对属于同一物体的单元格的空间一致性损失，以及用于约束连续两帧之间运动的前景时间一致性损失。作为一种自监督框架，PillarMotion [109] 提出了一种自监督结构一致性损失，用于近似柱体运动场，并将跨感官损失作为辅助正则项，以弥补稀疏激光雷达输入所提供的结构一致性。

激光雷达与摄像头的相互监督有助于学习几何与运动。PillarMotion [109]在激光雷达分支中计算柱体运动，并利用自车位姿补偿光流。光流会引发概率运动掩码。光流与柱体流经过跨传感器调节，以提升结构一致性。对 PillarMotion 进行微调也能改善鸟瞰视图网格的语义与运动。

视觉模型。 视觉模型首先通过图像骨干网络和视角变换模块处理所有帧。随后，时序模块融合过去多帧的鸟瞰视图特征，如第 IV-A 节所述。占用流预测需要未来鸟瞰视图特征的状态表示。预测模块的主要组成部分是循环神经网络（RNN）的变体。FIERY [29] 提出了空间门控循环单元（SpatialGRU），用于将当前鸟瞰视图状态传播至近未来。ST-P3 [111]提出了双路径概率未来建模（Dual-GRU），以当前鸟瞰视图状态的两种不同分布为输入，从而获得更强的预测特征。BEVerse [112] 采用迭代流进行高效的未来预测，将上一帧鸟瞰视图特征输入当前帧预测。StretchBEV采用一种基于神经常微分方程（Neural-ODE）的变分自编码器，通过生成式方法学习时序动态。除时序模块设计外，视频Transformer也能作为良好的时序预测器。TBP-Former [113]提出了一种基于 Swin Transformer 的时空金字塔预测模型，可全面提取多尺度鸟瞰视图特征并预测未来鸟瞰视图状态。FipTR [114]提出了一种用于未来鸟瞰视图特征预测的流感知鸟瞰视图预测器，该预测器由流感知可变形注意力组成，并利用反向流引导偏移采样。

视觉预测头由一个轻量级BEV编码器（例如 Res Net18 [115]）和四个鸟瞰视图解码器组成。五个独立的解码器分别输出中心度、鸟瞰视图分割、到中心点的偏移量以及未来流向量。后处理单元将偏移量与中心点关联起来，根据分割结果形成实例，并根据多帧实例输出实例流。空间回归损失以 L1 或均方误差范数的方式回归中心点、偏移量和未来流向量。分类使用交叉熵损失。概率损失则对鸟瞰视图特征之间的 Kullback-Leibler散度进行回归。

C. 将占用流扩展到长期

给定真实历史数据作为输入，长期占用流可作为轨迹函数的有用补充。占用流任务通过序列流向量，将占用从遥远未来的网格追溯至当前时刻的位置。DRF [116] 使用自回归序列网络预测占用残差。ChauffeurNet [117] 通过占用多任务学习，为更安全的轨迹规划提供补充。道路规则 [118] 提出了一种动态框架，用于从占用流中解码轨迹。MP3 [119] 预测

运动向量及其对应的每个网格的可能性。Waymo占用流挑战赛的前三名参赛者是 HOPE [120],VectorFlow [121], 和 STrajNet[122]。HOPE [120] 是一种新颖的分层时空网络，配备了由潜变量增强的多尺度聚合器。VectorFlow [121] 得益于结合向量化和栅格化表示。STrajNet [122] 的特点是采用交互感知Transformer，在轨迹特征与栅格化特征之间进行交互。

D. 四维占用预测

四维占用预测旨在预测未来的占据状态，因此这项任务也被称为预测式世界模型。除 OccWorld [123] 会在目标坐标系中联合预测未来占用和自车轨迹外，大多数方法都会将未来标签转换到当前坐标系，以消除未来自车轨迹的影响。

启动四维占用预测任务的一种简单方法，是对鸟瞰图占用流进行维度扩展。Occ4cast [124]提出了首个 4DOcc基准 OCFBench。Occ4cast 以二值体素化点云为输入，并验证了三种时序基线的性能：4d-occ-forcasting [36],中的长短期记忆网络模块和三维卷积模块。Cam4DOcc 以多视角图像为输入，并验证了四种基线的性能：静态模型、点云预测、采用 3D提升的 BEV实例预测，以及采用 Conv2D时序预测器的端到端预测。ViDAR [37] 采用预测Transformer方法来获取用于解码的未来鸟瞰视图特征。

时序网络在短期预测方面表现有效，但其预测并非多模态，即只能预测一种可能的未来。生成式人工智能（GenAI）的兴起为多模态四维预测带来了新思路。在生成式人工智能的背景下，四维占用预测被表述为视频预测任务，其中每个柱体都被视为一个像素。COPILOT4D [125] 是一个两阶段生成式框架。第一阶段采用与 UltraLiDAR [126]类似的鸟瞰视图编码器、量化器和解码器。第二阶段将扩散策略 [127] 应用于 MaskGIT[128] Transformer。COPILOT4D 使用三种掩码策略来学习预测任务。结果显示，其在 1 秒预测中减少了 65% 的误差，并能在 3 秒预测中生成多样化的未来。作为一种类似的方法，OccWorld [123] 以自回归Transformer取代耗时的扩散过程，从而实时生成未来预测。

V. 以网格为中心的感知的标签高效学习

由于占用标签标注成本高昂，亟需开发标签高效技术，利用易于获取的弱标签或无标签来训练占用网络。另一方面，实践表明，当模型在占用场景下进行预训练时，预训练策略能够提升模型在所有感知任务上的表现。

A. 将占用作为下游任务的预训练

随着自然语言处理（NLP）领域大规模预训练取得巨大成功，自监督视觉学习受到了越来越多的关注。在二维领域，基于对比学习 [129],和基于掩码图像建模 [130], [131] 的自监督模型发展迅速，甚至能够超越全监督模型。在三维领域，研究人员已在激光雷达点云 [132]–[135]上开展自监督预训练。自监督任务的核心问题是设计预定义任务，以获得更强的特征表示。预定义任务可以源自时间一致性、判别式对比学习和生成式掩码学习。

研究人员发现，与在原始传感器数据上进行自监督学习相比，将占用预测用作代理预训练任务，是一种更稳健、泛化能力更强的预训练技术，适用于各种下游任务。SPOT表明，占用预训练是一种通用且可扩展的范式，能够在不同数据集之间迁移，因为占用并不局限于特定的激光雷达，而且在不同数据集中可以用相同的方式定义占用。因此，二维或三维网格可作为自监督学习三维几何和运动的理想中间表示。Voxel-MAE [136], [137] 定义了一种基于体素的任务，将 90% 的非空体素掩码，并以补全这些体素为目标。这种预训练提升了下游三维目标检测的性能。同样，BEV-MAE [138] 提出对鸟瞰视图网格进行掩码，并将其恢复作为预定义任务。MAELi [139]区分空闲空间与遮挡空间，并采用一种新颖的掩码策略，以适应激光雷达固有的球面投影。与其他基于掩码图像建模的预训练方法相比，MAELi显著提升了下游检测任务的性能。ALSO [140] 设定了一种新颖的预定义任务：预测查询点的三维占用，这些查询点沿着从原点到反射点的每条射线采样。对于每条射线，都会在反射点附近采样两个点作为查询点，其中一个位于表面外侧，标记为空闲；另一个位于表面内侧，标记为占用。该预定义任务能够补全障碍物表面，并提升三维检测和激光雷达分割任务的性能。

尽管激光雷达网络能够从占用预训练中学习，但视觉网络从三维和四维占用学习中获益更多。UniScene [141] 和 UniWorld [142] 使用当前帧和未来帧的体素化点云作为占用监督，并提升下游鸟瞰视图检测任务的性能。Drive World [143] 利用状态空间模型学习未来占用标签，并提升检测、跟踪、规划等所有驾驶任务的性能。ViDAR [37] 是一种基于 Transformer 的预测器，通过潜在BEV渲染学习未来点云，并显著提升所有下游驾驶任务的性能。

B. 自监督三维占用预测

全监督三维占用网络高度依赖于基于昂贵激光雷达分割标签构建的训练标签，其可扩展性极为有限，难以扩展至

没有激光雷达的自动驾驶车辆。为此，提出了弱监督和自监督占用网络，以摆脱来自其他传感器的监督。

深度估计是占用预测的一项重要前置任务。在三维视觉领域，自监督深度估计由来已久。对于单目情况，MonoDepth2[144] 以新视角合成的方式，从单目视频中联合预测自车位姿和深度图。对于多视角情况，SurroundDepth [145] 使用跨视角变换器（CVT）捕捉不同摄像头之间的线索，并利用运动恢复结构算子生成的伪深度。NeRF [66] 也是一种颇有前景的深度几何自监督方法。Behind the scenes [146] 和 SceneRF [147] 提出了一种新颖的深度合成方法：通过优化多层感知机辐射密度场，使其能够利用同一序列中的其他帧推断源帧图像的深度。采用神经渲染训练的深度网络表现出更小的相对深度误差和绝对深度误差。

利用图像进行重建是三维几何领域中一项广泛研究的课题。将几何知识注入实时感知网络是一项新兴挑战。神经渲染可作为从二维信息到三维几何的弱信号，用于监督三维几何。其基本思路是利用多层感知机构建体素特征的密度场和语义场，沿射线采样属性，将属性整合到像平面，并根据像平面信息计算损失。UniOcc [57] 和 RenderOcc [47] 使用图像RGB像素、激光雷达深度以及激光雷达分割作为弱监督，从而提升占用预测性能。S4C [148],SelfOcc [149]和OccNeRF[150] 仅使用视频序列结合NeRF进行几何重建。此外，OccNeRF使用grounding-dino和Segment Anything模型（SAM）[151]生成高质量的图像分割结果。由于多视图立体视觉受到静态场景的严重限制，大多数基于NeRF的自监督学习方法会利用相邻帧中的静态类别，并仅利用当前帧中的动态类别。这种方法极大地限制了动态物体的几何重建。为更好地利用物体轨迹，OccFlowNet [152]会根据真实物体轨迹预先计算场景流，然后利用流偏移移动动态体素索引，最后使用RenderOcc [47]进行弱监督占用学习。

C. 自监督四维占用预测

四维预测标签成本高昂，而且在大多数情况下无法收集，因为传感器数据只能反映世界表面。为此，4d-occ-forecasting 使用未来点云作为未来占用的替代数据，因此可以在任何具有激光雷达点云序列的数据集上进行评估。其思路很简单：给定未来传感器原点，预测的未来点云就是沿真实射线投影到最近的、预测为占用区域的点。作者提出了一种可微体素渲染层，通过比较同一方向上的真实射线长度和预测射线长度来计算损失。ViDAR [37] 进一步将二值体素射线渲染发展为鸟瞰图潜在射线渲染，即沿射线对体素值进行积分，作为射线的长度与 4d-occ-forecasting 相比，ViDAR 的方法并未显式构建二值四维占用网格，而是更多地作为强大的鸟瞰视角预训练策略，以提升下游任务的性能。

D. 开放词汇三维占用预测

大多数占用网络会在封闭集类别中预测语义类别，并将其他物体归为一个通用类别，但道路上出现的障碍物种类层出不穷。开放词汇占用网络旨在根据输入的文本提示检测任意障碍物。常见方法是将体素特征蒸馏到视觉语言教师模型中，例如 CLIP 及其变体。OVO [153]是借助高质量体素选择进行开放词汇语义场景补全的首次尝试。POP-3D [154]基于 TPVFormer [43]构建，并利用 MaskCLIP [155]教师模型开展三平面自监督训练。

VI. 驾驶系统中的栅格中心感知

作为空间环境的一种高层表示，占用栅格隐含着丰富的感知信息，可作为自动驾驶规划的基础，例如无碰撞的可行驶区域、道路结构信息、空间风险分布等。

具体而言，以栅格为中心的感知结果可以通过量化代价或对规划结果进行安全检查，充当轨迹评估器。此外，通过基于网格构建目标函数和约束，它还可用于轨迹优化。另外，将栅格化环境表示作为驾驶策略网络的输入也是一种常见的规划范式，能够对环境进行统一描述。

由于网格具有稠密表示和丰富的空间语义描述能力，因此在自动驾驶的下游任务中展现出通用性、空间细节丰富、可解释性强等优势。

A. 栅格中心轨迹评估

基于栅格的环境表示对离散空间位置的安全相关信息进行编码，例如智能体占用和道路区域。因此，它天然适合作为航点评估器：可用于对采样轨迹执行碰撞检查，以筛除不可行轨迹；也可用于对候选轨迹进行评分，从而选出最优方案。传统的基于规则的规划算法，如基于采样的 [156], [157] 或基于搜索的 [158] 方法，通常利用占用栅格检查轨迹的碰撞风险。近年来，基于学习的方法逐渐出现，用于构建逐栅格环境表示，以评估并为一系列候选轨迹进行评分。NMP [159] 提出了一种可解释的端到端规划器，其中骨干网络输出代价体，用于表示时空占用，并作为安全指标从采样轨迹中进行选择。类似的规划

框架也被P3 [160], MP3 [119], ST-P3[111], OccNet [161], 采用，它们均构建语义占用栅格的时间序列，用于指示动态物体和静态地图，并利用这些栅格设计代价函数，对采样的候选轨迹进行评估。QuAD [162] 不再预测显式的稠密三维栅格地图，而是采用以鸟瞰图潜在表示为输入的隐式占用模型，以查询感兴趣航点的代价，从而显著降低计算负担。UAP-BEV [163] 进一步学习鸟瞰图预测的不确定性，并对扰动进行采样，以提升驾驶安全。

此外，占用在自动驾驶规划中的另一种常见应用，是在训练过程中构建安全感知损失。例如，PLUTO [164] 通过辅助损失惩罚超出栅格化可行驶区域的轨迹，而 NMP [159] 则引入最大间隔损失，以促使专家轨迹具有最低成本。由于损失的构造形式也是对规划结果进行评估的标准，因此，这类方法也被归类为轨迹评估。

B. 以栅格为中心的轨迹优化

不同于对采样轨迹进行评估的方法，轨迹优化是指基于以栅格为中心的环境表示构建数值优化问题，以求解最优轨迹。对于这一非凸非线性问题，大多数算法都需要一个初始解作为优化的起点。因此，许多传统的运动规划算法会基于采样算法（如快速扩展随机树）提供的初始路径对轨迹进行优化，从而确保轨迹的运动学可行性，并提供可解释的安全保证。这通常需要一个可微占用栅格 [165], , 该栅格也常被称为风险势场 [166], [167] 或热力图。

出于相同原因，在许多基于学习的规划框架中，通常会采用轨迹优化器作为神经规划器的后处理组件，以细化轨迹。基于占用预测，UniAD [168]和 Hoplan [169] 均利用高斯分布构建一个与碰撞风险相关的目标函数，并在优化过程中使用该函数提升网络生成的初始轨迹的安全性。Gameformer规划器[170] 定义了一种占用栅格，仅表示纵向维度上的时间占用，并在后处理优化中确保车辆位置不超出安全区域。OPGP [171] 根据占用预测推导出潜在碰撞代价，并对低于安全阈值的纵向距离施加惩罚。

C. 神经网络策略的状态表示

以网格为中心的感知结果能够持续描述环境的空间信息，并常被直接用作神经网络规划器的输入。它们通常组织为多通道栅格图像，每个通道具有不同的物理含义，或表示不同的历史步。例如， [172] 将观测封装为一个三通道栅格图像，分别表示占用、速度和航向，随后由 [173]添加另一个位移通道。 [174] 定义了一种归一化多通道栅格状态，其中不同的运动学特征按给定范围进行缩放。 [175] 则将连续两帧的占用情况作为状态输入。此外，TOFG [176]提出为基于图注意力网络的模型构建时序占用流图，其中不仅明确建模了网格之间的连接，也建模了不同时间步之间的连接。ChauffeurNet [117] 将所有观测渲染为栅格化鸟瞰图像，作为模型输入，其中包括道路地图、路线、动态框等。You 等人 [177]没有采用高分辨率网格，而是关注周围九个栅格单元的占用情况，并以车辆大小作为粗粒度尺度进行状态表示。

D. 生成式世界模型中的以网格为中心的表征

自动驾驶领域的最新世界模型 [178]–[182]可以通过生成式框架模拟未来的多视角图像。它们通常在扩散框架中使用高精度矢量地图和三维边界框作为控制输入。然而，这些紧凑元素限制了场景元素的多样性。将占用作为中间表征或最终输出，可以为仿真引入更多样性。XCube [183]提出了一种新颖的生成方法，采用稀疏体素层级，从 Waymo 数据集的单帧点云生成高分辨率网格曲面。WoVoGen [184]将三维占用标签用作世界体积，并联合生成未来占用和未来多视角图像。UrbanDiffusion [185] 提出了一种新颖的扩散模型，以鸟瞰图布局为输入，在三维占用层级生成无界城市场景，以保留场景几何和语义信息。SemCity [186]提出了一种三平面扩散网络，可通过图像内补和图像外补实现大规模城市语义场景补全。

这些世界模型通常与基于模型的规划方法相结合，通过数据增强在想象中学习驾驶策略。例如，MILE [187] 提出了一种基于模型的模仿学习框架，可迭代生成未来状态和动作，并将其解码为鸟瞰图占用；而 Think2Drive [188] 则利用该模型合成经验数据，用于强化学习训练。

分析：与以对象为中心的方法相比，以网格为中心的表征在规划中具有优势。与以对象为中心的处理流程相比，以网格为中心的感知无需显式进行对象不确定性建模，并能自然地表示全空间的不确定性；与在预测或规划中直接回归矢量化轨迹相比，它能够内在地处理多模态问题。作为一种状态表示，它消除了对象列表所带来的排列不变性问题 [189]；除了上述优势之外，基于以网格为中心的感知进行规划也具有很强的可解释性。

七、讨论

本节将深入总结以网格为中心的感知当前发展趋势，并对其未来进一步发展的方向作出一些展望。

A. 当前趋势的观察与局限

与目标级感知流程相比，以网格为中心的方法对物体形状的几何假设较少，并且在描述处于任意遮挡状态的物体时更加灵活。以网格为中心的方法已成为车辆感知系统不可或缺的组成部分。以下从三个角度概述当前趋势。

二维/三维/四维占用网格上的特征表示。与传统的基于概率的占据栅格地图相比，深度学习显著增强了描述网格语义和运动的能力。特征表示能力在很大程度上取决于网络结构。过去两年中，研究人员广泛探索了如何将激光雷达、视觉和雷达原始数据中的特征表示为鸟瞰视图和三维占用网格。相较于未来帧预测，时序模块在历史信息融合方面得到了更多研究。

高效学习。 在主流的占用监督学习中，目前的标签生成成本很高，这限制了占用模型的性能上限。由于网格标签通常代价高昂，人们期待标签高效学习（例如半监督、弱监督或自监督学习）能够加速未来解决方案的发展，以应对开放世界交通场景；不过，这类方法在三维领域仍处于起步阶段。将占用学习用作预训练任务，为下游感知与规划任务提供基础模型以提升性能的趋势日益明显。

在驾驶系统中的应用。我们观察到，以网格为中心的感知在整个自动驾驶系统中发挥着越来越重要的作用。以网格为中心的感知具备强大的障碍物表示能力，并能直观地指示碰撞风险。依赖网格的规划在下游任务中有着悠久的历史。一种新兴趋势是，端到端规划方法展现出巨大潜力，能够将网格特征传递给环境认知模块，并构建更准确的安全场。

当前方法的局限性。除了该领域的快速发展之外，我们还总结了当前方法的局限性。

在三维占用预测中，我们需要可扩展且经济实惠的占用标签生成流程。神经渲染 [68],扩散[184] 或来自仿真的虚拟数据 [190]等技术可能有所帮助。对于基于激光雷达或相机数据的占用模型，需要通过轻量化设计提升模型效率，以满足实时部署要求。利用其他传感器的三维占用网络鲜有研究。

在时间预测和四维预测中，大多数占用预测方法侧重于当前帧及近期未来的运动估计，因为占用预测主要依据惯性进行预测。预测长期（最长可达 8 秒）的未来时，还应考虑更多社会交互因素，而这正是轨迹预测的优势。兼顾所有物体的惯性与已知智能体之间交互的占用预测和轨迹预测相结合的方法，仍未得到充分探索。

在标签高效学习中，大多数自监督学习技术仍使用激光雷达点作为监督信号。然而，许多车辆仅配备多视角相机。使用激光雷达作为预训练目标会限制数据和模型的可扩展性。另一方面，目前仅依赖视频序列的占用网络性能较弱，在 Occ3D-nuScenes 基准上的表现约为 $10mIoU$，仅达到全监督视觉模型的约 25%。如何利用不含激光雷达的大规模视频数据训练占用网络，仍有待探索。

在基于以网格为中心的感知进行规划时，如何设计从占用网格到最新决策技术的适配器仍然颇具挑战。这主要是因为占用网格的维度过高。除了用于实现碰撞规避的后处理模块外，规划模块仍难以充分利用占用网格中丰富的感知信息。

B. 未来展望

以网格为中心的感知有望为三维视觉、几何和移动机器人学带来跨学科研究机遇。例如，统一的占用三维基础模型有望在面向不同用途的机器人学任务中实现良好的泛化。

1）相关新兴技术：占用任务与三维视觉和三维几何密切相关。它还有望结合生成式人工智能技术，用于预测和仿真任务。本节简要介绍可能相关的新兴技术。

高斯泼溅。继 NeRF 取得成功之后，三维高斯泼溅（3DGS） [67] 作为一种新型体积渲染技术，将三维场景表示为一组高斯。泼溅过程会在适当的位置放置这些高斯，并设置其强度，以渲染新视角，从而呈现场景的几何和外观。3DGS 的渲染速度显著更快，且新视角合成质量更高。SuGaR [191] 首次提出了表面对齐的 3D-GS，以获得场景的网格表示。

利用生成模型进行视频生成。 视频生成是一种新兴技术，可根据图像或文本自动创建视频内容。视频生成与世界模型这一概念密切相关。这些模型通常基于深度学习框架，并利用生成对抗网络（GANs）、变分自编码器（VAEs）、循环神经网络（RNNs）和潜在扩散模型等技术，生成一系列视频帧。受 OpenAI 的 SORA [192] 启发——SORA 使用基于视频 Transformer 的扩散技术生成时长 60 秒且几何一致的长视频——OccSora [193] 是一个结合轨迹提示的四维世界模型，能够生成未来

最长 16 秒的占用视频。值得注意的是，视频生成与神经渲染的结合将实现几何一致且逼真的三维场景生成。

2）迈向三维占用的高效表示：在真实世界驾驶场景中，周围近处环境通常比远处环境具有更高的风险，因此障碍物逼近时需要提高警惕。在以网格为中心的感知中，一个自然的需求是：近处网格的分辨率应高于远处网格。进一步而言，在无需重点关注的区域（例如远处或被遮挡的区域），网格特征表示存在冗余。视觉Transformer和隐式表示是实现按需可变粒度的有前景技术。另一个需要考虑的问题是如何为下游任务确定安全粒度，因为这与感知需求分析密切相关。

3）迈向无标签的可扩展占用预训练：与以目标为中心的感知相比，以网格为中心的感知需要更严格的标注。激光雷达点云受最小角分辨率和最大高度的限制，这意味着远处的障碍物可能无法被准确标注。针对网格的三维视觉标签高效学习亟待开展。目前，自监督学习仅在小型数据集上经过测试，且精度非常有限。扩大占用预训练的规模，是构建三维视觉基础模型的一种有前景的解决方案。

4）迈向四维占用世界模型：四维占用是一个新兴领域，目前相关实践有限。现有方法中，时序模型难以处理长序列，而生成模型难以实现实时部署。四维占用世界模型在对感知的最终输出进行建模方面潜力巨大，能够涵盖其中的所有内容。此外，世界模型通常采用自监督方式训练，更有利于扩大规模。

5）借助占用先验迈向开放世界语义理解：占用网格提供了强大的尺度感知几何先验，有助于实现开放世界语义理解。GOOD [194] 是一个利用尺度模糊的深度信息来更好地检测未知物体的例子。另一种方法是将开放词汇语言感知特征蒸馏为体素特征。另一方面，当不可通行障碍物（例如躺倒的人）与地面融为一体时，占用栅格也面临棘手情况；此时需要准确的开放世界语义理解，以区分带有凸起或凹陷、但仍可通行的空闲区域与不可通行的路面异常。重建和语义理解能够从彼此的改进中获益良多。

6）探索更适用于以网格为中心的感知的规划适配器：由于难以从栅格图像中提取实例级特征信息，现有方法在规划时往往缺乏社会交互处理。如何显式或隐式地对网格区域之间的交互模式进行建模，值得深入探讨。此外，网格表示天然适合对不确定性进行建模并将其应用于感知和预测，但事实上，很少有规划算法关注这一优势。基于不确定性感知占用网格的应急规划可能是一个有前景的研究方向。此外，采用高维网格状态表示的神经网络策略可能难以实现良好的泛化，还会带来巨大的计算负担，尤其是在使用网格时间序列时。

C. 以网格为中心的感知的更广泛影响

我们讨论以网格为中心的感知可能带来的伦理影响和更广泛的影响。以网格为中心的感知可用于广泛的自动化应用，包括移动机器人和自动驾驶车辆。与目标检测模型相比，以网格为中心的感知模型在匿名化和隐私保护方面具有更好的表现，因为它不需要在标注文件中对每个个体进行详细标注，而是以占用网格的形式存在。

VIII. 结论

本文全面回顾和分析了自动驾驶领域成熟且新兴的以网格为中心的感知技术。本文首先介绍以网格为中心的感知任务的背景，包括数据集、基准和评估指标。我们从多模态传感器的鸟瞰图表示逐步介绍到三维网格表示。针对以网格为中心的感知中的时序学习进展，我们回顾了历史信息融合、鸟瞰图运动和四维占用流。随后，我们深入研究标签高效学习，以利用少量标签学习占用网络，从而实现通用预训练。我们回顾了规划方法如何受益于以网格为中心的感知结果。最后，我们总结当前的研究趋势和当前技术的局限性，并介绍以网格为中心的感知的新兴技术、未来展望及其更广泛的影响。作者希望本文能够展望自动驾驶车辆中以网格为中心的感知的未来发展。

附录

A. 占用栅格地图构建基础

1）自动驾驶车辆的传感器输入：自动驾驶车辆在环境感知中高度依赖多个摄像头、激光雷达传感器和雷达传感器。摄像头系统可由单目摄像头、双目摄像头或两者组成。其成本相对较低，并可提供高分辨率图像 $I_{cam} \in \mathbb{R}^{H\times W\times 3}$，包括纹理和颜色信息。然而，摄像头无法获取直接的三维结构信息和深度估计。此外，图像质量高度依赖于环境条件。

激光雷达传感器以点云的形式生成场景的三维表示 $I_{LiDAR} \in \mathbb{R}^{N\times 3}$，其中 N 是场景中的点数，每个点包含x，y，z坐标，以及反射强度等额外属性。由于具备深度感知、更宽的视场和更大的探测范围，激光雷达传感器在自动驾驶中的应用更加

广泛，并且较少受到环境条件的影响。遗憾的是，其应用主要受成本限制。

雷达传感器是自动驾驶领域最重要的传感器之一，这得益于其成本低、探测范围远，以及在恶劣环境中检测运动目标的能力。雷达传感器返回包含目标相对位置和速度的点。然而，雷达数据较为稀疏，且对噪声更为敏感。因此，自动驾驶车辆经常将雷达数据与其他传感器输入结合起来，以提供额外的几何信息。

2) 占用栅格地图：占用栅格地图源自机器人环境表示领域 [195],**占用栅格地图（OGM）**主要用于处理雷达测量的不确定性，并以稳健、统一的方式对复杂环境建模。早期的机器人通常配备单扫描激光雷达或多扫描激光雷达。由于雷达测量中不可避免地存在误差，因此提出了 OGM，将环境划分为离散且相互独立的网格，并通过逆传感器模型估计每个网格的占据概率 [196] [197]。

经典占用栅格地图将时间 k 的网格状态 O_k 描述为两种互斥的二元事件：占用或空闲，即 $O_k \in \{0, F\}$。当时间 $k+1$ 的测量到达时，OGM 会更新网格的状态。对于户外场景，多扫描激光雷达需要使用随机抽样一致算法进行地面去除预处理 [198] ，并采用成熟的聚类技术 [199] ，以获取仅表示非地面场景的点云，作为输入测量。随后，逆传感器模型根据时间 $k+1$ 的测量z_{k+1}，为每个网格分配离散的二元占据概率$p_{z_{k+1}}(o_{k+1} \mid z_{k+1})$；这一结果称为测量栅格地图。

图 5 展示了一个简单的逆传感器示例。网格的占据概率是距离的函数 r，假设测量结果的不确定性会随着距离增大而增加。因此，逆传感器模型会为反射点附近的网格分配较大的占据概率，而为机器人与反射点之间区域内的网格分配较小的占据概率，该区域称为自由空间。该算法对每束雷达波束分别建模，但这并不总是符合实际情况。研究人员通常会根据具体任务需求设计不同的逆传感器模型 [200] [201]，或使用机器学习算法推导模型参数 [202]。

占据栅格地图依据贝叶斯滤波机制更新后验占据概率。给定来自传感器逆向模型的时刻$k+1$测量值，通过二值贝叶斯滤波器（BBF）[197]估计每个网格在时刻 $k+1$ 的二值状态：

$$p_{o,k+1}(o_{k+1}) = \frac{p_{z_{k+1}}(o_{k+1} \mid z_{k+1}) \cdot p_{o,k}(o_k)}{p_{z_{k+1}}(o_{k+1} \mid z_{k+1}) \cdot p_{o,k}(o_k) + p_{z_{k+1}}(f_{k+1} \mid z_{k+1}) \cdot p_{o,k}(f_k)} \tag{9}$$

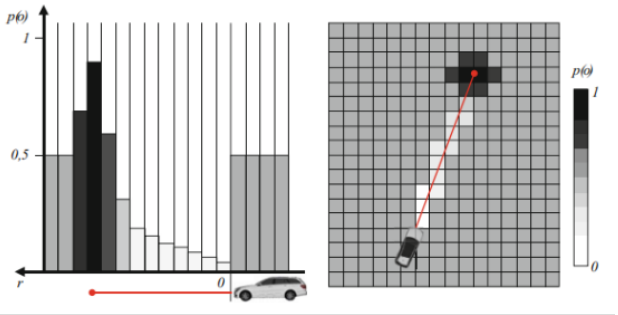

图 5. 单次激光雷达反射的逆传感器模型示意图，展示侧视图（左）和鸟瞰图（右） [203]

$$p_{o,k+1}(o_{k+1}) = \frac{p_{z_{k+1}}(o_{k+1} \mid z_{k+1}) \cdot p_{o,k}(o_k)}{p_{z_{k+1}}(o_{k+1} \mid z_{k+1}) \cdot p_{o,k}(o_k) + p_{z_{k+1}}(f_{k+1} \mid z_{k+1}) \cdot p_{o,k}(f_k)} \tag{10}$$

其中，p（o）和 p（f）分别表示占用状态或空闲状态的概率。式 9 成立的前提条件包括：占用状态和空闲状态的先验概率相等，各网格的测量值相互独立，且状态不随时间变化。

占据栅格地图具有以下优点：(a) 以统一的概率形式进行环境感知，并具有坚实的理论基础；(b) 能够对任意形状的复杂障碍物建模；(c) 直观、清晰地估计自由空间（可行驶空间）；(d) 为融合不同传感器数据提供便捷的框架。

然而，固定分辨率占据栅格地图存在明显缺陷，阻碍了其在自动驾驶领域的应用。(a) 需要合理设置分辨率，以平衡精确度与计算负担。(b) 稀疏且含噪的激光输入会导致大量未知区域 [204]。公式 9 的一个重要前提是，网格的测量彼此完全独立，也就是说，当网格中没有雷达射线反射点时，该网格将被视为未知状态。点云的稀疏性意味着环境中的大多数网格都处于未知状态。(c) 静态占据栅格地图无法应对动态场景。移动车辆会在静态占据栅格地图上留下轨迹。

3）连续占据地图：连续占据地图是一种新颖的方法，它使用连续占据概率核，从而允许对占据进行任意分辨率的表示。连续占据地图首先将点云划分为空闲或占用部分，并根据这些具有已知状态的点或片段更新核函数中的参数，然后利用获得的核函数估计附近任意位置的占据状态，实现对整个空间连续占据率的估计。高斯过程占用地图（GPOMs） [205] 采用经修改的高斯过程分类器模型，作为一种非参数贝叶斯技术，同时引入地图中各点之间的相关性，以相关方差地图对真实环境中占据估计的连续概率表示进行建模，尤其适用于那些

具有稀疏点云的区域。但其时间复杂度相对于训练点数量为 $\mathcal{O}\left(N^3\right)$，这给实时应用带来了困难。希尔伯特地图 [206] 基于核特征近似将原始数据投影到希尔伯特特征空间，并使用随机梯度下降训练逻辑回归分类器。该算法对离群点具有鲁棒性，且更新时长随数据点的增长呈线性变化，在连续占据估计效果与实时性能之间取得了良好平衡，但它无法像GPOM那样计算估计结果的不确定性。

4）动态占据地图：在交通场景中，占据地图需要表示行人和其他车辆等动态障碍物的占用情况，同时估计其运动状态，以进一步预测未来轨迹并规避驾驶风险。

最直观的方法是将静态栅格地图与多目标跟踪技术相结合：使用占据栅格地图表示静态障碍物，并通过**运动物体检测与跟踪 (DATMO)** 对动态障碍物进行建模。早期方法通过检查栅格地图与新增观测之间的冲突来获取动态障碍物的范围，并将这些冲突区域作为多目标跟踪算法的输入，再从静态栅格地图中相应地移除，从而生成稳健的静态占据栅格地图[207]–[211]。然而，采用两种不同的环境表示来建模统一场景，主要难点在于数据关联。关联时，静态栅格结构与动态障碍物的多边形列表可能不一致。此外，对动态障碍物的形状假设限制了系统的估计精度和适用范围，这与占据栅格地图能够直观表示任意障碍物的优势相悖。针对数据关联问题，有研究尝试对冲突栅格进行聚类，并将其直接关联到多物体跟踪，以分别跟踪各个部分 [212],；也有研究为每个动态物体构建专属的局部栅格地图 [207]。

基于运动物体检测与跟踪（DATMO）的方法在表示动态物体时面临的另一个关键问题是，动态障碍物的检测模型和形状模型都必须经过充分训练[200], ，因为此阶段产生的误检和漏检会沿着整个系统流程逐级传播，导致实际应用中的未知物体和特征可能对算法性能产生重大影响。因此，更合理的方法是直接改进占据地图本身，从而实现对动态物体和静态物体统一且一致的表示。

动态占据栅格地图（DOGM）将动态目标纳入栅格地图，并估计与网格对应的障碍物占据概率及其速度状态，从而提升机器人或自动驾驶车辆在复杂动态环境中的感知和避障能力。

这些出版物中的方法各不相同。对于连续占据地图，一些研究将动态障碍物假设应用于现有的静态占据图结构。例如，[213] 使用基于光流的运动地图来估计网格的速度，并改进了 GPOM[205] ，使其能够适应动态环境。在 [214], 中，对激光雷达点云进行聚类和筛选，以估计动态障碍物的速度，并用该速度在希尔伯特空间中生成非平稳核，从而构建动态希尔伯特地图 [206]。

贝叶斯占据滤波器（BOF） [215], [216] 完全基于网格，不依赖任何物体假设。BOF 使用四维网格估计二维环境，其中两个网格维度表示空间位置，另外两个网格维度表示障碍物的二维速度。因此，BOF 会估计物体的运动，并在其过程模型中显式考虑该运动。它利用每个网格的直方图估计离散网格速度分布，这意味着在预测过程中 [216], 速度等于精确的整数网格位移。然而，这种速度离散化会产生伪影并限制准确度。此外，每个网格都需要对所有可能的前驱网格及其速度求和，计算开销很大，尤其是在需要高精度表示环境时。

近年来出现的一类方法以粒子作为地图的基础，从一开始就是为动态环境而设计的。在基于粒子的方法中，障碍物被视为一组点目标，并使用带有速度的粒子来模拟点目标。基于粒子的方法在理论上源自序贯蒙特卡洛（SMC）滤波器 [217], [218]。整个动态网格由一个粒子群表示，其中每个粒子代表一个可从一个网格移动到另一个网格的独立假设。粒子状态由位置和速度定义，并根据其运动模型和参数在网格中进行预测。物体根据其位置直接关联到相应网格，从而影响该网格的占用和速度分布。占据概率由网格中的粒子数量描述。

文献 [219] 将随机有限集（RFS）理论引入基于粒子的地图，并利用概率假设密度（PHD）滤波器和伯努利滤波器推导出一种地图构建流程。随后，采用邓普斯特-谢弗（DS）证据理论更新动态网格的占据状态，从而显著降低计算开销。

在同时存在动态和静态障碍物的杂乱环境中，有必要对动态或静态多点目标进行建模，并且可能会出现不可忽略的噪声。为减少噪声，通常采用两种方法。第一种方法是使用混合模型 [220]–[222], ，其中包括一个独立的静态模型和一个动态模型，以分别更新静态点目标和动态点目标的状态。这种混合模型可用于双重PHD滤波器 [220] 或网格级推理 [221], [222]。另一种方法是在更新过程中应用额外信息以减少噪声，包括物体高度 [223],、灰度强度 [223], [224],、语义信息 [224] 和物体标识符 [225], [226]。

总之，现有的基于粒子的 DOGM 前景广阔，但存在两个局限。第一个局限是

大多数解决方案都利用粒子的整体（多模态概率密度），来估计多个被视为相互独立的网格的状态（每个网格由其自身的概率密度描述）。粒子可以脱离原先的源网格密度，从一个网格迁移到另一个网格，并被重新分配到新的目标网格密度。由于每次迭代都会将粒子重新组合成新的独立网格估计器，因此会丢失被跟踪网格的身份。不过，如果采用额外的启发式方法，就可以保留这类信息。换言之，粒子通常能够识别物体的存在，却无法识别“特定”的物体部分或网格。为了改进估计和数据关联， [225], [226] 的作者采用了将粒子与物体关联起来的思路，为粒子添加物体 ID。需要指出的是，所提出的方法仍然忽略了网格识别，必须将每个粒子重新映射到每个网格。

第二个局限是，属于大型均匀网格区域的网格通常难以准确估计，且不确定性较高。例如，预测位于较大物体中间的粒子可能会附着在该物体的任意一个占用网格上。这会导致数据相关性模糊，最终增加不确定性。为了提高均匀网格区域（即具有相同占用值的大型物体中间区域）的估计精度，需要使用来自不同传感器的额外特征、更大的斑块或更多粒子。

5）内存高效三维网格映射：在分辨率较低的大规模场景中，内存是三维占用映射的主要瓶颈。显式地图表示方法有多种，例如体素、网格曲面、曲面元、体素哈希、截断符号距离场（TSDF）和欧几里得符号距离场（ESDF）。传统的体素占用网格地图按索引查询存储，需要占用大量内存，因此在映射方法中并不常用。网格曲面存储障碍物的表面信息。曲面元由点和斑块组成，其中包含半径和法向量。体素哈希是对传统体素方法的一种内存高效改进，它只对摄像头测得的场景表面进行体素划分，并以哈希表的形式存储场景表面的体素块，以便查询体素块。Octomap [227]提出了一种基于八叉树的高效概率三维映射框架。Octomap 会迭代地将立方体空间划分为八个小立方体，其中大立方体成为父节点，小立方体成为子节点；八叉树可以不断向下扩展，直到达到最小分辨率，此时的节点称为叶节点。Octomap 使用概率描述，以便根据传感器数据轻松更新节点状态。

连续映射算法是另一种替代方案，可用于构建兼具计算和内存效率、分辨率任意的三维占用描述。高斯过程占用地图（GPOM）采用改进的高斯过程作为一种非参数贝叶斯学习技术，通过在地图上的点之间引入依赖关系来保证连续性。希尔伯特地图 [206] 将原始数据投影到希尔伯特空间，并在其中训练逻辑回归分类器。BGKOctoMap-L [228] 扩展了传统计数模型 CSM，并通过核函数对其进行平滑后，周围体素的观测也可以纳入考虑。AKIMap [229]基于 BGKOctoMap，其改进之处在于核函数不再基于径向，而是能够自适应地改变方向并适应边界。DSP-map [230]将基于粒子的地图推广到连续空间，并构建了可适用于室内和室外场景的连续三维局部地图。广义而言，NeRF 系列中的多层感知机结构也是一种用于三维几何的隐式连续映射，几乎不需要存储空间。

B. 鸟瞰图二维网格表示

鸟瞰视图网格是道路车辆障碍物检测的一种常见表示。以网格为中心的感知的基本技术是将原始传感器信息映射到鸟瞰视图网格单元，不同传感器模态的映射机制各不相同。激光雷达点云天然以三维空间形式表示，因此在BEV地图上提取点特征或体素特征由来已久 [231], [232]。摄像头富含语义线索，但缺少几何线索，因此三维重建是一个不适定问题。视觉鸟瞰感知的兴起为透视视图到鸟瞰视图的视图变换（PV2BEV）带来了新的思路。

1）基于激光雷达的网格映射：激光雷达点云的特征提取遵循以下范式：点、体素、柱体、距离视图或上述混合特征 [1]。本节重点介绍点云到鸟瞰视图网格的特征映射。

在三维空间中采集的激光雷达数据可以轻松转换为鸟瞰视图，并与多视角相机的信息融合。激光雷达点云稀疏且密度不均，使卷积神经网络效率低下。一些方法 [233]–[235] 将点云体素化为均匀网格，并以手工设计特征对每个网格单元进行编码。MV3D [233]、AVOD [234] 通过对每个网格的高度、强度和密度特征进行编码来生成鸟瞰图表示。PIXOR[235]中的鸟瞰图表示由三维占用张量和二维反射率图组合而成，并将高度信息保留为通道。BEVDetNet [236] 进一步将基于鸟瞰视图的模型延迟缩短至Nvidia Xavier嵌入式平台上的2毫秒。针对网格上的高级时序任务，MotionNet提出了一种新颖的时空编码器STPN [11]，用于将过去的点云与当前自车位姿对齐。网络设计如图7所示。

然而，这些固定编码器无法有效利用点云中包含的所有信息。学习到的特征逐渐成为一种趋势。VoxelNet [237] 堆叠体素特征编码（VFE）层，对体素内的点间交互进行编码，并生成稀疏四维逐体素特征。随后，VoxelNet 使用三维卷积中间层聚合并重塑该特征，再将其输入二维检测架构。为避免对硬件不友好的三维卷积，PointPillars [238] 和 EfficientPillarNet [239]中的基于柱体的编码器在点云的柱体上学习特征。这些特征可以散射回原始柱体位置，以生成二维伪图像。PillarNet [240] 进一步发展了柱体表示：在颈部模块中融合稠密化的柱体语义特征与空间特征，用于

最终检测，并采用方向解耦的 IoU 回归损失。PillarNet [240] 的编码器如图 6 所示。

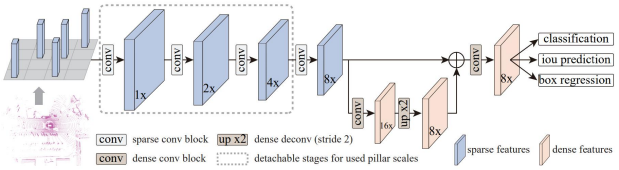

图 6. PillarNet [240]：PillarNet 中提出的基于柱体的编码器不仅实现了最先进的性能，还具备超快速的计算速度。

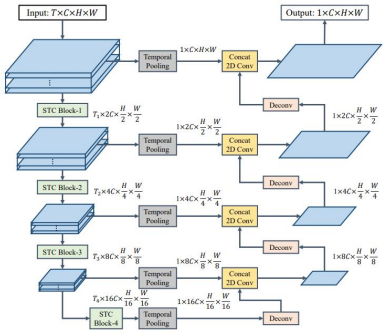

图 7. STPN [11]：仅使用二维卷积神经网络模块提取多帧体素特征的时空编码器，用于 [11], [65], [241], [242]

2）基于相机的网格映射：相机能够提供包含丰富语义线索的信息，但几何线索存在损失。早期研究人员在透视视图（PV）上开展三维视觉任务，而透视视图缺乏足够的几何引导。近年来，基于视觉BEV的方法取得了巨大成功，因为BEV为来自多视角相机的像素特征提供了结构化的坐标编码。从透视视图（PV）到鸟瞰视图（BEV）的视图投影是以视觉为中心的网格感知中的一个主要问题。

基于几何的视图变换基于几何的方法利用PV与BEV之间天然的几何投影关系进行变换。这些方法还可进一步分为两类：基于单应性的方法和基于深度的方法。由于采用了严格的平坦地面假设，基于单应性的方法能够高效地学习自动驾驶中的路面布局。基于深度的方法则更常用于表示道路上的障碍物，并实现最先进的性能。

基于单应性的方法。 逆透视映射（IPM） [243] 是最早且最重要的研究，它使用单应性矩阵来建模变换，并假设对应点位于同一水平面上。单应性矩阵可以根据摄像头的内参和外参通过数学方法推导得出。除了图像投影，Cam2BEV [244], 3D-LaneNet[245] 还使用 IPM 转换特征图，以用于下游感知任务。Gu 等人 [246] 提出了用于网络训练的单应性损失。为减少 IPM 导致的地平面以上区域的畸变，OGM[247], SHOT [248] 使用语义信息和 MonoLayout [249] 利用生成对抗网络（GAN） [250]来改进 BEV特征。

基于深度的方法。 这些方法高度依赖体素，这是一种与端到端设计兼容的显式三维表示。如图 8 所示，基于深度的方法利用每个网格的预测深度分布，将二维特征提升到三维体素空间，然后生成鸟瞰视图二维网格表示。该提升过程如下：

$$\mathcal{F}_{3D}(x,y,z) = [\mathcal{F}^*_{2D}(u,v) \otimes \mathcal{D}^*(u,v)]_{x,y,z}, \qquad (11)$$

其中，(u,v) 是三维空间中对应于 (x,y,z) 的像素坐标，$\mathcal{F}$ 是特征，$\mathcal{D}$ 是预测深度分布，* 表示一幅或多幅图像，$\otimes$ 表示外积。

深度通常被表示为一种分布。OFT [251],M2BEV [95] 假设深度分布均匀，这意味着沿同一射线的所有 3D 特征都相同。其他一些基于深度的方法 [91], [92], [252]–[255] 则使用不同的可学习网络结构和参数来建模深度分布。

基于深度的方法通常使用下游任务的标签进行监督，但逐像素深度监督也能提高深度估计的准确度。CaDDN [252], DfM[253],BEVDepth [91] 和 STS [254] 可以选择在训练阶段使用来自激光雷达点云的深度监督。

一些结构能够在环视摄像头设置中捕获立体视觉线索。MV-FCOS3D++ [255] 利用预训练二维特征提取器提取少量立体视觉线索。BEVDepth提出了一种深度校正模块，用于解决自车运动导致的图像特征图与激光雷达生成的真实标注之间的空间错位。结合BEVDepth中的单目深度模块，BEVStereo[254] 利用时序动态立体相机之间的几何对应关系来促进深度学习。

基于深度的方法在视觉中心感知中仍占主导地位。近期工作，如FIERY [29], 、M2BEV [95]、StretchBEV [256] 和BEVerse [112], 均基于由基于深度的LSS [92]生成的鸟瞰图表示。

鱼眼鸟瞰图算法。与针孔相机相比，鱼眼相机具有更宽的视场和更严重的畸变。大多数鱼眼检测算法都基于透视视图。Wu等人 [257] 提出了一个环视鱼眼数据集FPD和一个鸟瞰图感知多任务模型FPNet。该模型包含一个共享特征提取器、二维和三维检测头，以及一个逐像素深度头。

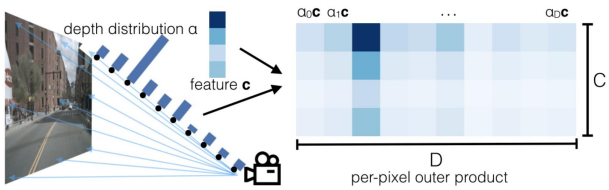

图8. LSS [92]：将图像反投影至BEV 空间时进行概率深度估计，应用于 [55], [90], [104], [112], [258]

基于网络的视图变换

基于网络的变换方法利用深度神经网络作为映射函数来建模视图变换，并隐式使用相机几何。近期方法还可进一步分为两类：基于多层感知机的和基于Transformer的。前者以正向方式进行变换，后者则构建BEV查询，并通过交叉注意力机制在透视视图地图上搜索对应特征。

基于多层感知机的方法。多层感知机用于建模从透视视图到鸟瞰视图的变换，且不需要相机参数。因此，这类方法避免了标定相机系统带来的归纳偏置。VED [259]采用变分编码器-解码器多层感知机网络，从前视单目图像实时预测鸟瞰视图语义占用栅格地图。为利用不同视角的信息实现环视感知，VPN [260] 设计了一种包含两层多层感知机的视图Transformer模块（VTM），将各透视视图特征图融合为鸟瞰图特征图。FishingNet [261]通过改进VTM，将摄像头特征转换为鸟瞰视图地图，再与雷达和激光雷达数据融合，以实现多模态感知和预测。为从单目图像中预测更准确的鸟瞰视图地图，PON [262]和STA-ST [263] 引入特征金字塔网络（FPN） [264] ，以提供具有丰富上下文信息的高分辨率特征。PON中的稠密Transformer层通过多层感知机沿高度轴压缩透视视图特征，并沿深度轴扩展这些特征，随后重新采样以获得笛卡尔坐标系中的BEV特征。HDMapNet [103] 提出了一种新颖的视图变换模块，由基于多层感知机的神经特征变换和几何投影组成。该模块能够根据环视相机拍摄的图像生成矢量化局部语义地图。由于将鸟瞰视图特征投影回透视视图地图可形成自监督信号，PYVA [265] 引入了跨视图变换模块，通过基于多层感知机的循环保留与视图变换相关的特征。为兼具基于单应性的方法和基于多层感知机的方法的优势，HFT [266] 采用两个分支，分别利用几何先验和全局上下文。

基于多层感知机的方法倾向于分别转换多视角图像，并忽略重叠区域所带来的几何潜力。这些缺点使其应用不如后续的基于Transformer的方法广泛。

基于Transformer的方法。Vaswani等 [267]提出了Transformer，这是一种促进自然语言处理（NLP）发展的基于注意力的模型。视觉Transformer凭借其捕捉长距离依赖关系和全局感受野的能力，在各种视觉任务中展现出卓越性能。与基于多层感知机的方法相比，基于Transformer的方法在视图变换中的透视视图到鸟瞰图关系建模能力更为出色。基于查询的元素提取始于二维目标检测器DETR[268]。

为满足以栅格为中心的感知任务的要求，Transformer 解码器中的查询粒度应更细，并且每个查询都根据其在三维空间或 BEV 空间中的空间位置进行位置编码。特斯拉 [269]首先通过稠密鸟瞰图查询与多视图图像特征之间的交叉注意力来进行视图变换。CVT [270] 指出，交叉注意力Transformer的几何推理表现良好，且推理速度快于几何模型。因此，它采用跨视角Transformer进行视图变换，并使用根据标定相机内外参生成的位置嵌入。除了前文提到的基于MLP的循环视角投影外，PYVA 的 [265] 变换模块还包含一个跨视角Transformer，用于关联视图投影前后的特征。

为避免Transformer注意力模块中的高计算复杂度，一种方法是使用稀疏注意力。可变形注意力 [271] 仅关注相关区域并捕获信息丰富的特征，从而减少键的数量，以避免过多的注意力计算。许多工作的视图变换模块都采用这种稀疏注意力，例如图9所示的BEVSegFormer [272] 、PersFormer [273]和BEVFormer [53]。请注意，BEVSegFormer中的查询直接预测图像上的参考点，而PersFormer和BEVFormer则依赖相机参数来计算参考点。除可变形注意力外，还提出了多种稀疏注意力。GKT [89] 在投影参考点周围展开核区域，并使其与稠密鸟瞰图查询交互，以生成鸟瞰图表示。ViT-BEVSeg [274] 使用更简单、朴素的视觉Transformer进行基于图块嵌入的视图变换。

另一种方法是利用三维几何约束简化交叉注意力计算。Image2Map [275]利用相机几何，规定图像中的垂直扫描线与BEV地图中的极线射线一一对应，从而构建平面间注意力机制。视图变换被视为一组序列到序列转换。类似地，PolarFormer [276] 也采用这种对应关系，将全局注意力简化为列注意力，并使用极坐标对齐模块聚合来自多个摄像头视角的射线，以生成结构化的极坐标BEV特征图。

PETRv2 [277]由PETR [278] 发展而来，后者包含用于以对象为中心任务的稀疏对象查询。为构建统一的三维感知框架，PETRv2添加了一组BEV分割查询，以支持高质量的BEV分割这一重要的网格中心型任务。

几何线索和时间线索对基于Transformer的方法都很重要。作为几何线索，标定相机参数被许多基于Transformer的方法用于降低内存和计算复杂度，或提高数据效率。如上所述， [270], [277] 根据相机参数生成三维位置嵌入，帮助Transformer学习视图变换。 [53], [89], [273]依赖相机参数计算用于特征采样的参考点。

[275], [276] 利用相机几何规定图像中的每条垂直扫描线与BEV特征之间的关系。然而，使用这类几何线索可能会引入标定相机系统中包含的归纳偏置。 [53], [276] 中使用的时间线索可以提升方法的性能。

极坐标。 常见做法是采用均匀分布的矩形网格笛卡尔坐标。然而，

PolarBEV [279] 认为，相比远处的网格，应更加关注周围的网格，因此采用分布不均匀的极坐标栅格化可以降低计算开销。PolarDETR[280] 基于极坐标参数化对多视角 BEV 检测变换器进行建模。PolarFormer [276] 在极坐标下提取稠密的 BEV 特征，其性能超过了 BEVFormer [53]中的笛卡尔坐标方法。

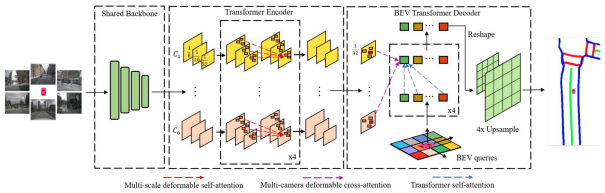

图 9. BEVSegFormer [272]：BEVSegFormer 中提出的基于Transformer的 BEV 分割模型。

3) 栅格深度融合：多传感器、多模态融合一直是汽车感知领域长期存在的难题。融合框架通常分为早期融合、深度融合和后期融合。其中，深度融合在端到端框架中表现最佳。以网格为中心的表征为多个传感器和智能体之间的深度融合提供了统一的特征嵌入空间。

多传感器融合不同传感器本质上具有互补性：摄像头几何信息较弱但语义信息丰富，而激光雷达语义信息较弱但几何信息丰富。雷达的几何和语义信息较为稀疏，但能够适应不同的天气条件。深度融合可跨模态融合潜在特征，弥补各类传感器的局限性。

激光雷达-摄像头融合。一些方法在更高的三维层级执行融合操作，并支持在三维空间中进行特征交互。UVTR [281] 根据预测深度分数从图像中采样特征，并根据点云的准确位置将其特征与体素关联。因此，可以引入用于体素空间跨模态交互的体素编码器。AutoAlign [282] 设计了一种跨注意力特征对齐模块（CAFA），使点云的体素化特征能够感知整幅图像并执行特征聚合。AutoAlign [282],通过网络学习对齐；与之不同，AutoAlignV2[283] 包含一个跨域 DeformCAFA，并利用相机投影矩阵获取图像特征图中的参考点。FUTR3D [284] 和 TransFusion[285] 基于注意力机制和查询融合特征。FUTR3D采用基于查询的模态无关特征采样器（MAFS），根据三维参考点提取多模态特征。TransFusion依靠激光雷达BEV特征和图像引导生成目标查询，并将这些查询与图像特征融合。一种简单且稳健的方法是统一在BEV特征上进行融合。图10所示的BEVFusion [90], [258], 的两种实现方式，将多模态输入的特征统一到共享 BEV 空间中。DeepInteration [286] 和 MSMDFusion [287] 在BEV空间和体素空间中设计了多模态交互，以更好地对齐来自不同传感器的空间特征。

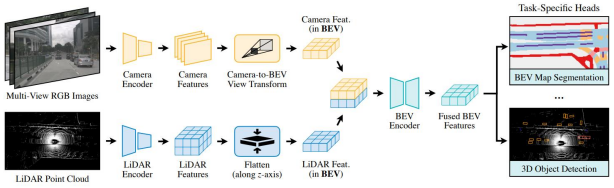

图 10. BEVFusion [90]：一种简单统一的鸟瞰视图特征级融合框架，用于融合激光雷达、摄像头和雷达点。

摄像头-雷达融合。雷达传感器最初是为高级驾驶辅助系统（ADAS）任务设计的，因此其准确度和密度不足以用于高阶自动驾驶任务。OccupancyNet [288]和NVRadarNet [289] 仅使用雷达执行实时障碍物和自由空间检测。摄像头-雷达融合是一种很有前景的低成本感知方案，可利用语义信息补充雷达几何信息。Simple-BEV [290], RCBEVDet[291], 和 CramNet [292] 研究了不同的雷达特征表达方法，包括在BEV上表达雷达特征并与视觉BEV特征融合。RCBEVDet [291] 使用PointNet++ [293]网络处理多帧聚合的雷达点云。CramNet [292] 将摄像头特征作为查询、雷达特征作为值，沿三维空间中的像素射线提取雷达特征。Simple-BEV [290]将多帧雷达点云体素化为二值占用图像，并将元数据作为附加通道。RRF [294]通过投影和采样，从每个摄像头生成一个三维特征体，然后拼接栅格化雷达鸟瞰图。最后，通过压缩垂直维度得到鸟瞰图特征图。

激光雷达-摄像头-雷达融合。激光雷达、雷达和摄像头融合是一种适用于各种天气的稳健融合策略。RaLiBEV[295] 采用交互式基于Transformer的鸟瞰视图融合方法，将激光雷达点云与雷达距离-方位角热力图进行融合。FishingNet [261] 以自顶向下语义网格作为通用输出接口，对激光雷达、雷达和摄像头进行后期融合，并对语义网格进行短期预测。

多智能体融合 近期以网格为中心的感知研究大多基于单智能体系统，在复杂交通场景中存在局限。车辆间通信（V2V）技术的发展使车辆能够共享其传感信息。CoBEVT [296]是首个能够协同生成鸟瞰视图分割地图的多智能体多摄像头感知框架。在该框架中，自车根据发送方的位姿对接收到的BEV特征进行几何变换，然后使用具有融合轴向注意力（FAX）的Transformer将其融合。动态占据栅格地图（DOGM）也展现出降低多车协同感知融合平台中不确定性的能力 [297]–[299]。

C. 相关感知任务

1）同步定位与建图：同步定位与建图（SLAM）技术对于移动机器人在未知环境中导航至关重要。SLAM 与几何建模密切相关。在激光雷达

SLAM 领域中，高阶条件随机场 [300] 提出了一种增量构建的三维滚动占据栅格地图，可用于高效表示大规模场景。SUMA++ [301] 直接采用 RangeNet++ [302]进行激光雷达分割，仅在静态环境中使用语义迭代最近点[107] ，并采用基于语义的动态滤波器进行面元地图重建。在视觉 SLAM 领域，ORB-SLAM [303] 使用点、线或平面存储地图。将空间划分为离散网格是稠密建图和语义建图算法 [304]中的常见做法。一种新趋势是将神经场与 SLAM 相结合，这有两大优势：NeRF 模型可直接处理原始像素值，无需进行特征提取；NeRF 模型能够同时区分分隐式表示和显式表示，从而实现对三维几何的全稠密优化。NICE-SLAM [305] 和 NeRF-SLAM [306] 能够生成稠密且无孔洞的地图。NeRF-SLAM 会生成一种体积 NeRF，其稠密深度损失根据深度的边缘协方差进行加权。

2) 地图元素检测：检测地图元素是制作高精地图的关键步骤。传统的全局地图构建需要通过离线全局 SLAM，获得全局一致的点云以及厘米级定位。近年来，一种新方法是基于鸟瞰视图分割和后处理技术进行局部地图学习的端到端在线学习方法，然后将不同帧中的局部地图连接起来，生成全局高精地图 [103]。整个流程如图11所示。

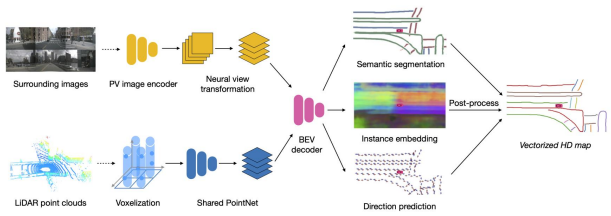

图11. HDMapNet [103]：端到端的地图元素与方向提取网络，其最终矢量化基于后处理技术。

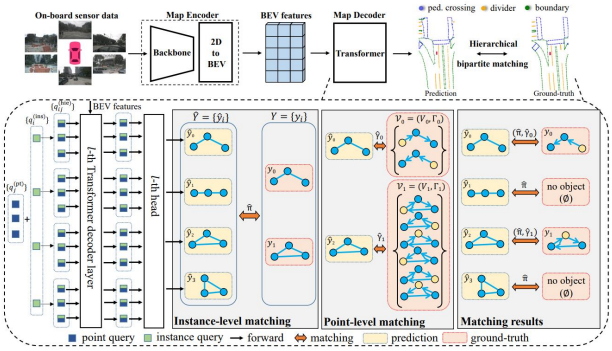

图12. MAPTR [310]：端到端的矢量地图提取方法，将矢量表示为查询，并在基于Transformer的框架中寻找相关性。

基于高精地图的应用（如定位或规划）通常需要矢量化地图元素。在 HDMapNet [103], 中，可通过对地图元素进行鸟瞰视图分割并加以处理，生成矢量化地图元素；不过，近年来端到端方法日益受到青睐。端到端流程包括对第 B 节介绍的车载激光雷达和摄像头进行特征提取，以及基于Transformer的头部模块：这些模块将矢量元素候选作为查询进行回归，并与鸟瞰图特征图中的值交互。STSU [102] 利用 Polyline-RNN [307] 从结构化交通场景中提取道路拓扑；该网络通过提取初始点估计来构建中心线曲线。VectorMapNet[308] 直接预测一组稀疏的折线基元，以表示高精地图的几何结构。InstaGraM [309] 提出了一种结合卷积神经网络和图神经网络（GNN）的混合架构，可从BEV特征中提取顶点位置和隐式边映射。该图神经网络用于将高精地图元素矢量化并连接起来。如图12所示，MAPTR[310]提出了一种层次化查询嵌入方案，通过实例级和点级二分匹配来进行地图元素学习。

3）多任务模型：许多研究表明，在多任务模型中同时预测几何、语义和时序任务，可以提高各个模型的准确度。最新进展使得一个基础框架能够处理以网格为中心的任务之外的更多感知任务。基于BEV网格的统一框架对于汽车感知系统而言十分高效。本节将介绍一些常用的多任务学习框架。

联合鸟瞰视图分割与预测在BEV网格中准确识别运动物体，是鸟瞰视图运动预测的重要前提条件。因此，已有实践表明，准确的语义识别有助于运动和速度估计。常见做法包括空间-时间特征提取骨干网络以及任务专用头：用于分类网格所属类别的分割头、用于分类网格处于静止状态还是动态状态的状态头、用于预测每个网格到实例中心偏移量的实例头，以及用于预测短期运动位移的运动头。以视觉为中心的BEV模型通常会联合优化实例的类别、位置和覆盖范围，FIERY[29]引入不确定性损失 [311] ，以平衡分割、中心度、偏移和流损失的权重。

与基于激光雷达和摄像头的BEV分割及运动预测的比较。一个明显的区别是，激光雷达模型只能估计激光扫描可触及的网格。换言之，基于激光雷达的方法无法补全未观测的网格区域，也无法补全动态物体中未观测的部分。相反，基于摄像头的方法可以采用LSS [92] 中的概率深度等技术，推断观测背后某些被遮挡的几何结构。观测之外的幻觉 [312] 就是一种从图像推断车辆三维形状的实践。另一个区别在于对开放世界未知障碍物的泛化能力。MotionNet[11]指出，尽管MotionNet是在封闭集标签上训练的，但它能够预测未知标签的运动，这些标签均被归入“其他”类别。然而，基于摄像头的方法严格依赖于对车辆、行人等定义明确的语义类别进行分类。摄像头对开放世界语义的适应能力仍是一个未解的问题。

联合三维目标检测与鸟瞰视图分割联合三维目标检测与鸟瞰视图分割是一种常见的组合，可在统一框架中处理动态物体的感知以及

静态道路布局。这也是 SSLAD2022 工作坊挑战赛[7]设立的赛道之一。给定共享的鸟瞰视图特征表示，目标检测常用的预测头包括 CenterPoint[313]中提出的中心点头，以及 Deformable DETR [314],中提出的 DETR 头；分割常用的头包括简单轻量的卷积头（例如）以及 BEVFormer [53], 中的 SegFormer [315] 或 Panoptic SegFormer [316]，也可以轻松扩展到更复杂的分割技术。BEVFormer 的流程如图 13 所示。旷视提出了 SSLAD2022 多任务挑战赛中的第一名方案 [317] 。他们提出了一种多模态多任务BEV模型作为基础模型。该模型在 ONCE数据集上预训练，并在 AutoScenes数据集上微调，采用了半监督标签校正和模块级指数移动平均（EMA）等技术。

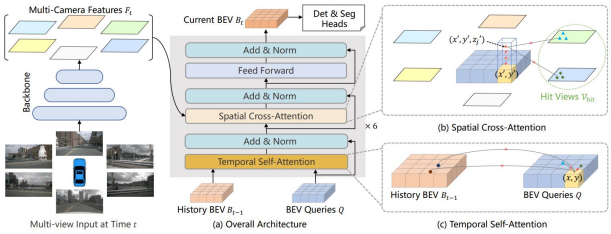

图 13. BEVFormer [105]: 基于 BEV特征的典型多任务任务。BEVFormer 使用可变形 DETR 检测头进行检测，并使用全景 SegFormer 检测头进行车道线分割。

面向更多任务的多任务处理 近期研究将更多重要的感知任务纳入基于BEV的多任务框架。BEVerse [112] 展示了BEV特征的多面性，涵盖三维目标检测、道路布局分割和占用流预测。Perceive Interact Predict [318]基于与地图元素的交互进行端到端轨迹预测，这些地图元素通过共享的BEV特征在线提取。UniAD[168] 在一个以视觉为中心的端到端框架中，全面整合了目标检测、跟踪、轨迹预测、地图分割、占用与流预测以及规划。为获得更稳定的性能，UniAD采用两阶段训练：第一阶段训练跟踪和制图模块，第二阶段训练整个模型。

分析：多任务BEV模型中的对立性。统一的BEV特征表示与任务特定预测头共同构成了一种高效的框架设计，广泛应用于工业领域。然而，一个值得关注的问题是，共享骨干网络是否能够增强各项任务的性能。联合BEV分割与运动预测研究 [110] 报告了多任务学习的积极影响：更好的分割能够带来更优的运动预测。然而，大多数联合BEV检测与分割模型 [53], [317], [318]报告了这两项任务之间的对立性。一个合理的解释是，这两项任务并不相关，因为它们需要不同高度上的特征，分别位于地面表面和地面上方。共享BEV特征如何在适应各项任务需求的同时实现良好的泛化，仍是一个尚未得到充分探索的问题。

[7]https://sslad2022.github.io/pages/challenge.html

致谢

本研究部分获得了中国国家自然科学基金项目（批准号：52394264、52372414、U22A20104）的资助。本研究由清华大学—纵目科技联合研究中心赞助。

参考文献

[1]J. Mao, S. Shi, X. Wang, 和 H. Li, 《自动驾驶三维目标检测：综述与新展望》, *arXiv*预印本 *arXiv:2206.09474*, 2022年。[2] R. Qian, X. Lai, 和 X. Li, 《自动驾驶三维目标检测：综述》, 模式识别, 第130卷, 24, 2022年8月, 1号, 第–页。[3] K. Gao, Y. Gao, H. He, D. Lu, L. Xu, 和 J. Li, 《NeRF：三维视觉中的神经辐射场，全面综述》, 第–页。在线。获取地址：http://arxiv.org/abs/。 00379

[4] Y. Ma, T. Wang, X. Bai, H. Yang, Y. Hou, Y. Wang, Y. Qiao, R. Yang, D. Manocha, 和 X. Zhu, 《视觉中心化鸟瞰图感知：综述》, *arXiv*预印本 *arXiv:2208.02797*, 2022年。[5] H. Li, C. Sima, J. Dai, W. Wang, L. Lu, H. Wang, E. Xie, Z. Li, H. Deng, H. Tian, 等, 《深入探究鸟瞰图感知中的难题：综述、评估与实践指南》, *arXiv*预印本 *arXiv:2209.05324*, 2022年。[6] Y. Zhang, J. Zhang, Z. Wang, J. Xu, 和 D. Huang, 《自动驾驶中的基于视觉的三维占用预测：综述与展望》, 2024年。[7] H. Xu, J. Chen, S. Meng, Y. Wang, 和 L.-P. Chau, 《自动驾驶占据感知综述：信息融合视角》, 2024年。[8] A.-Q. Cao 和 R. de Charette, 《MonoScene：单目三维语义场景补全》, 载于*IEEE/CVF*计算机视觉与模式识别会议论文集, 2022年, 第3991–4001页。[9] M. Schreiber, V. Belagiannis, C. Gl¨aser, 和 K. Dietmayer, 《用于端到端动态占用栅格地图构建的多任务循环神经网络》, 2022年。 [在线]。获取地址：http://arxiv.org/abs/2202.04461[10] J. Ma, X. Chen, J. Huang, J. Xu, Z. Luo, J. Xu, W. Gu, R. Ai, 和 H. Wang, 《Cam4DOcc：面向自动驾驶应用的纯摄像头四维占用预测基准》, *arXiv*预印本 *arXiv:2311.17663*, 2023年。[11] P. Wu, S. Chen, 和 D. N. Metaxas, 《MotionNet：基于鸟瞰视角地图的自动驾驶联合感知与运动预测》, *IEEE*计算机学会计算机视觉与模式识别会议论文集, 第11 382–11 392页, 2020年。[12] K. Joseph, S. Khan, F. S. Khan, 和 V. N. Balasubramanian, 《迈向开放世界目标检测》, 载于*IEEE/CVF*计算机视觉与模式识别会议论文集, 2021年, 第5830–5840页。[13] A. Gupta, S. Narayan, K. Joseph, S. Khan, F. S. Khan, 和 M. Shah, 《Ow-DETR：开放世界检测变换器》, 载于*IEEE/CVF*计算机视觉与模式识别会议论文集, 2022年, 第9235–9244页。[14] O. Zohar, K.-C. Wang, 和 S. Yeung, 《PROB：用于开放世界目标检测的概率目标性》, 第31–33页, 2022年。

[15] [ ] 2212 01424 在线。获取地址：http://arxiv.org/abs/。J. Cen, P. Yun, J. Cai, M. Y. Wang, 和 M. Liu, 《开放集三维目标检测》, 载于*2021*年国际三维视觉会议（*3DV*）。IEEE, 2021年, 第869–878页。[16] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, 等, 《Segment Anything》, 载于*IEEE/CVF*国际计算机视觉会议论文集, 2023年, 第4015–4026页。[17] A. Geiger, P. Lenz, 和 R. Urtasun, 《我们准备好迎接自动驾驶了吗？KITTI视觉基准套件》, 载于*2012*年*IEEE*计算机视觉与模式识别会议, 2012年, 第3354–3361页。[18] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Kr-ishnan, Y. Pan, G. Baldan, 和 O. Beijbom, 《nuScenes：面向自动驾驶的多模态数据集》, 载于*2020*年*IEEE/CVF*计算机视觉与模式识别会议（*CVPR*）, 2020年, 第11 618–11 628页。

[19] M.-F. Chang, J. Lambert, P. Sangkloy, J. Singh, S. Bak, A. Hartnett, D. Wang, P. Carr, S. Lucey, D. Ramanan, 和 J. Hays, 《Argoverse：利用丰富地图进行三维跟踪与预测》，载于2019年IEEE/CVF计算机视觉与模式识别会议（CVPR），2019年，第8740–8749页。[20] R. Kesten, M. Usman, J. Houston, T. Pandya, K. Nadhamuni, A. Fer-reira, M. Yuan, B. Low, A. Jain, P. Ondruska, S. Omari, S. Shah, A. Kulkarni, A. Kazakova, C. Tao, L. Platinsky, W. Jiang, 和 V. Shet, 《Level 5 感知数据集 2020》，https://level-5.global/level5/data/，2019年。[21] J. Behley, M. Garbade, A. Milioto, J. Quenzel, S. Behnke, C. Stachniss, 和 J. Gall, 《SemanticKITTI：用于激光雷达序列语义场景理解的数据集》，载于2019年IEEE/CVF国际计算机视觉会议（ICCV），2019年，第9296–9306页。[22] Y. Liao, J. Xie, 和 A. Geiger, 《KITTI-360：用于二维和三维城市场景理解的新型数据集与基准》，IEEE模式分析与机器智能汇刊，第1 1页，2022年–。[23] P. Sun, H. Kretzschmar, X. Dotiwalla, A. Chouard, V. Patnaik, P. Tsui, J. Guo, Y. Zhou, Y. Chai, B. Caine, V. Vasudevan, W. Han, J. Ngiam, H. Zhao, A. Timofeev, S. Ettinger, M. Krivokon, A. Gao, A. Joshi, Y. Zhang, J. Shlens, Z. Chen, 和 D. Anguelov, 《自动驾驶感知的可扩展性：Waymo 开放数据集》，载于2020年IEEE/CVF计算机视觉与模式识别会议（CVPR），2020年，第2443–2451页。[24] B. Wilson, W. Qi, T. Agarwal, J. Lambert, J. Singh, S. Khandelwal, B. Pan, R. Kumar, A. Hartnett, J. K. Pontes, D. Ramanan, 和 J. Hays, 《Argoverse 2：面向自动驾驶感知与预测的下一代数据集》，载于NeurIPS数据集与基准，2021年。[25] J. Mao, M. Niu, C. Jiang, X. Liang, Y. Li, C. Ye, W. Zhang, Z. Li, J. Yu, C. Xu，等，《自动驾驶的百万场景：ONCE数据集》，2021年。[26] Y. Li, S. Li, X. Liu, M. Gong, K. Li, N. Chen, Z. Wang, Z. Li, T. Jiang, F. Yu，等，《SSCBench：面向自动驾驶的大规模三维语义场景补全基准》，arXiv预印本 arXiv:2306 09001，2023年。[27] X. Tian, T. Jiang, L. Yun, Y. Mao, H. Yang, Y. Wang, Y. Wang, 和 H. Zhao, 《Occ3D：面向自动驾驶的大规模三维占用预测基准》，神经信息处理系统进展，第36卷，2024年。[28] X. Wang, Z. Zhu, W. Xu, Y. Zhang, Y. Wei, X. Chi, Y. Ye, D. Du, J. Lu, 和 X. Wang, 《OpenOccupancy：用于周围环境语义占用感知的大规模基准》，载于IEEE/CVF国际计算机视觉会议论文集，2023年，第17 850–17 859页。[29] A. Hu, Z. Murez, N. Mohan, S. Dudas, J. Hawke, V. Badrinarayanan, R. Cipolla, 和 A. Kendall, 《FIERY：利用环视单目相机进行鸟瞰图中的未来实例预测》，IEEE国际计算机视觉会议论文集，第15 253–15 262页，2021年。[30] W. Tong, C. Sima, T. Wang, L. Chen, S. Wu, H. Deng, Y. Gu, L. Lu, P. Luo, D. Lin，等，《场景即占用》，载于IEEE/CVF国际计算机视觉会议论文集，2023年，第8406–8415页。[31] J. Wang, X. Li, A. Sullivan, L. Abbott, 和 S. Chen, 《PointMotionNet：面向大规模激光雷达点云序列的逐点运动学习》，载于IEEE/CVF计算机视觉与模式识别会议论文集，2022年，第4419–4428页。[32]D. Kim, S. Woo, J.-Y. Lee, 和 I. S. Kweon, 《视频全景分割》，载于2020年IEEE/CVF计算机视觉与模式识别会议（CVPR），2020年，第9856–9865页。[33] G. M´attyus, W. Luo, 和 R. Urtasun, 《DeepRoadMapper：从航空影像中提取道路拓扑》，载于2017年IEEE国际计算机视觉会议（ICCV），2017年，第3458–3466页。[34] H. Liu, H. Wang, Y. Chen, Z. Yang, J. Zeng, L. Chen, 和 L. Wang, 《完全稀疏的三维全景占用预测》，arXiv预印本 arXiv:2312.17118，2023年。[35] X. Liu, M. Gong, Q. Fang, H. Xie, Y. Li, H. Zhao, 和 C. Feng, 《基于激光雷达的4D占用补全与预测》，arXiv预印本 arXiv:2310.11239，2023年。[36] ——, 《基于激光雷达的4D占用补全与预测》，arXiv预印本 arXiv:2310.11239，2023年。[37] Z. Yang, L. Chen, Y. Sun, 和 H. Li, 《视觉点云预测助力可扩展自动驾驶》，arXiv预印本 arXiv:2312 17655，2023年。

[38] R. Cheng、C. Agia、Y. Ren、X. Li 和 L. Bingbing, 《S3CNet：面向激光雷达点云的稀疏语义场景补全网络》，arXiv预印本 arXiv:2012.09242，2020年。[39] L. Rold˜ao、R. de Charette 和 A. Verroust-Blondet, 《LMSCNet：轻量级多尺度3D语义补全》，载于《2020年国际三维视觉会议（3DV）》，2020年，第111–119页。[40] X. Yan、J. Gao、J. Li、R. Zhang、Z. Li、R. Huang 和 S. Cui, 《通过从场景补全中学习上下文形状先验实现稀疏单帧激光雷达点云分割》，载于《AAAI人工智能会议论文集》，第35卷第4期，2021年，第3101–3109页。[41] C. B. Rist、D. Emmerichs、M. Enzweiler 和 D. M. Gavrila, 《利用激光雷达数据上的局部深度隐式函数进行语义场景补全》，《IEEE模式分析与机器智能汇刊》，第44卷第10期，2022年，第7205–7218页。[42] J. Pan、Z. Wang 和 L. Wang, 《Co-Occ：将显式特征融合与体渲染正则化相结合的多模态三维语义占用预测》，《IEEE机器人与自动化快报》，2024年。[43] Y. Huang、W. Zheng、Y. Zhang、J. Zhou 和 J. Lu, 《面向基于视觉的三维语义占用预测的三视角视图》，2023年。[ ] 2302 07817 在线。可访问：[44] https://arxiv.org/abs/ 。Y. Zhang、Z. Zhu 和 D. Du, 《OccFormer：面向基于视觉的三维语义占用预测的双路径Transformer》，载于《IEEE/CVF国际计算机视觉会议论文集》，2023年，第9433–9443页。[45] Y. Wei、L. Zhao、W. Zheng、Z. Zhu、J. Zhou 和 J. Lu, 《SurroundOcc：面向自动驾驶的多摄像头三维占用预测》，载于《IEEE/CVF国际计算机视觉会议论文集》，2023年，第21729–21740页。[46] J. Yao、C. Li、K. Sun、Y. Cai、H. Li、W. Ouyang 和 H. Li, 《NDC-Scene：在归一化设备坐标空间中提升单目三维语义场景补全》，载于《2023年IEEE/CVF国际计算机视觉会议（ICCV）》。IEEE计算机学会，2023年，第9421–9431页。[47] M. Pan、J. Liu、R. Zhang、P. Huang、X. Li、L. Liu 和 S. Zhang, 《RenderOcc：具有二维渲染监督的以视觉为中心的三维占用预测》，arXiv预印本 arXiv:2309.09502，2023年。[48] H. Jiang、T. Cheng、N. Gao、H. Zhang、W. Liu 和 X. Wang, 《利用上下文实例查询对三维语义场景补全进行Symphonize》，arXiv预印本 arXiv:2306.15670，2023年。[49] S. Wang、J. Yu、W. Li、W. Liu、X. Liu、J. Chen 和 J. Zhu, 《并非所有体素都相同：采用自蒸馏的难度感知语义场景补全》，arXiv预印本 arXiv:2404.11958，2024年。[50] B. Li、Y. Sun、Z. Liang、D. Du、Z. Zhang、X. Wang、Y. Wang、X. Jin 和 W. Zeng, 《通过可靠的相互交互连接双目几何与鸟瞰图表示，实现语义场景补全》，2024年。[51] Y. Li、Z. Yu、C. Choy、C. Xiao、J. M. Alvarez、S. Fidler、C. Feng 和 A. Anandkumar, 《VoxFormer：面向基于摄像头的三维语义场景补全的稀疏体素Transformer》，2023年。 [在线]。可访问：https://arxiv.org/abs/ 2302 12251 。[52] Y. Zheng、X. Li、P. Li、Y. Zheng、B. Jin、C. Zhong、X. Long、H. Zhao 和 Q. Zhang, 《MonoOcc：深入研究单目语义占用预测》，arXiv预印本 arXiv:2403.08766，2024年。[53] Z. Li、W. Wang、H. Li、E. Xie、C. Sima、T. Lu、Y. Qiao 和 J. Dai, 《BEVFormer：通过时空Transformer从多摄像头图像中学习鸟瞰图表示》，载于《计算机视觉—ECCV 2022：第17届欧洲会议》，以色列特拉维夫，2022年10月23–27日，论文集，第IX部分。Springer，2022年，第1–18页。[54] Y. Wang、Y. Chen、X. Liao、L. Fan 和 Z. Zhang, 《PanoOcc：面向基于摄像头的三维全景分割的统一占用表示》，arXiv预印本 arXiv:2306.10013，2023年。[55] J. Huang、G. Huang、Z. Zhu 和 D. Du, 《BEVDet：鸟瞰视图中的高性能多摄像头三维目标检测》，arXiv预印本 arXiv:2112.11790，2021年。[56] H. Zhang、X. Yan、D. Bai、J. Gao、P. Wang、B. Liu、S. Cui 和 Z. Li, 《通过渲染辅助蒸馏学习跨模态占用知识的RadOcc》，载于《AAAI人工智能会议论文集》，第38卷第7期，2024年，第7060–7068页。[57] M. Pan、L. Liu、J. Liu、P. Huang、L. Wang、S. Zhang、S. Xu、Z. Lai 和 K. Yang, 《UniOcc：通过几何与语义渲染统一以视觉为中心的三维占用预测》，arXiv预印本 arXiv:2306.09117，2023年。[58] T. V. J.-H. K. Myeongjin 和 K. S. J. S.-G. Jeong, 《MiLO：通过定位歧义抑制进行多任务学习——CVPR 2023占用挑战赛报告》，arXiv预印本 arXiv:2306.11414，2023年。

[59] Z. Li、Z. Yu、D. Austin、M. Fang、S. Lan、J. Kautz 和 J. M. Alvarez，《FB-OCC：基于前向-后向视图变换的三维占用预测》，arXiv 预印本 arXiv:2307.01492，2023年。[60] X. Chen、K.-Y. Lin、C. Qian、G. Zeng 和 H. Li，《通过半监督结构先验实现3D草图感知语义场景补全》，载于《IEEE/CVF计算机视觉与模式识别会议论文集》，2020年，第4193–4202页。[61] J. Li、K. Han、P. Wang、Y. Liu 和 X. Yuan，《用于三维语义场景补全的各向异性卷积网络》，载于《IEEE/CVF计算机视觉与模式识别会议论文集》，2020年，第3351–3359页。[62] S. Zuo、W. Zheng、Y. Huang、J. Zhou 和 J. Lu，《PointOcc：面向基于点的3D语义占用预测的圆柱三视角视图》，arXiv预印本 arXiv:2308.16896，2023年。[63] J. Zhang 和 Y. Ding，《OccFusion：无需深度估计的多传感器融合三维占用预测》，arXiv预印本 arXiv: . 3 05329，2024年 2403。[64] G. Wang、Z. Wang、P. Tang、J. Zheng、X. Ren、B. Feng 和 C. Ma，《OccGen：面向自动驾驶的生成式多模态3D占用预测》，arXiv预印本 arXiv:2404.15014，2024年。[65] J. Wilson、J. Song、Y. Fu、A. Zhang、A. Capodieci、P. Jayakumar、K. Barton 和 M. Ghaffari，《MotionSC：用于动态环境中实时语义建图的数据集与网络》，《IEEE机器人与自动化快报》，第7卷第3期，2022年，第8439–8446页。[66] B. Mildenhall、P. P. Srinivasan、M. Tancik、J. T. Barron、R. Ramamoor[67] thi 和 R. Ng，《NeRF：将场景表示为用于视图合成的神经辐射场》，《计算机科学讲义》（包括《人工智能讲义》和《生物信息学讲义》子系列），第12346卷 LNCS，2020年，第405–421页。¨B. Kerbl、G. Kopanas、T. Leimk[68] uhler 和 G. Drettakis，《用于实时辐射场渲染的三维高斯泼溅》，《ACM图形学汇刊》，第42卷第4期，2023年，第1–14页。[68] J. Yang、B. Ivanovic、O. Litany、X. Weng、S. W. Kim、B. Li、T. Che、D. Xu、S. Fidler、M. Pavone 等，《Emernerf：通过自监督实现涌现式时空场景分解》，arXiv预印本 arXiv:2311.02077，2023年。[69] B. Graham、M. Engelcke 和 L. Van Der Maaten，《利用子流形稀疏卷积网络进行3D语义分割》，载于《IEEE计算机视觉与模式识别会议论文集》，2018年，第9224–9232页。[70] C. Choy、J. Gwak 和 S. Savarese，《4D时空卷积网络：Minkowski卷积神经网络》，载于《IEEE/CVF计算机视觉与模式识别会议论文集》，2019年，第3075–3084页。[71] A. Dosovitskiy、G. Ros、F. Codevilla、A. Lopez 和 V. Koltun，《CARLA：开放式城市驾驶模拟器》，载于《机器人学习会议》。PMLR，2017年，第1–16页。[72] F. Ding、X. Wen、Y. Zhu、Y. Li 和 C. X. Lu，《RadarOcc：利用四维成像雷达进行鲁棒3D占用预测》，2024年。[73] Z. Li、Z. Yu、W. Wang、A. Anandkumar、T. Lu 和 J. M. Alvarez，《FB-BEV：来自前向-后向视图变换的鸟瞰图表示》，载于《IEEE/CVF国际计算机视觉会议论文集》，2023年，第6919–6928页。[74] W. Wang、J. Dai、Z. Chen、Z. Huang、Z. Li、X. Zhu、X. Hu、T. Lu、L. Lu、H. Li 等，《Internimage：利用可变形卷积探索大规模视觉基础模型》，载于《IEEE/CVF计算机视觉与模式识别会议论文集》，2023年，第14408–14419页。[75] R. Miao、W. Liu、M. Chen、Z. Gong、W. Xu、C. Hu 和 S. Zhou，《OccDepth：用于三维语义场景补全的深度感知方法》，

2023 [ ] 2302 13540。在线。可访[76] 问：https://arxiv.org/abs/ 。A.-Q. Cao、A. Dai 和 R. de Charette，《PaSco：具有不确定性感知的城市3D全景场景补全》，2023年。[77] B. Cheng、I. Misra、A. G. Schwing、A. Kirillov 和 R. Girdhar，《用于通用图像分割的掩码注意力掩码变换器》，载于《IEEE/CVF计算机视觉与模式识别会议论文集》，2022年，第1290–1299页。[78] Z. Ming、J. S. Berrio、M. Shan 和 S. Worrall，《OccFusion：用于三维占用预测的简洁有效的多传感器融合框架》，arXiv预印本 arXiv:2403.01644，2024年。[79] Y. Yan、B. Liu、J. Ai、Q. Li、R. Wan 和 J. Pu，《PointSSC：用于语义场景补全的车路协同点云基准》，arXiv预印本 arXiv:2309.12708，2023年。[80] R. Song、C. Liang、H. Cao、Z. Yan、W. Zimmer、M. Gross、A. Festag 和 A. Knoll，《网联自动驾驶车辆中的混合特征融合协同语义占用预测》，arXiv预印本 arXiv:2402.07635

，2024年。

[81] V. Sitzmann、J. Martel、A. Bergman、D. Lindell 和 G. Wetzstein，《采用周期激活函数的隐式神经表示》，神经信息处理系统进展，第33卷，第7462–7473页，2020年。[82] P. Li、R. Zhao、Y. Shi、H. Zhao、J. Yuan、G. Zhou 和 Y.-Q. Zhang，《LODE：基于稀疏激光雷达的局部条件化程函隐式场景补全》，载于2023年IEEE机器人与自动化国际会议（ICRA）。IEEE，2023年，第8269–8276页。[83] L. Liu、B. Wang、H. Xie、D. Liu、L. Liu、Z. Tian、K. Yang 和 B. Wang，《SurroundSDF：基于有符号距离场的隐式三维场景理解》，arXiv预印本，arXiv:2403.14366，2024年。[84] A. Elluswamy。占用网络、自动驾驶辅助系统、特斯拉。YouTube。 [在线]。网址：https://www.youtube.com/watch?v=jPCV4GKX9Dw& list=PLvXze1V52Yy3YfsHjqkKTijvYDPOvy9L2[85] CVPR 2022自动驾驶研讨会（WAD），IEEE/CVF计算机视觉与模式识别会议研讨会，CVPR 2022研讨会，美国路易斯安那州新奥尔良，2022年6月19日至20日。IEEE，2022年。[在线]。网址：https://cvpr2022.wad.vision/ [86] 特斯拉。特斯拉AI日2022。YouTube。 [在线]。网址：https: //www.youtube.com/watch?v=ODSJsviD SU [87] I. Radosavovic、R. P. Kosaraju、R. Girshick、K. He 和 P. Doll´ar，《设计网络设计空间》，载于IEEE/CVF计算机视觉与模式识别会议论文集，2020年，第10 428–10 436页。[88] M. Tan、R. Pang 和 Q. V. Le，《EfficientDet：可扩展且高效的目标检测》，载于IEEE/CVF计算机视觉与模式识别会议论文集，2020年，第10 781–10 790页。[89] S. Chen、T. Cheng、X. Wang、W. Meng、Q. Zhang 和 W. Liu，《通过几何引导的核Transformer实现高效且稳健的二维到鸟瞰图表示学习》，arXiv预印本，arXiv:2206.04584，2022年。[90] Z. Liu、H. Tang、A. Amini、X. Yang、H. Mao、D. Rus 和 S. Han，《BEVFusion：采用统一鸟瞰图表示的多任务多传感器融合》，arXiv预印本，arXiv:2205.13542，2022年。[91] Y. Li、Z. Ge、G. Yu、J. Yang、Z. Wang、Y. Shi、J. Sun 和 Z. Li，《BEVDepth：为多视角三维目标检测获取可靠深度》，arXiv预印本，arXiv:2206.10092，2022年。[92] J. Philion 和 S. Fidler，《Lift, Splat, Shoot：通过隐式反投影至三维，对来自任意相机系统的图像进行编码》，《计算机科学讲义》（含《人工智能讲义》和《生物信息学讲义》子系列），第12359卷（LNCS），第194–210页，2020年。[93] H. Zhou、Z. Ge、Z. Li 和 X. Zhang，《MatrixVT：用于三维感知的高效多相机到鸟瞰视图变换》，载于IEEE/CVF国际计算机视觉会议论文集，2023年，第8548–8557页。[94] B. Huang、Y. Li、E. Xie、F. Liang、L. Wang、M. Shen、F. Liu、T. Wang、P. Luo 和 J. Shao，《Fast-BEV：迈向实时车载鸟瞰图感知》，arXiv预印本，arXiv:2301.07870，2023年。[95] E. Xie、Z. Yu、D. Zhou、J. Philion、A. Anandkumar、S. Fidler、P. Luo 和 J. M. Alvarez，《Mˆ 2bev：采用统一鸟瞰图表示的多相机联合三维检测与分割》，arXiv预印本，arXiv:2204.05088，2022年。[96] C. Yang、T. Lin、L. Huang 和 E. J. Crowley，《WidthFormer：迈向高效的基于Transformer的鸟瞰图视图变换》，arXiv预印本，arXiv:2401.03836，2024年。[97] J. Hou、X. Li、W. Guan、G. Zhang、D. Feng、Y. Du、X. Xue 和 J. Pu，《FastOcc：通过融合二维鸟瞰图和透视视图加速三维占用预测》，arXiv预印本，arXiv:2403 02710，2024年。[98] P. Tang、Z. Wang、G. Wang、J. Zheng、X. Ren、B. Feng 和 C. Ma，《SparseOcc：重新思考用于基于视觉的语义占用预测的稀疏潜在表示》，arXiv预印本，arXiv:2404 09502，2024年。[99] Z. Yu、C. Shu、J. Deng、K. Lu、Z. Liu、J. Yu、D. Yang、H. Li 和 Y. Chen，《FlashOcc：通过通道到高度插件实现快速且节省内存的占用预测》，arXiv预印本，arXiv:2311.12058，2023年。[100] J. Lee、J. Koh、Y. Lee 和 J. W. Choi，《D-Align：基于多帧点云序列的三维目标检测双查询协同注意力网络》，载于2023年IEEE机器人与自动化国际会议（ICRA）。IEEE，2023年，第9238–9244页。[101] Y. Zhang、Z. Zhu 和 J. Hou，《面向多帧三维目标检测的时空增强Transformer》，arXiv预印本，arXiv:2307.00347，2023年。[102] Y. B. Can、A. Liniger、D. P. Paudel 和 L. Van Gool，《基于车载图像的结构化鸟瞰图交通场景理解》，IEEE国际计算机视觉会议论文集，第Iccv期

，第 15 641–15650 页，2021年。

[103] Q. Li、Y. Wang、Y. Wang 和 H. Zhao，《HDMapNet：一种在线高清地图构建与评估框架》，载于《2022年国际机器人与自动化会议（ICRA）》。IEEE，2022年，第4628–4634页。[104] J. Huang 和 G. Huang，《BEVDet4D：在多摄像头三维目标检测中利用时间线索》，arXiv 预印本 arXiv:2203.17054，2022年。[105] Z. Qin、J. Chen、C. Chen、X. Chen 和 X. Li，《UniFormer：用于鸟瞰视图时空表征的统一多视角融合Transformer》，arXiv 预印本 arXiv:2207.08536，2022年。[106] J. Park、C. Xu、S. Yang、K. Keutzer、K. M. Kitani、M. Tomizuka 和 W. Zhan，《时间会给出答案：时序多视角三维目标检测的新视角与基线》，载于《第十一届国际学习表征会议》，2022年。[107] S. Rusinkiewicz 和 M. Levoy，《ICP算法的高效变体》，载于《第三届三维数字成像与建模国际会议论文集》。IEEE，2001年，第145–152页。[108] P. Biber 和 W. Straßer，《正态分布变换：一种新的激光扫描匹配方法》，载于《2003年IEEE/RSJ国际智能机器人与系统会议论文集（IROS 2003）（目录号03CH37453）》，第3卷。IEEE，2003年，第2743–2748页。[109] C. Luo、X. Yang 和 A. Yuille，《面向自动驾驶的自监督柱体运动学习》，《IEEE计算机学会计算机视觉与模式识别会议论文集》，第3182–3191页，2021年。[110] Y. Wang、H. Pan、J. Zhu、Y.-H. Wu、X. Zhan、K. Jiang 和 D. Yang，《BE-STI：通过双向增强实现类别无关运动预测的时空集成网络》，载于《IEEE/CVF计算机视觉与模式识别会议论文集》，2022年，第17 093–17 102页。[111] S. Hu、L. Chen、P. Wu、H. Li、J. Yan 和 D. Tao，《ST-P3：通过时空特征学习实现端到端的基于视觉的自动驾驶》，载于《计算机视觉—ECCV 2022：第17届欧洲会议，特拉维夫，以色列，2022年10月23日至27日，论文集，第XXXVIII部分》。Springer，2022年，第533–549页。[112] Y. Zhang、Z. Zhu、W. Zheng、J. Huang、G. Huang、J. Zhou 和 J. Lu，《BEVerse：面向以视觉为中心的自动驾驶，在鸟瞰视图中实现统一感知与预测》，arXiv 预印本 arXiv:2205 09743，2022年。[113] S. Fang、Z. Wang、Y. Zhong、J. Ge 和 S. Chen，《TBP-Former：学习时序鸟瞰视角金字塔，以实现以视觉为中心的自动驾驶中的联合感知与预测》，载于《IEEE/CVF计算机视觉与模式识别会议论文集》，2023年，第1368–1378页。[114] X. Gui、T. Huang、H. Shao、H. Yao 和 C. Zhang，《FipTR：一种简单而有效的Transformer框架，用于自动驾驶中的未来实例预测》，arXiv 预印本 arXiv:2404.12867，2024年。[115] S. Targ、D. Almeida 和 K. Lyman，《Resnet in resnet：泛化残差架构》，arXiv 预印本 arXiv:1603.08029，2016年。[116] A. Jain、S. Casas、R. Liao、Y. Xiong、S. Feng、S. Segal 和 R. Urtasun，《用于概率行人行为预测的离散残差流》，载于《机器人学习会议》。PMLR，2020年，第407–419页。[117] M. Bansal、A. Krizhevsky 和 A. S. Ogale，《ChauffeurNet：通过模仿优秀驾驶并合成糟糕驾驶来学习驾驶》，CoRR，第abs/1812.03079卷，2018年。 [在线]。可访问：http://arxiv.org/abs/ 1812 03079 。J. Hong、[118] B. Sapp 和 J. Philbin，《道路规则：利用语义交互卷积模型预测驾驶行为》，载于《IEEE/CVF计算机视觉与模式识别会议论文集》，2019年，第8454–8462页。[119] S. Casas、A. Sadat 和 R. Urtasun，《MP3：一种用于地图构建、感知、预测和规划的统一模型》，载于《IEEE/CVF计算机视觉与模式识别会议论文集》，2021年，第14 403–14 412页。[120] Y. Hu、W. Shao、B. Jiang、J. Chen、S. Chai、Z. Yang、J. Qian、H. Zhou 和 Q. Liu，《HOPE：用于占用流预测的分层时空网络》，arXiv 预印本 arXiv:2206.10118，2022年。[121] X. Huang、X. Tian、J. Gu、Q. Sun 和 H. Zhao，《VectorFlow：结合图像与向量进行交通占用和流预测》，第2–5页，2022年。 [在线]。可访问：http://arxiv.org/abs/ 2208.04530[122] H. Liu、Z. Huang 和 C. Lv，《STrajNet：通过多模态Swin Transformer进行占用流预测》，arXiv 预印本 arXiv:2208 00394，2022年。[123] W. Zheng、W. Chen、Y. Huang、B. Zhang、Y. Duan 和 J. Lu，《OccWorld：学习面向自动驾驶的三维占用世界模型》，arXiv 预印本 arXiv:2311.16038，2023年。

[124] X. Liu、M. Gong、Q. Fang、H. Xie、Y. Li、H. Zhao 和 C. Feng，《基于激光雷达的4D占用补全与预测》，arXiv预印本 arXiv:2310.11239，2023年。[125] L. Zhang、Y. Xiong、Z. Yang、S. Casas、R. Hu 和 R. Urtasun，《Copilot4d：通过离散扩散学习用于自动驾驶的无监督世界模型》，载于《第十二届国际学习表征会议》，2024年。[126] Y. Xiong、W.-C. Ma、J. Wang 和 R. Urtasun，《UltraLiDAR：学习用于激光雷达补全与生成的紧凑表示》，arXiv预印本 arXiv:2311.01448，2023年。[127] R. Rombach、A. Blattmann、D. Lorenz、P. Esser 和 B. Ommer，《使用潜在扩散模型进行高分辨率图像合成》，载于《IEEE/CVF计算机视觉与模式识别会议论文集》，2022年，第10 684–10 695页。[128] H. Chang、H. Zhang、L. Jiang、C. Liu 和 W. T. Freeman，《MaskGIT：掩码生成式图像变换器》，载于《IEEE/CVF计算机视觉与模式识别会议论文集》，2022年，第11 315–11 325页。[129] X. Chen、S. Xie 和 K. He，《训练自监督视觉Transformer的实证研究》，《IEEE国际计算机视觉会议论文集》，第9620–9629页，2021年。[130] K. He、X. Chen、S. Xie、Y. Li、P. Doll´ar 和 R. Girshick，《掩码自编码器是可扩展的视觉学习器》，载于《IEEE/CVF计算机视觉与模式识别会议论文集》，2022年，第16 000–16 009页。[131] Z. Xie、Z. Zhang、Y. Cao、Y. Lin、J. Bao、Z. Yao、Q. Dai 和 H. Hu，《SimMIM：用于掩码图像建模的简单框架》，载于《IEEE/CVF计算机视觉与模式识别会议论文集》，2022年，第9653–9663页。[132] S. Xie、J. Gu、D. Guo、C. R. Qi、L. Guibas 和 O. Litany，《PointContrast：用于3D点云理解的无监督预训练》，《计算机科学讲义》（包括《人工智能讲义》和《生物信息学讲义》子系列），第12348卷LNCS，第574–591页，2020年。[133] Z. Zhang、R. Girdhar、A. Joulin 和 I. Misra，《在任意点云上进行3D特征的自监督预训练》，载于《IEEE/CVF国际计算机视觉会议论文集》，2021年，第10 252–10 263页。[134] Y. Pang、W. Wang、F. E. Tay、W. Liu、Y. Tian 和 L. Yuan，《用于点云自监督学习的掩码自编码器》，arXiv预印本 arXiv:2203.06604，2022年。[135] J. Yin、D. Zhou、L. Zhang、J. Fang、C.-Z. Xu、J. Shen 和 W. Wang，《ProposalContrast：用于基于激光雷达的3D目标检测的无监督预训练》，载于《计算机视觉—ECCV 2022：第17届欧洲会议》，以色列特拉维夫，2022年10月23日至27日，论文集，第XXXIX部分。Springer，2022年，第17–33页。[136] G. Hess、J. Jaxing、E. Svensson、D. Hagerman、C. Petersson 和 L. Svensson，《用于汽车点云自监督学习的掩码自编码器》，arXiv预印本 arXiv:2207.00531，2022年。[137] C. Min、D. Zhao、L. Xiao、Y. Nie 和 B. Dai，《Voxel-MAE：用于大规模点云预训练的掩码自编码器》，arXiv预印本 arXiv:2206.09900，2022年。[138] Z. Lin 和 Y. Wang，《BEV-MAE：用于户外点云预训练的鸟瞰视图掩码自编码器》，arXiv预印本 arXiv:2212 05758，2022年。。[139] G. Krispel、D. Schinagl、C. Fruhwirth-Reisinger、H. Possegger 和 H. Bischof，《MAELi—用于大规模激光雷达点云的掩码自编码器》，arXiv预印本 arXiv:2212.07207，2022年。[140] A. Boulch、C. Sautier、B. Michele、G. Puy 和 R. Marlet，《ALSO：基于占用估计的汽车激光雷达自监督》，2022年。 [ ] 2212 05867 在线。获取地址：http://arxiv.org/abs/。C. Min、L. Xiao、D. Zhao、Y. Nie 和 B. Dai，《通过3D场景重建进行多摄像头统一预训练》，《IEEE机器人与自动化快报》，2024年。[142] C. Min、D. Zhao、L. Xiao、Y. Nie 和 B. Dai，《UniWorld：通过世界模型进行自动驾驶预训练》，arXiv预印本 arXiv:2308.07234，2023年。[143] C. Min、D. Zhao、L. Xiao、J. Zhao、X. Xu、Z. Zhu、L. Jin、J. Li、Y. Guo、J. Xing 等，《DriveWorld：通过世界模型进行自动驾驶的4D预训练场景理解》，arXiv预印本 arXiv:2405.04390，2024年。[144] C. Godard、O. Mac Aodha、M. Firman 和 G. J. Brostow，《深入研究自监督单目深度估计》，载于《IEEE/CVF国际计算机视觉会议论文集》，2019年，第3828–3838页。[145] Y. Wei、L. Zhao、W. Zheng、Z. Zhu、Y. Rao、G. Huang、J. Lu [141]

，和 J. Zhou ，“Surrounddepth：融合环绕视图

以实现自监督多摄像头深度估计，” arXiv 预印本 arXiv:2204.03636，2022年。[146] F. Wimbauer、N. Yang、C. Rupprecht 和 D. Cremers，“幕后：用于单视图重建的密度场”，2023年。

[147] [ ] 2301 07668 在线。可访问：http://arxiv.org/abs/。A.-Q. Cao 和 R. de Charette，“SceneRF：基于辐射场的自监督单目3D场景重建”，arXiv 预印本 arXiv:2212.02501，2022年。[148] A. Hayler、F. Wimbauer、D. Muhle、C. Rupprecht 和 D. Cremers，“S4C：基于神经场的自监督语义场景补全”，arXiv 预印本 arXiv:2310.07522，2023年。[149] Y. Huang、W. Zheng、B. Zhang、J. Zhou 和 J. Lu，“SelfOcc：自监督视觉三维占用预测”，arXiv 预印本 arXiv:2311.12754，2023年。[150] C. Zhang、J. Yan、Y. Wei、J. Li、L. Liu、Y. Tang、Y. Duan 和 J. Lu，“OccNeRF：基于神经辐射场的自监督多摄像头占用预测”，arXiv 预印本 arXiv:2312.09243，2023年。[151] A. Kirillov、E. Mintun、N. Ravi、H. Mao、C. Rolland、L. Gustafson、T. Xiao、S. Whitehead、A. C. Berg、W.-Y. Lo 等，“Segment Anything”，载于《IEEE/CVF国际计算机视觉会议论文集》，2023年，第4015–4026页。[152] S. Boeder、F. Gigengack 和 B. Risse，“OccFlowNet：通过可微渲染和占用流实现自监督占据估计”，arXiv 预印本 arXiv:2402.12792，2024年。[153] Z. Tan、Z. Dong、C. Zhang、W. Zhang、H. Ji 和 H. Li，“OVO：开放词汇占用”，arXiv 预印本 arXiv:2305.16133，2023年。[154] A. Vobecky、O. Simeoni、D. Hurych、S. Gidaris、A. Bursuc、P. Pérez 和 J. Sivic，“POP-3D：从图像进行开放词汇三维占用预测”，《神经信息处理系统进展》，第36卷，2024年。[155] X. Dong、J. Bao、Y. Zheng、T. Zhang、D. Chen、H. Yang、M. Zeng、W. Zhang、L. Yuan、D. Chen 等，“MaskCLIP：掩码自蒸馏推动对比式语言图像预训练”，载于《IEEE/CVF计算机视觉与模式识别会议论文集》，2023年，第10 995–11 005页。[156] R. Chen、J. Hu 和 W. Xu，“一种基于RRT-Dijkstra的自动驾驶车辆路径规划策略”，《应用科学》，第12卷，第23期，第11982页，2022年。[157] K. Lee 和 D. Kum，“碰撞规避／缓解系统：通过预测占用地图进行自动驾驶车辆运动规划”，《IEEE Access》，第7卷，第52 846–52 857页，2019年。[158] M. Montemerlo、J. Becker、S. Bhat、H. Dahlkamp、D. Dolgov、S. Ettinger、D. Haehnel、T. Hilden、G. Hoffmann、B. Huhnke 等，“Junior：斯坦福大学在城市挑战赛中的参赛车辆”，《野外机器人学期刊》，第25卷，第9期，第569–597页，2008年。[159] W. Zeng、W. Luo、S. Suo、A. Sadat、B. Yang、S. Casas 和 R. Urtasun，“端到端可解释神经运动规划器”，载于《IEEE/CVF计算机视觉与模式识别会议论文集》，2019年，第8660–8669页。[160] A. Sadat、S. Casas、M. Ren、X. Wu、P. Dhawan 和 R. Urtasun，“感知、预测与规划：通过可解释语义表征实现安全运动规划”，载于《计算机视觉—ECCV 2020：第16届欧洲计算机视觉会议，格拉斯哥，英国，2020年8月23–28日，论文集，第XXIII部分16》，Springer，2020年，第414–430页。[161] W. Tong、C. Sima、T. Wang、L. Chen、S. Wu、H. Deng、Y. Gu、L. Lu、P. Luo、D. Lin 等，“场景即占用”，载于《IEEE/CVF国际计算机视觉会议论文集》，2023年，第8406–8415页。[162] S. Biswas、S. Casas、Q. Sykora、B. Agro、A. Sadat 和 R. Urtasun，“QuAD：面向自动驾驶的基于查询的可解释神经运动规划”，arXiv 预印本 arXiv:2404.01486，2024年。[163] V. Dewangan、B. Sharma、T. Choudhary、S. Sharma、A. Aanegola、A. K. Singh 和 K. M. Krishna，“UAP-BEV：利用环视单目图像生成的鸟瞰视图进行不确定性感知规划”，载于2023年IEEE第19届自动化科学与工程国际会议（CASE），2023年，第1–8页。[164] J. Cheng、Y. Chen 和 Q. Chen，“PLUTO：突破基于模仿学习的自动驾驶规划上限”，arXiv 预印本 arXiv:2404.14327，2024年。[165] M. Cho、Y. Lee 和 K.-S. Kim，“利用集成屏障和占用栅格地图对自动驾驶车辆进行模型预测控制”，《IEEE机器人与自动化快报》，第8卷，2013年，2023年4月，2006，无页码。[166] B. Lu、G. Li、H. Yu、H. Wang、J. Guo、D. Cao 和 H. He，“面向复杂自动驾驶场景的基于自适应势场的路径规划”，《IEEE Access》，第8卷，第225 294–225 305页，2020年。[167] P. Raksincharoensak、T. Hasegawa 和 M. Nagai，《基于风险势优化框架的自动驾驶智能系统运动规划与控制》，International Journal of Automotive Engineering，第 7 卷，AVEC14 期，第 53–60 页，2016年。[168] Y. Hu、J. Yang、L. Chen、K. Li、C. Sima、X. Zhu、S. Chai、S. Du、T. Lin、W. Wang、L. Lu、X. Jia、Q. Liu、J. Dai、Y. Qiao 和 H. Li，《面向规划的自动驾驶》，载于 2023 IEEE/CVF计算机视觉与模式识别会议（CVPR），2023年，第 17 853–17 862 页。[169] Y. Hu、K. Li、P. Liang、J. Qian、Z. Yang、H. Zhang、W. Shao、Z. Ding、W. Xu 和 Q. Liu，《结合时空热力图的模仿：nuplan挑战赛第二名方案》，arXiv 预印本 arXiv:2306.15700，2023年。[170] Z. Huang、H. Liu、X. Mo 和 C. Lyu，《GameFormer规划器：面向自动驾驶车辆的学习型交互式预测与规划框架》，网址 https://opendrivelab. com/e2ead/AD23Challenge/Track 4 AID. pdf。[171] H. Liu、Z. Huang 和 C. Lv，《占用预测引导的自动驾驶神经规划器》，载于 2023 IEEE 第26届智能交通系统国际会议（ITSC）。IEEE，2023年，第 4859–4865 页。[172] D. Isele、R. Rahimi、A. Cosgun、K. Subramanian 和 K. Fujimura，《利用深度强化学习驾驶自动驾驶车辆通过遮挡交叉口》，载于 2018 IEEE国际机器人与自动化会议（ICRA），2018年，第 2034–2039 页。[173] D. M. Saxena、S. Bae、A. Nakhaei、K. Fujimura 和 M. Likhachev，《利用无模型强化学习在密集交通中驾驶》，载于 2020 IEEE国际机器人与自动化会议（ICRA）。IEEE，2020年，第 5385–5392 页。[174] G. Chen、Y. Zhang 和 X. Li，《基于深度强化学习的自动驾驶注意力式高速公路安全规划器》，IEEE车辆技术汇刊，2023年。[175] K. Rezaee、P. Yadmellat 和 S. Chamorro，《在不确定性条件下利用强化学习进行自动驾驶车辆运动规划》，载于 2021 IEEE/RSJ国际智能机器人与系统会议（IROS）。IEEE，2021年，第 3506–3511 页。[176] Z. Wen、Y. Zhang、X. Chen、J. Wang、Y.-H. Li 和 Y.-K. Huang，《TOFG：用于自动驾驶预测与规划的时序占用流图》，IEEE智能车辆汇刊，2023年。[177] C. You、J. Lu、D. Filev 和 P. Tsiotras，《利用强化学习和深度逆强化学习实现自动驾驶车辆的高级规划》，Robotics and Autonomous Systems，第 114 卷，第 1–18 页，2019年。[178] X. Li、Y. Zhang 和 X. Ye，《DrivingDiffusion：布局引导的潜在扩散模型多视角驾驶场景视频生成》，arXiv 预印本 arXiv:2310.07771，2023年。[179] X. Wang、Z. Zhu、G. Huang、X. Chen 和 J. Lu，《DriveDreamer：迈向由真实世界驱动的自动驾驶世界模型》，arXiv 预印本 arXiv:2309.09777，2023年。[180] G. Zhao、X. Wang、Z. Zhu、X. Chen、G. Huang、X. Bao 和 X. Wang，《DriveDreamer-2：大语言模型增强的多样化驾驶视频生成世界模型》，arXiv 预印本 arXiv:2403.06845，2024年。[181] A. Hu、L. Russell、H. Yeo、Z. Murez、G. Fedoseev、A. Kendall、J. Shotton 和 G. Corrado，《GAIA-1：用于自动驾驶的生成式世界模型》，arXiv 预印本 arXiv:2309.17080，2023年。[182] J. Yang、S. Gao、Y. Qiu、L. Chen、T. Li、B. Dai、K. Chitta、P. Wu、J. Zeng、P. Luo 等，《自动驾驶广义预测模型》，arXiv 预印本 arXiv:2403.09630，2024年。[183] X. Ren、J. Huang、X. Zeng、K. Museth、S. Fidler 和 F. Williams，《XCube：利用稀疏体素层级结构进行大规模三维生成建模》，arXiv 预印本 arXiv:2312.03806，2023年。[184] J. Lu、Z. Huang、J. Zhang、Z. Yang 和 L. Zhang，《WoVoGen：用于可控多摄像头驾驶场景生成的世界体积感知扩散》，arXiv 预印本 arXiv:2312.02934，2023年。[185] J. Zhang、Q. Zhang、L. Zhang、R. R. Kompella、G. Liu 和 B. Zhou，《通过语义占用图实现城市场景扩散》，arXiv 预印本 arXiv:2403.11697，2024年。[186] J. Lee、S. Lee、C. Jo、W. Im、J. Seon 和 S.-E. Yoon，《SemCity：利用三平面扩散进行语义场景生成》，arXiv 预印本 arXiv:2403.07773，2024年。[187] A. Hu、G. Corrado、N. Griffiths、Z. Murez、C. Gurau、H. Yeo、A. Kendall、R. Cipolla 和 J. Shotton，《用于城市驾驶的基于模型的模仿学习》，神经信息处理系统进展，第 35 卷，第 20 703–20 716 页，2022年。[188] Q. Li、X. Jia、S. Wang 和 J. Yan，《Think2Drive：在潜在世界模型中通过思考实现准真实自动驾驶的高效强化学习（CARLA-v2 16720，2024 2402）》，arXiv 预印本 arXiv: 。

[189] E. Leurent，《自动驾驶状态-动作表示综述》，2018年。[190] LightwheelAI 和 L. contributors，《LightwheelOcc：自动驾驶中的三维占用合成数据集》，https://github.com/ OpenDriveLab/LightwheelOcc，2024年。[191] A. Gu´edon 和 V. Lepetit，《SUGAR：用于高效三维网格重建和高质量网格渲染的表面对齐高斯泼溅》，arXiv 预印本 arXiv:2311.12775，2023年。[192] O. contributors，《视频生成模型作为世界模拟器》，https: // openai.com/index/video-generation-models-as-world-simulators/，2024年。[193] L. Wang、W. Zheng、Y. Ren、H. Jiang、Z. Cui、H. Yu 和 J. Lu，《 OccSora：作为自动驾驶世界模拟器的四维占用生成模型》，2024年。[194] H. Huang、A. Geiger 和 D. Zhang，《GOOD：探索开放世界物体检测的几何线索》，arXiv 预印本 arXiv:2212.11720，2022年。[195] H. Moravec 和 A. Elfes，《利用广角声呐获取高分辨率地图》，载于《论文集：1985 IEEE国际机器人与自动化会议》，第 2 卷。IEEE，1985年，第 116–121 页。[196] A. Elfes，《利用占用栅格实现移动机器人感知与导航》，Computer，第 22 卷，第 6 期，第 46–57 页，1989年。[197] S. Thrun，《概率机器人学》，ACM 通讯，第 45 卷，第 3 期，第 52–57 页，2002年。[198] R. Schnabel、R. Wahl 和 R. Klein，《用于点云形状检测的高效随机抽样一致算法》，载于《计算机图形学论坛》，第 26 卷，第 2 期。Wiley Online Library，2007年，第 214–226 页。[199] S. Kato、S. Tokunaga、Y. Maruyama、S. Maeda、M. Hirabayashi、Y. Kitsukawa、A. Monrroy、T. Ando、Y. Fujii 和 T. Azumi，《车载Autoware：利用嵌入式系统实现自动驾驶车辆》，载于 2018 ACM/IEEE 第9届国际信息物理系统会议（ICCPS）。IEEE，2018年，第 287–296 页。[200] S. Richter、Y. Wang、J. Beck、S. Wirges 和 C. Stiller，《利用单目和双目摄像头进行语义证据栅格地图构建》，Sensors，第 21 卷，第 10 期，文章号 3380，2021年。[201] T. Q. Tran、A. Becker 和 D. Grzechca，《利用二维激光扫描仪与三维超声波传感器的传感器融合为真实移动机器人构建环境地图》，Sensors，第 21 卷，第 9 期，文章号 3184，2021年。[202] R. Van Kempen、B. Lampe、T. Woopen 和 L. Eckstein，《用于证据占用栅格地图构建的仿真式端到端学习框架》，载于 2021 IEEE 智能车辆研讨会（IV）。IEEE，2021年，第 934–939 页。[203] K. C. Dietmayer、S. Reuter 和 D. Nuss，《融合环境数据的表示》，载于《驾驶辅助系统手册》。Springer，2016年，第 567–603 页。[204] K. Doherty、T. Shan、J. Wang 和 B. Englot，《利用贝叶斯广义核推断进行学习辅助的三维占用地图构建》，IEEE 机器人学汇刊，第 35 卷，第 4 期，第 953–966 页，2019年。[205] S. T. O'Callaghan 和 F. T. Ramos，《高斯过程占用地图》，《国际机器人研究杂志》，第 31 卷，第 1 期，第 42–62 页，2012年。[206] F. Ramos 和 L. Ott，《希尔伯特地图：利用随机梯度下降实现可扩展的连续占用地图构建》，《国际机器人研究杂志》，第 35 卷，第 14 期，第 1717–1730 页，2016年。[207] C.-C. Wang、C. Thorpe、S. Thrun、M. Hebert 和 H. Durrant-Whyte，《同时定位、地图构建与运动目标跟踪》，《国际机器人研究杂志》，第 26 卷，916，2007 9，889no. pp. – 。[208] M. Schreier、V. Willert 和 J. Adamy，《动态道路环境中的栅格地图构建：基于跟踪的动态单元假设分类》，载于 2014 IEEE国际机器人与自动化会议（ICRA）。IEEE，2014年，第 3995–4002 页。[209] D. Hahnel、R. Triebel、W. Burgard 和 S. Thrun，《移动机器人在动态环境中的地图构建》，载于 2003 IEEE国际机器人与自动化会议（目录号 03CH37422），第 2 卷。IEEE，2003年，第 1557–1563 页。[210] T.-D. Vu、O. Aycard 和 N. Appenrodt，《动态室外环境中结合运动目标跟踪的在线定位与地图构建》，载于 2007 IEEE 智能车辆研讨会。IEEE，2007年，第 190–195 页。[211] J. Saarinen、H. Andreasson 和 A. J. Lilienthal，《用于表示动态环境的独立马尔可夫链占用栅格地图》，载于 2012 IEEE/RSJ国际智能机器人与系统会议。IEEE，2012年，第 3489–3495 页。[212] M. E. Bouzouraa 和 U. Hofmann，《融合占用栅格地图构建与基于模型的目标跟踪：利用激光和雷达传感器实现驾驶辅助系统》，载于 2010 IEEE 智能车辆研讨会。IEEE，2010年，第 294–300 页。[213] S. T. O'Callaghan 和 F. T. Ramos，"面向动态环境的高斯过程占用地图"，载于《实验机器人学》。Springer，2016年，第791–805页。[214] V. Guizilini、R. Senanayake 和 F. Ramos，"动态希尔伯特地图：变化环境中的实时占用预测"，载于2019年国际机器人与自动化会议（ICRA）。IEEE，2019年，第4091–4097页。[215] C. Cou´e、C. Pradalier、C. Laugier、T. Fraichard 和 P. Bessi`ere，"用于多目标跟踪的贝叶斯占用滤波：一种汽车应用"，《国际机器人研究杂志》，第25卷，第1期，第19–30页，2006年。[216] M. Tay、K. Mekhnacha、C. Chen、M. Yguel 和 C. Laugier，"用于动态环境中目标跟踪的贝叶斯占用滤波器高效算法"，《国际车辆自主系统杂志》，第6卷，第1期，第155页，2008年。[217] R. Danescu、F. Oniga 和 S. Nedevschi，"用于基于立体视觉的环境感知的粒子网格跟踪系统"，载于2010年IEEE智能车辆研讨会。IEEE，2010年，第987–992页。[218] ——，"利用基于粒子的占用栅格对驾驶环境建模与跟踪"，《IEEE智能交通系统汇刊》，第12卷，第4期，第1331–1342页，2011年。[219] D. Nuss、S. Reuter、M. Thom、T. Yuan、G. Krehl、M. Maile、A. Gern 和 K. Diet mayer，"用于实时应用的动态占用栅格地图随机有限集方法"，《国际机器人研究杂志》，第37卷，第8期，第841–866页，2018年。[220] H. Fan、T. P. Kucner、M. Magnusson、T. Li 和 A. J. Lilienthal，"用于高度动态环境中有效占用滤波的双重PHD滤波器"，《IEEE智能交通系统汇刊》，第19卷，第9期，第2977–2993页，2017年。[221] S. Steyer、G. Tanzmeister 和 D. Wollherr，"利用证据映射和粒子跟踪进行基于栅格的环境估计"，《IEEE智能车辆汇刊》，第3卷，第3期，第384–396页，2018年。[222] G. Tanzmeister 和 D. Wollherr，"基于证据的栅格跟踪与映射"，《IEEE智能交通系统汇刊》，第18卷，第6期，第1454–1467页，2016年。[223] R. Danescu 和 S. Nedevschi，"用于动态数字高程地图建模与跟踪的基于粒子的方法"，《IEEE智能交通系统汇刊》，第15卷，第3期，第1002–1015页，2013年。[224] A. Vatavu、R. Danescu 和 S. Nedevschi，"利用策略树、占用栅格块和贝叶斯滤波器对拥挤交通场景进行建模与跟踪"，载于第17届IEEE智能交通系统国际会议（ITSC）。IEEE，2014年，第1948–1955页。[225] S. Steyer、G. Tanzmeister、C. Lenk、V. Dallabetta 和 D. Wollherr，"利用粒子标记进行基于栅格的目标跟踪中的数据关联"，载于2018年第21届智能交通系统国际会议（ITSC）。IEEE，2018年，第3036–3043页。[226] L. Rummelhard、A. N`egre 和 C. Laugier，"条件蒙特卡罗稠密占用跟踪器"，载于2015年IEEE第18届智能交通系统国际会议。IEEE，2015年，第2485–2490页。[227] A. Hornung、K. M. Wurm、M. Bennewitz、C. Stachniss 和 W. Bur- gard，"OctoMap：一种基于八叉树的高效概率三维映射框架"，《自主机器人》，第34卷，206，2013年3月，第189页。[228] L. Gan、R. Zhang、J. W. Grizzle、R. M. Eustice 和 M. Ghaffari，"用于可扩展稠密语义地图构建的贝叶斯空间核平滑"，《IEEE机器人与自动化快报》，第5卷，第2期，第790–797页，2020年。[229] Y. Kwon、B. Moon 和 S.-E. Yoon，"面向稠密锐利占用栅格的自适应核推断"，载于2020年IEEE/RSJ智能机器人与系统国际会议（IROS）。IEEE，2020年，第4712–4719页。[230] G. Chen、W. Dong、P. Peng、J. Alonso-Mora 和 X. Zhu，"利用粒子在动态环境中进行连续占用地图构建"，arXiv预印本 arXiv:2202.06273，2022年。[231] S. Shi、X. Wang 和 H. Li，"PointRCNN：基于点云的三维目标提议生成与检测"，载于IEEE/CVF计算机视觉与模式识别会议论文集，2019年，第770–779页。[232] S. Shi、C. Guo、L. Jiang、Z. Wang、J. Shi、X. Wang 和 H. Li，"PV-RCNN：面向三维目标检测的点 - 体素特征集抽象 "，载于IEEE/CVF计算机视觉与模式识别会议论文集，2020年，第10 529–10 538页。[233] X. Chen、H. Ma、J. Wan、B. Li 和 T. Xia，"面向自动驾驶的多视图三维目标检测网络"，载于IEEE计算机视觉与模式识别会议论文集，2017年，第1907–1915页。

[234] J. Ku、M. Mozifian、J. Lee、A. Harakeh 和 S. L. Waslander，“通过视图聚合联合生成三维候选并检测目标”，载于2018年IEEE/RSJ智能机器人与系统国际会议（IROS）。IEEE，2018年，第1–8页。[235] B. Yang、W. Luo 和 R. Urtasun，“PIXOR：基于点云的实时三维目标检测”，载于IEEE计算机视觉与模式识别会议论文集，2018年，第7652–7660页。[236] S. Mohapatra、S. Yogamani、H. Gotzig、S. Milz 和 P. Mader，“BEVDetNet：基于鸟瞰视图激光雷达点云的自动驾驶实时三维目标检测”，载于2021年IEEE智能交通系统国际会议（ITSC）。IEEE，2021年，第2809–2815页。[237] Y. Zhou 和 O. Tuzel，“VoxelNet：基于点云进行三维目标检测的端到端学习”，载于IEEE计算机视觉与模式识别会议论文集，2018年，第4490–4499页。[238] A. H. Lang、S. Vora、H. Caesar、L. Zhou、J. Yang 和 O. Beijbom，“PointPillars：用于点云目标检测的快速编码器”，载于IEEE/CVF计算机视觉与模式识别会议论文集，2019年，第12 697–12 705页。[239] S. Li、T. Wang、S. Yang 和 Y. Yuan，“EfficientPillarNet：用于激光雷达大规模点云语义分割的快速深度网络”，载于2021年光学成像与测量国际会议（ICOIM）。IEEE，2021年，第295–300页。[240] G. Shi、R. Li 和 C. Ma，“PillarNet：高性能基于柱体的三维目标检测”，arXiv预印本 arXiv:2205.07403，2022年。[241] Y. H. Khalil 和 H. T. Mouftah，“Licanet：基于多模态融合进一步提升联合感知与运动预测”，《IEEE智能交通系统开放期刊》，第3卷，第222–235页，2022年。[242] ——，“Licanext：纳入序列距离残差，进一步推进联合感知与运动预测”，IEEE Access，第9卷，第146 244–146 255页，2021年。[243] H. A. Mallot、H. H. B¨ulthoff、J. J. Little 和 S. Bohrer，“逆透视映射简化光流计算与障碍物检测”，《生物控制论》，第64卷，第3期，第177–185页，1991年。[244] L. Reiher、B. Lampe 和 L. Eckstein，“一种仿真到现实的深度学习方法：将多车载摄像头图像转换为鸟瞰视图中的语义分割图像”，2020年IEEE第23届智能交通系统国际会议（ITSC 2020），2020年。[245] N. Garnett、R. Cohen、T. Pe’er、R. Lahav 和 D. Levi，“3D-LaneNet：端到端三维多车道线检测”，载于2019年IEEE/CVF国际计算机视觉会议（ICCV），2019年，第2921–2930页。[246] J. Gu、B. Wu、L. Fan、J. Huang、S. Cao、Z. Xiang 和 X.-S. Hua，“用于单目三维目标检测的单应性损失”，载于IEEE/CVF计算机视觉与模式识别会议论文集，2022年，第1080–1089页。[247] A. Loukkal、Y. Grandvalet、T. Drummond 和 Y. Li，“在扁平移动体之间行驶：利用单目摄像头生成鸟瞰视图占用栅格以进行整体轨迹规划”，载于2021年IEEE计算机视觉应用冬季会议（WACV），2021年，第51–60页。[248] L. Song、J. Wu、M. Yang、Q. Zhang、Y. Li 和 J. Yuan，“用于多视图行人检测的堆叠单应变换”，载于2021年IEEE/CVF国际计算机视觉会议（ICCV），2021年，第6029–6037页。[249] K. Mani、S. Daga、S. Garg、N. S. Shankar、J. Krishna Murthy 和 K. M. Krishna，“Mono lay out：从单幅图像推断非模态场景布局”，《2020年IEEE计算机视觉应用冬季会议论文集》，WACV 2020，第1678–1686页，2020年。[250] A. Creswell、T. White、V. Dumoulin、K. Arulkumaran、B. Sengupta 和 A. A. Bharath，“生成对抗网络综述”，《IEEE信号处理杂志》，第35卷，第1期，第53–65页，2018年。[251] T. Roddick、A. Kendall 和 R. Cipolla，“用于单目三维目标检测的正交特征变换”，arXiv预印本 arXiv:1811.08188，2018年。[252] C. Reading、A. Harakeh、J. Chae 和 S. L. Waslander，“用于单目三维目标检测的类别深度分布网络”，载于2021年IEEE/CVF计算机视觉与模式识别会议（CVPR），2021年，第8551–8560页。[253] T. Wang、J. Pang 和 D. Lin，“利用运动深度估计进行单目三维目标检测”，arXiv预印本 arXiv:2207.12988，2022年。[254] Z. Wang、C. Min、Z. Ge、Y. Li、Z. Li、H. Yang 和 D. Huang，“STS：用于多视图三维检测的环视时序立体视觉”，arXiv预印本 arXiv:2208.10145，2022年。[255] T. Wang、Q. Lian、C. Zhu、X. Zhu 和 W. Zhang，“Mv-fcos3d++：基于预训练单目骨干网络的纯摄像头多视图四维目标检测”，arXiv预印本 arXiv:2207.12716，2022年。

[256] A. K. Akan 和 F. G¨uney，“StretchBEV：在空间和时间维度上拓展未来实例预测”，载于《计算机视觉—ECCV 2022：第17届欧洲会议，特拉维夫，以色列，2022年10月23日至27日，论文集，第XXXVIII部分》。Springer，2022年，第444–460页。[257] Z. Wu、Y. Gan、X. Li、Y. Wu、X. Wang、T. Xu 和 F. Wang，“用于代客泊车的环视鱼眼鸟瞰视图感知：数据集、基线与畸变不敏感多任务框架”，《IEEE智能车辆汇刊》，第i期，第1–12页，2022年。[258] T. Liang、H. Xie、K. Yu、Z. Xia、Z. Lin、Y. Wang、T. Tang、B. Wang 和 Z. Tang，“BEVFusion：一种简单且稳健的激光雷达-摄像头融合框架”，arXiv预印本 arXiv:2205.13790，2022年。[259] C. Lu、M. J. G. van de Molengraft 和 G. Dubbelman，“基于卷积变分编码器—解码器网络的单目语义占用栅格映射”，《IEEE机器人与自动化快报》，第4卷，第2期，第445–452页，2019年。[260] B. Pan、J. Sun、H. Y. T. Leung、A. Andonian 和 B. Zhou，“用于感知周围环境的跨视角语义分割”，《IEEE机器人与自动化快报》，第5卷，第3期，第4867–4873页，2020年。[261] N. Hendy、C. Sloan、F. Tian、P. Duan、N. Charchut、Y. Xie、C. Wang 和 J. Philbin，“Fishing net：网格中语义热力图的未来推断”，arXiv预印本 arXiv:2006.09917，2020年。[262] T. Roddick 和 R. Cipolla，“使用金字塔占据网络从图像预测语义地图表示”，载于《2020年IEEE/CVF计算机视觉与模式识别会议（CVPR）》，2020年，第11 135–11 144页。[263] A. Saha、O. Mendez、C. Russell 和 R. Bowden，“实现鸟瞰视角车辆估计中的时空聚合”，载于《2021年IEEE国际机器人与自动化会议（ICRA）》，2021年，第5133–5139页。[264] T.-Y. Lin、P. Doll´ar、R. Girshick、K. He、B. Hariharan 和 S. Belongie，“用于目标检测的特征金字塔网络”，载于《IEEE计算机视觉与模式识别会议论文集》，2017年，第2117–2125页。[265] W. Yang、Q. Li、W. Liu、Y. Yu、Y. Ma、S. He 和 J. Pan，“专注地投射你的视野：通过跨视角变换进行单目道路场景布局估计”，《IEEE计算机学会计算机视觉与模式识别会议论文集》，第15 531–15 540页，2021年。[266] J. Zou、J. Xiao、Z. Zhu、J. Huang、G. Huang、D. Du 和 X. Wang，“HFT：通过混合特征变换提升透视表示”，arXiv预印本 arXiv:2204.05068，2022年。[267] A. Vaswani、N. Shazeer、N. Parmar、J. Uszkoreit、L. Jones、A. N. Gomez、Ł. Kaiser 和 I. Polosukhin，“注意力就是一切”，《神经信息处理系统进展》，第30卷，2017年。[268] N. Carion、F. Massa、G. Synnaeve、N. Usunier、A. Kirillov 和 S. Zagoruyko，“使用Transformer实现端到端目标检测”，载于《欧洲计算机视觉会议》。Springer，2020年，第213–229页。[269] “特斯拉AI日2022”，2021年，http://www.npac.syr.edu/projects/cpsedu/summer98summary/examples/hpf/hpf.html。[270] B. Zhou 和 P. Kr¨ahenb¨uhl，“用于实时地图视角语义分割的跨视角Transformer”，载于《IEEE/CVF计算机视觉与模式识别会议论文集》，2022年，第13 760–13 769页。[271] Z. Xia、X. Pan、S. Song、L. E. Li 和 G. Huang，“具有可变形注意力的视觉Transformer”，载于《IEEE/CVF计算机视觉与模式识别会议论文集》，2022年，第4794–4803页。[272] L. Peng、Z. Chen、Z. Fu、P. Liang 和 E. Cheng，“BEVSegFormer：基于任意相机系统的鸟瞰视角语义分割”，arXiv预印本 arXiv:2203.04050，2022年。[273] L. Chen、C. Sima、Y. Li、Z. Zheng、J. Xu、X. Geng、H. Li、C. He、J. Shi、Y. Qiao 等，“PersFormer：通过透视Transformer进行三维车道线检测及OpenLane基准”，arXiv预印本 arXiv:2203 11089，2022年。。[274] P. Dutta、G. Sistu、S. Yogamani、E. Galv´an 和 J. McDonald，“Vit-bevseg：用于单目鸟瞰视图分割的分层Transformer网络”，载于《2022年国际神经网络联合会议（IJCNN）》。IEEE，2022年，第1–7页。[275] A. Saha、O. Mendez、C. Russell 和 R. Bowden，“将图像转换为地图”，载于《2022年国际机器人与自动化会议（ICRA）》，2022年，第9200–9206页。[276] Y. Jiang、L. Zhang、Z. Miao、X. Zhu、J. Gao、W. Hu 和 Y.-G. Jiang，“PolarFormer：使用极坐标Transformer进行多摄像头三维目标检测

[277] Y. Liu、J. Yan、F. Jia、S. Li、Q. Gao、T. Wang、X. Zhang 和 J. Sun，《Petrv2：基于多摄像头图像进行三维感知的统一框架》，arXiv 预印本 arXiv:2206.01256，2022年。[278] Y. Liu、T. Wang、X. Zhang 和 J. Sun，《Petr：用于多视角三维目标检测的位置嵌入变换》，arXiv 预印本 arXiv:2203.05625，2022年。[279] Z. Liu、S. Chen、X. Guo、X. Wang、T. Cheng、H. Zhu、Q. Zhang、W. Liu 和 Y. Zhang，《基于视觉的非均匀鸟瞰图表示学习：极坐标栅格化与表面估计》，arXiv 预印本 arXiv:2207.01878，2022年。[280] S. Chen、X. Wang、T. Cheng、Q. Zhang、C. Huang 和 W. Liu，《面向基于视觉的环视三维检测的极坐标参数化》，arXiv 预印本 arXiv:2206.10965，2022年。[281] Y. Li、Y. Chen、X. Qi、Z. Li、J. Sun 和 J. Jia，《利用 Transformer 统一基于体素的表示与三维目标检测》，arXiv 预印本 arXiv:2206.00630，2022年。[282] Z. Chen、Z. Li、S. Zhang、L. Fang、Q. Jiang、F. Zhao、B. Zhou 和 H. Zhao，《AutoAlign：面向多模态三维目标检测的像素-实例特征聚合》，arXiv 预印本 arXiv:2201.06493，2022年。[283] Z. Chen、Z. Li、S. Zhang、L. Fang、Q. Jiang 和 F. Zhao，《AutoAlignV2：面向动态多模态三维目标检测的可变形特征聚合》，arXiv 预印本 arXiv:2207.10316，2022年。[284] X. Chen、T. Zhang、Y. Wang、Y. Wang 和 H. Zhao，《FUTR3D：用于三维检测的统一传感器融合框架》，arXiv 预印本 arXiv:2203.10642，2022年。[285] X. Bai、Z. Hu、X. Zhu、Q. Huang、Y. Chen、H. Fu 和 C.-L. Tai，《TransFusion：利用 Transformer 进行稳健的激光雷达-摄像头融合以实现三维目标检测》，载于《IEEE/CVF计算机视觉与模式识别会议论文集》，2022年，第1090–1099页。[286] Z. Yang、J. Chen、Z. Miao、W. Li、X. Zhu 和 L. Zhang，《DeepInteraction：通过模态交互实现三维目标检测》，arXiv 预印本 arXiv:2208.11112，2022年。[287] Y. Jiao、Z. Jie、S. Chen、J. Chen、X. Wei、L. Ma 和 Y.-G. Jiang，《MSMDFusion：利用多深度种子在多个尺度融合激光雷达与摄像头以实现三维目标检测》，arXiv 预印本 arXiv: . .[288] L. Sless、B. E. Shlomo、G. Cohen 和 S. Oron，《通过语义分割从稀疏雷达簇进行占据栅格学习，实现道路场景理解》，载于《2019年IEEE/CVF国际计算机视觉研讨会》（ICCVW），2019年，第867–875页。[289] A. Popov、P. Gebhardt、K. Chen、R. Oldja、H. Lee、S. Murray、R. Bhargava 和 N. Smolyanskiy，《NVRadarNet：面向自动驾驶的实时雷达障碍物与自由空间检测》，2022年。 [ ] 2209 14499 在线。可访问：http://arxiv.org/abs/ 。[290] A. W. Harley、Z. Fang、J. Li、R. Ambrus 和 K. Fragkiadaki，《Simple-BEV：对于多传感器鸟瞰图感知，真正重要的是什么？》，载于 arXiv:2206.07959，2022年。[291] T. Zhou、J. Chen、Y. Shi、K. Jiang、M. Yang 和 D. Yang，《弥合雷达与摄像头特征之间的视角差异，实现多模态融合三维目标检测》，《IEEE智能车辆汇刊》，第1–14页，2023年。[292] J.-J. Hwang、H. Kretzschmar、J. Manela、S. Rafferty、N. Armstrong-Crews、T. Chen 和 D. Anguelov，《CramNet：利用射线约束交叉注意力实现稳健的摄像头-雷达融合三维目标检测》，第1图，第388–405页，2022年。[293] C. R. Qi、L. Yi、H. Su 和 L. J. Guibas，《Pointnet++：度量空间中点集的深层层次化特征学习》，arXiv 预印本 arXiv:1706.02413，2017年。[294] A. W. Harley、Z. Fang、J. Li、R. Ambrus 和 K. Fragkiadaki，《无需激光雷达的鸟瞰图感知简单基线》，arXiv 预印本 arXiv:2206.07959，2022年。[295] Y. Yang、J. Liu、T. Huang、Q.-L. Han、G. Ma 和 B. Zhu，《RaLiBEV：用于无锚框目标检测系统的雷达与激光雷达鸟瞰图融合学习》，第1–12页，2022年。 [在线]。可访问：http://arxiv.org/abs/ 2211 06108 。R. Xu、Z. Tu、H. Xiang、W. Shao、B. Zhou 和 J. Ma，《CoBEVT：利用稀疏 Transformer 进行协同鸟瞰视角语义分割》，arXiv 预印本 arXiv:2207.02202，2022年。[296] H. Li 和 F. Nashashibi，《一种占用栅格地图合并的新方法：应用于多车辆协同局部建图和室外环境中的运动目标检测》，载于《2012年第12届控制、自动化、机器人与视觉国际会议》（ICARCV）。IEEE，2012年，第632–637页。[297] X. Zheng、Y. Li、D. Duan、L. Yang、C. Chen 和 X. Cheng，《用于自动驾驶的多车辆多传感器占用栅格地图（MVMS-OGM）》，《IEEE物联网期刊》，2022年。[299] Y. Yuan、H. Cheng、M. Y. Yang 和 M. Sester，《在连续驾驶空间中生成证据式鸟瞰图》，arXiv 预印本 arXiv:2302.02928，2023年。[300] S. Yang、Y. Huang 和 S. Scherer，《通过高效高阶条件随机场进行语义三维占用映射》，载于2017年 IEEE/RSJ国际智能机器人与系统会议（IROS）。IEEE，2017年，第590–597页。[301] X. Chen、A. Milioto、E. Palazzolo、P. Giguere、J. Behley 和 C. Stach- niss，《Suma++：高效的基于激光雷达的语义SLAM》，载于2019年 IEEE/RSJ国际智能机器人与系统会议（IROS）。IEEE，2019年，第4530–4537页。[302] A. Milioto、I. Vizzo、J. Behley 和 C. Stachniss，《Rangenet++：快速且准确的激光雷达语义分割》，载于2019年 IEEE/RSJ国际智能机器人与系统会议（IROS）。IEEE，2019年，第4213–4220页。[303] R. Mur-Artal、J. M. M. Montiel 和 J. D. Tardos，《Orb-slam：一种通用且准确的单目SLAM系统》，《IEEE机器人学汇刊》，第31卷，第5期，第1147–1163页，2015年。[304] V. Vineet、O. Miksik、M. Lidegaard、M. Nießner、S. Golodetz、V. A. Prisacariu、O. K¨ahler、D. W. Murray、S. Izadi、P. P´erez 和 P. H. S. Torr，《用于大规模语义场景重建的增量式稠密语义立体融合》，载于2015年 IEEE国际机器人与自动化会议（ICRA），2015年，第75–82页。[305] Z. Zhu、S. Peng、V. Larsson、W. Xu、H. Bao、Z. Cui、M. R. Oswald 和 M. Pollefeys，《Nice-slam：用于SLAM的神经隐式可扩展编码》，载于IEEE/CVF计算机视觉与模式识别会议论文集，2022年，第12 786–12 796页。[306] A. Rosinol、J. J. Leonard 和 L. Carlone，《Nerf-slam：利用神经辐射场实现实时稠密单目SLAM》，arXiv 预印本 arXiv:2210.13641，2022年。[307] N. Homayounfar、W.-C. Ma、S. K. Lakshmikanth 和 R. Urtasun，《用于结构化在线地图的分层循环注意力网络》，载于IEEE计算机视觉与模式识别会议论文集，2018年，第3417–3426页。[308] Y. Liu、Y. Wang、Y. Wang 和 H. Zhao，《Vectormapnet：端到端矢量化高清地图学习》，arXiv 预印本 arXiv:2206.08920，2022年。[309] J. Shin、F. Rameau、H. Jeong 和 D. Kum，《InstaGraM：用于矢量化高清地图学习的实例级图建模》，2023年。 [在线]。可访问：http://arxiv.org/abs/ 2301 04470 。B. Liao、S. Chen、X. Wang、T. Cheng、Q. Zhang、W. Liu 和 C. Huang，《Maptr：用于在线矢量化高清地图构建的结构化建模与学习》，arXiv 预印本 arXiv:2208.14437，2022年。[311] R. Cipolla、Y. Gal 和 A. Kendall，《利用不确定性为场景几何与语义的损失加权的多任务学习》，载于2018年 IEEE/CVF计算机视觉与模式识别会议，2018年，第7482–7491页。[312] C. Lu 和 G. Dubbelman，《观测之外的幻觉：利用部分观测和非配对先验知识学习补全》，arXiv 预印本 arXiv:1907.09786，2019年。[313] T. Yin、X. Zhou 和 P. Kr¨ahenb¨uhl，《基于中心的三维目标检测与跟踪》，载于2021年 IEEE/CVF计算机视觉与模式识别会议（CVPR），2021年，第11 779–11 788页。[314] X. Zhu、W. Su、L. Lu、B. Li、X. Wang 和 J. Dai，《Deformable DETR：用于端到端目标检测的可变形Transformer》，载于第9届国际学习表征会议（ICLR 2021），虚拟会议，奥地利，2021年5月3日至7日。OpenReview.net，2021年。[在线]。可访问：https://openreview.net/forum?id=gZ9hCDWe6ke [315] E. Xie、W. Wang、Z. Yu、A. Anandkumar、J. M. ´Alvarez、T. Lu 和 P. Luo，《Segformer：利用Transformer进行语义分割的简洁高效设计》，载于《神经信息处理系统进展34：2021年神经信息处理系统进展年度会议》，NeurIPS 2021，2021年12月6日至14日，虚拟会议，M. Ranzato、A. Beygelzimer、Y. N. Dauphin、P. Liang 和 J. W. Vaughan 编，2021年，第12 077–12 090页。[在线]。可访问：https://proceedings.neurips.cc/paper/2021/hash/64f1f27bf1b4ec22924fd0acb550c235-Abstract.html[316] Z. Li、W. Wang、E. Xie、Z. Yu、A. Anandkumar、J. M. ´Alvarez、T. Lu 和 P. Luo，《Panoptic segformer：深入探究利用Transformer进行全景分割》，2022年 IEEE/CVF计算机视觉与模式识别会议（CVPR），第1270–1279页，2022年。[317] T. Wen、N. Wang、H. Wang、M. Wang、Z. Li 和 F. Tan，《SSLAD2022多任务学习统一模型第一名方案》，第1–7页。[318] B. Jiang、S. Chen、X. Wang、B. Liao、T. Cheng、J. Chen、H. Zhou、Q. Zhang、W. Liu 和 C. Huang，《感知、交互、预测：学习动态与静态线索进行端到端运动预测》，arXiv 预印本 arXiv:2212.02181，2022年。