基于纯视觉的多视角 4D 占用预测

摘要

在自动驾驶领域中，对周围环境的变化进行准确感知是实现安全驾驶的一个关键任务。最近的基于纯视觉的占用估计技术仅使用相机图像作为输入，可以提供大规模场景的密集占用表示。然而，现有的工作大多仅限于表示当前时刻的 3D 空间，而不考虑周围场景的未来状态。本文首先对现有的纯视觉占用预测工作进行了综述，接着在现有 4D 占用预测模型 OCFNet 的基础上进行了改进，在未来状态预测模块中对基于 3D 卷积的占用预测结果与根据 flow 预测的占用结果预测一个 Mask 进行结果融合，使用了对抗训练提出帧级对抗损失和序列级对抗损失，并提出了一个帧间语义完整性损失，本文对改进的模块做了消融实验验证了所改进的模块的有效性。

关键词： 自动驾驶；4D 占用预测

1 引言

精确地对周围的环境进行感知对自动驾驶以及机器人等领域是至关重要的，主流的方法包括基于纯视觉的方法、多模态的方法与纯 lidar 的方法。由于基于 lidar 的方法成本高昂，近年来基于纯视觉的方法得到了广泛的研究。传统的纯视觉感知方法如 3D 目标检测、语义分割等任务仅关注特定的预定义的类别，无法对预定义的类别之外的物体进行感知。最近的基于纯视觉的 3D 占用预测把周围的环境建模为体素网格，如图 1，并对每个网格赋予占用或者未占用两种状态，对于被占用的网格，还赋予了具体的类别，这种方法可以对任意物体的几何进行粗略的建模。

然而，现有的 3D 占用预测方法大多数只关注当前时刻的周围环境占的用状态，无法对未来时刻的空间占用状态进行建模。对于自动驾驶场景来说，对未来时刻的空间场景占用状态进行建模是非常有必要的，自动驾驶的防撞与轨迹优化策略需要能够感知未来时刻的空间

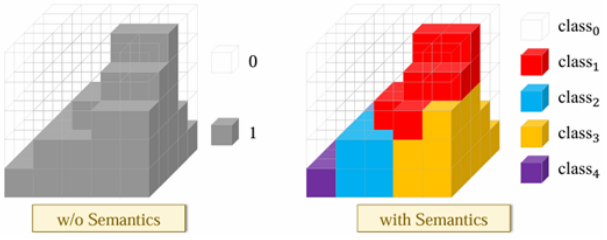

图 1. 占用网格

状态。最近在 3D 目标检测任务上已经有一些工作对 4D 感知进行了研究，如 BEVDet4D 等，但这种方法假设世界是一个平面，仅在 BEV 平面上进行感知，而忽略了物体的高度信息。4D 占用预测任务不仅能够估计当前时刻的空间占用状态还能够估计未来一段时间内的空间占用状态。

1.1 任务定义

4D 占用预测任务定义为给定 N_p 个过去和当前的连续多视角相机图像：$I = \{I_t\}_{t=-N_p}^{0}$，4D 占用预测任务旨在预测当前的占用状态:$O_c \in R^{1\times H\times W\times L}$ 以及未来一段时间 N_f 内的的占用状态 $O_f \in R^{N_f\times H\times W\times L}$，W，H，L 代表了体素空间的高度、宽度和长度。

2 相关工作

2.1 3D 占用预测

3D 占用预测把场景建模为体素空间，并对每个体素网格估计占用其状态。它用几何细节来表示空间，显著增强了复杂场景的表示能力。Cao 等人提出的 MonoScene [2] 首先解决了从相机图像中补全 3D 场景语义，但仅考虑前视图。相比之下，Huang 等人 [5] 用 TPVFormer 替换 MonoScene 的特征视线投影，以增强基于交叉注意机制的环绕视图占用预测的性能。Pan 等人的 UniOcc [13] 将基于体素的神经辐射场（NeRF）与占用预测相结合，以实现几何和语义渲染。Wang 等人 [18] 提出了一个名为 OpenOccupational 的大规模基准，它建立了具有高分辨率占用标签的 nuScenes-Occupational 数据集，并进一步提供了使用不同模式的几个 baseline。Tong 等人 [16] 还提出了一个占用预测基准 OpenOcc，并利用 OccNet 在各种任务上进行占用估计，包括语义场景补全，3D 对象检测，BEV 分割和运动规划。最近，Occ 3D [15] 利用遮挡推理和图像引导细化来进一步提高注释质量。与 OpenOcc 类似，Wei 等人的 SurroundOcc [19] 也产生密集的占用标签，并使用空间注意力将 2D 相机特征重新投影回 3D 体积。

2.2 4D 占用预测

4D 占用预测用于预测周围场景在不久的将来如何变化。现有的 4D 占用预测方法 [6,7,17] 主要使用 LiDAR 点云作为输入来捕捉周围结构的变化。例如，Khurana 等人。[7] 提出了一种可微光线投射方法，通过姿态对齐的 LiDAR 扫描来预测 2D 占用状态。最近，他们建议渲染未来的伪激光雷达点并估计占用率 [7]。其他点云预测方法 [3,9,10,12] 直接预测未来的激光点，可以将其体素化以进行未来的占用估计。然而，它们仍然需要连续的 LiDAR 点云，并且在预测过程中失去了语义一致性。与上述基于 LiDAR 的占用预测相比，在大规模场景中仅使用相机图像直接预测具有多个语义类别的未来 3D 占用仍然具有挑战性。因此，一些仅相机的语义/实例预测方法转向预测感兴趣对象的运动，例如，2D BEV 占用表示的一般车辆类别 [1, 4, 20]。例如，Hu 等人的 FIERY [4] 直接从多视图 2D 相机图像中提取 BEV 特征，然后结合时间卷积模型和递归网络来估计未来的实例分布。之后，提出了 StretchBEV [1] 和 BEVerse [20]，以进一步增强更长的时间范围。针对冗余输出的过度监督，最近提出了 PowerBEV [8]，以提高准确性和效率的预测性能。

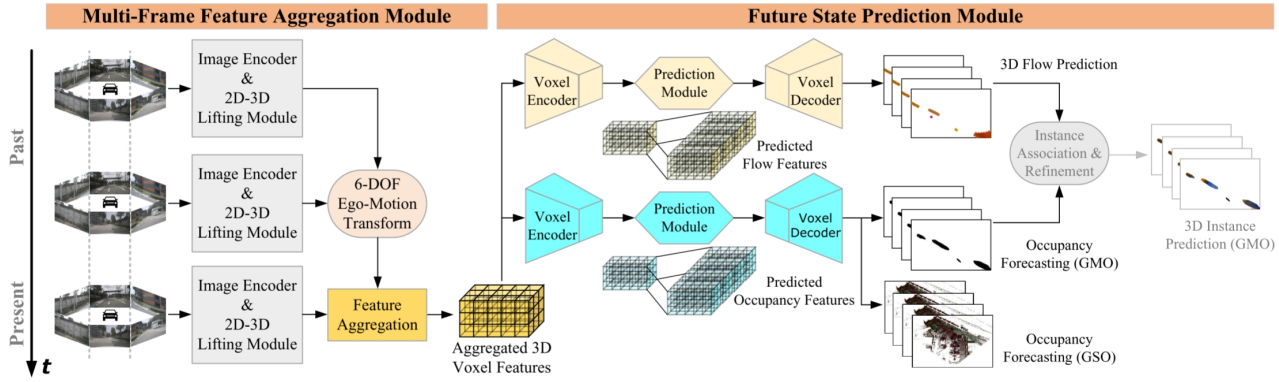

图 2. OCFNet 网络结构

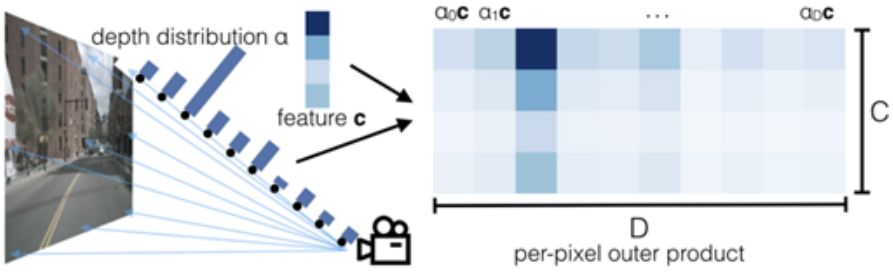

图 3. LSS 方法

3 本文方法

3.1 本文方法概述

OCFNet [11] 的模型结构如图 2所示，首先，其以多帧的图像作为输入，经过图像编码器得到多帧的图像特征，然后，通过 2D 到 3D 的特征提升模块将 2D 特征提升为 3D 特征，接着将多帧的 3D 特征根据自我运动对齐到当前帧并进行特征聚合。接着利用基于 3D 卷积的 occ 解码器和 flow 解码器分别预测出场景的占用状态和物体的向后向心 flow。

3.2 2D 到 3D 提升模块

为了将 2D 图像特征提升到 3D，本文采用了一种 LSS（Lift, Splat, Shoot）[14] 的方法，LSS 方法没有认为图像坐标系上的点与三维空间中的点是一一对应的。如图 3，他的做法是对每一个像素都估计一个 D 维深度分布 ，即每个像素可能对应三维空间中的 D 个点。然后每一个特征图上一个特征乘以这个 D 维深度分布进行加权被投影到三维空间中的 D 个点上。假设特征图的形状是 C×H×W 那么它被投影到 3d 空间中后会对应 H×W×D 个点，点的特征维度为 c。得到三维空间中的点及其特征后将其进行体素化处理。

3.3 多帧聚合模块

在得到 3D 特征之后，通过应用 6-DOF 自车姿态将 3D 特征体积变换到当前坐标系，产生聚合特征 $F_p \in \mathbb{R}^{(N_p+1)c\times h\times w\times l}$。在这里，我们将时间和特征维度压缩为一维，以实现以下

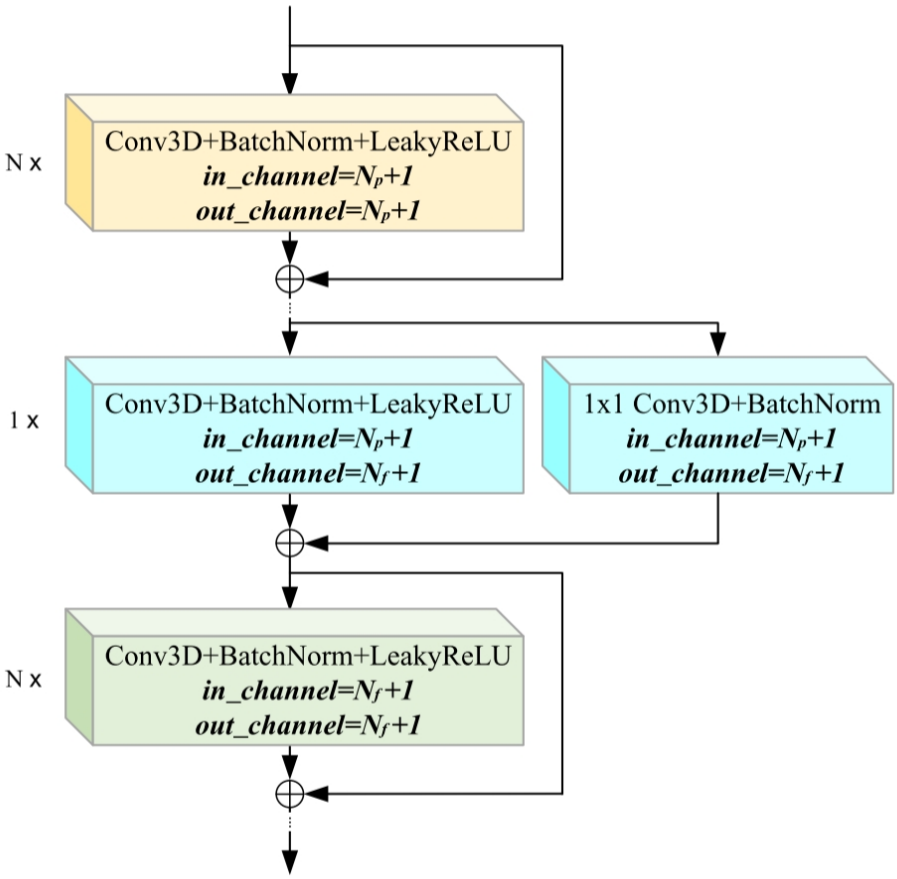

图 4. 预测模块

3D 时空卷积。随后，我们将其与相邻帧之间的 6-DOF 相对自车姿态连接，从而得到运动感知特征 $F_{pm} \in \mathbb{R}^{(N_p+1)(c+6)\times h\times w\times l}$

3.4 未来状态预测模块

经过聚合的特征通过使用两个预测头来同时预测网格的未来占用以及 flow。预测模块架构如图 4。首先，体素编码器将 F_{pm} 下采样为多尺度特征 $F_{pm}^{i} \in \mathbb{R}^{(N_p+1)c_i\times \frac{h}{2^i}\times \frac{w}{2^i}\times \frac{l}{2^i}}$，其中 i = 0, 1, 2, 3。然后，预测模块使用堆叠的 3D 残差卷积块将每个 F_{pm}^{i} 的信道维度扩展到 $(Nf+1)ci$，得到 $F_{pf}^{i} \in \mathbb{R}^{(N_f+1)c_i\times \frac{h}{2^i}\times \frac{w}{2^i}\times \frac{l}{2^i}}$。它们进一步与由体素解码器上采样的特征级联，之后在占用预测头中利用 softmax 函数来产生粗略的占用特征 $F_f^{occ} \in \mathbb{R}^{(N_f+1)\times cls\times h\times w\times l}$。在 flow 预测头中，用一个附加的 1×1 卷积层代替 softmax 函数来产生粗流特征 $F_f^{flow} \in \mathbb{R}^{(N_f+1)\times 3\times h\times w\times l}$。最后，我们利用 F_f^{occ} 和 F_f^{flow} 的三线性插值，以及占用状态维度上的附加 argmax 函数来生成最终的占用估计 $\hat{\mathbf{O}}_t \in \mathbb{R}^{(N_f+1)\times H\times W\times L}$ 和基于流的运动预测 $\hat{\mathbf{M}}_t \in \mathbb{R}^{(N_f+1)\times 3\times H\times W\times L}$。

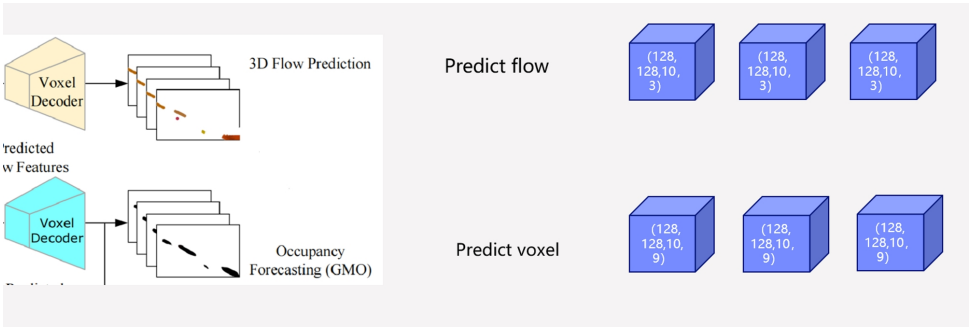

图 5. 原始预测模块

3.5 损失函数定义

我们使用交叉熵损失作为占用率预测损失 L_{occ}，并且使用平滑 l1 距离作为流量预测损失 L_{flow}。

$$L_{all} = \frac{1}{N_f + 1} \left(\sum_{t=0}^{N_f} \lambda_1 L_{occ}(\hat{\mathbf{O}}_t, \mathbf{O}_t) + \lambda_2 L_{flow}(\hat{\mathbf{M}}_t, \mathbf{M}_t) \right)$$

λ_1、λ_2 和 λ_3 是用于平衡占用预测、flow 预测的权重系数。

4 复现细节

4.1 与已有开源代码对比

该工作已开源了代码：https://github.com/haomo-ai/Cam4DOcc，本文在开源代码的基础上构建我们改进的模块，在未来状态预测模块中对基于 3D 卷积的占用预测结果与根据 flow 预测的占用结果预测一个 Mask 进行结果融合，使用了对抗训练提出帧级对抗损失和序列级对抗损失，并提出了一个帧间语义完整性损失。

4.1.1 两种结果的融合

原始基于 3D 卷积的解码器架构如图 5所示，多帧聚合得到的 3D 体素特征通过基于 3D 卷积的 voxel 解码器预测出未来两帧以及当前帧的空间占用状态和对应的 flow。占用预测和 flow 预测都由真值进行监督。其中预测的每帧的 flow 形状为 (H,W,D,3) 表示体素空间分辨率为 H,W,D 的每个体素的向后向心流 (x,y,z)。预测的每帧的占用状态形状 (H,W,D,numclass)。本文将 occ 的预测拆分成两个分支，一个分支保持原来的架构，另一个分支根据预测的向后向心 flow 从前一帧相应的位置利用三线性插值进行采样得到当前位置的结果，从而每一帧会有两种结果，一种是由原始预测模块预测的另一种是由 flow 采样生成的。接着我们预测一个值为 0 到 1 的 (H,W,D,1) 的 mask 对两种结果进行线性融合得到最终的输出。

4.1.2 对抗训练

我们首先对模型的输出使用 argmax 函数得到最终预测结果，形状为 (H,W,D,1) 我们设计了一个 4 层 3D 卷积的帧级判别器和序列级判别器进行对抗训练。对于帧级判别器，我们将时间维度视作 batch 维度，对于序列级判别器我们将时间维度视作通道维度，并且这两个

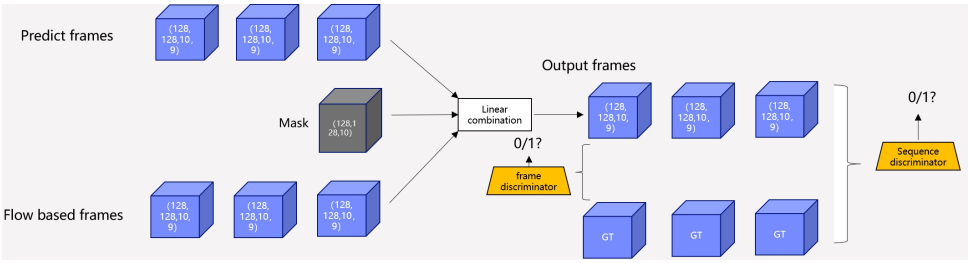

图 6. 改进后的预测模块

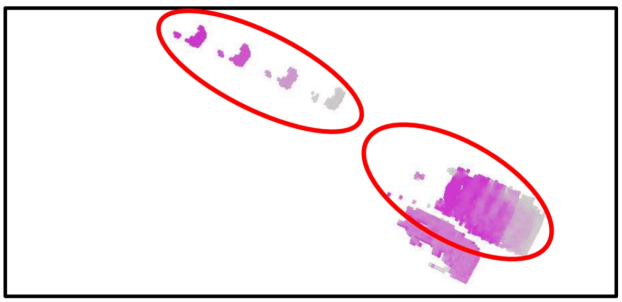

图 7. 运动时物体结构不完整

判别器之间不共享权重。因此我们有两种分类损失，一种是序列级判别器的分类损失，一种是帧级判别器的分类损失。改进后的未来状态预测模块如图 6所示。

4.1.3 帧间语义完整性损失

在原始的方法中，要对未来几帧的占用状态进行预测，但结果往往物体的结构信息不能被很好的保持，如图 7，为此我们提出了帧间语义完整性损失来解决这一问题。具体的做法是使用 GT 筛选出预测的三帧体素中真值为车的所有位置，并使他们的分布尽量的接近，这样可以更好的保持物体在运动过程中的结构完整。公式如下：

$$Loss_{integrity} = \frac{|Output \times class\ Mask - Class\ index|}{D}$$

4.2 创新点

本文改进的创新点主要有三个：

- 在未来状态预测模块中对基于 3D 卷积的占用预测结果与根据 flow 预测的占用结果预测一个 Mask 进行结果融合；
- 设计了一种帧级和序列级对抗训练的方法；
- 提出了一个帧间语义完整性损失，有效地缓解了运动中的结构不完整问题。

5 实验结果分析

使用了文中基于现有数据集 nuscene 和 occupancy-nuscene 提出的 4D 占用基准 Cam4Docc，这个基准除了提供占用表亲啊之外还提供了向后向心 flow 的标签。其中原始的训练配置是使

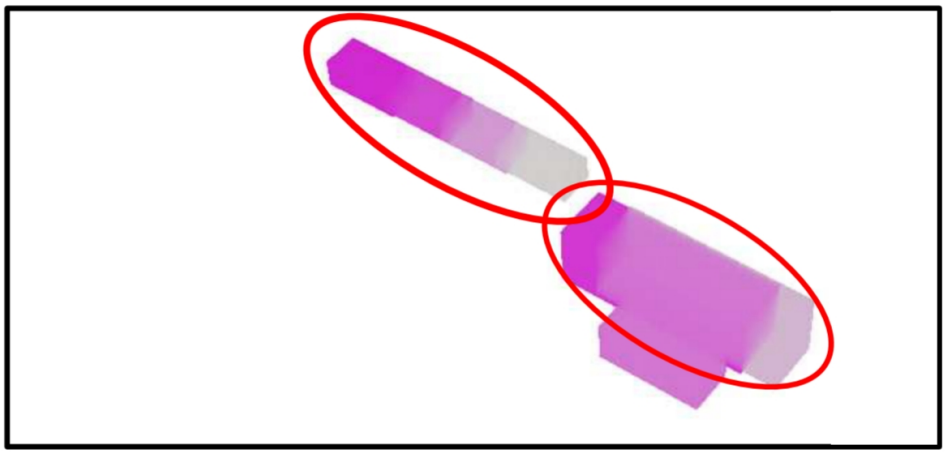

图 8. 加入帧间语义完整性损失后运动时结构保持完整

用 23930 个序列用于训练，5119 个序列用于测试，由于训练一次所需的时间较长，简单起见我们只使用了训练集和测试集的五分之一进行实验。这样，平均训练时间缩短为 20h。

本文所有实验都在 10 卡 v100 环境下完成。并在训练完成时评估他们的 mIOU，消融实验结果如下表所示，其中 origin 为原始论文中的 pipeline 结果，在我们加入对抗训练后性能提升了，mIOU 为 0.133，在我们对基于 flow 的预测结果和原始 occ 结果进行 mask 线性融合后性能继续提升至 0.134，我们还探索了其他的融合方式，如利用 Fumamba[21] 对特征进行融合等其他方式, 但效果并不如 mask 线性融合。在继续加入我们的帧间语义完整性损失后 mIOU 提升至 0.14。图 8为我们加入帧间语义完整性损失后的预测可视化效果，可以看到物体在运动时，物体能够较好的保持结构完整。

表 1. 消融实验结果

方法	free	bic.	bus	car	con.	moto.	truck	pede.	mean
origin	0.994	0.033	0.326	0.184	0.104	0.085	0.172	0.109	0.126
GAN	0.995	0.091	0.314	0.189	0.091	0.085	0.179	0.114	0.133
GAN+Mask	0.995	0.081	0.312	0.191	0.108	0.085	0.182	0.11	0.134
GAN+Fumamba	0.995	0.074	0.352	0.188	0.11	0.047	0.182	0.106	0.132
GAN+Mask+Loss	0.994	**0.102**	0.32	**0.18**	**0.11**	**0.106**	**0.194**	0.111	**0.14**

6 总结与展望

本文在现有 4D 占用预测模型 OCFNet(论文：Cam4DOcc: Benchmark for Camera-Only 4D Occupancy Forecasting in Autonomous Driving Applications [11]) 的基础上进行了改进，在未来状态预测模块中对基于 3D 卷积的占用预测结果与根据 flow 预测的占用结果预测一个 Mask 进行结果融合，使用了对抗训练提出帧级对抗损失和序列级对抗损失，并提出了一个帧间语义完整性损失，本文对改进的模块做了消融实验验证了所改进的模块的有效性。但本文的改进对模型性能的提升并不显著并且本文的方法只能预测未来几帧的占用状态，在未来可以充分考虑物体的运动特性进行长序列建模预测。

参考文献

[1] Adil Kaan Akan and Fatma Günbey. Stretchbev: Stretching future instance prediction spatially and temporally. In *ECCV*, pages 444–460, 2022.

[2] Anh-Quan Cao and Raoul de Charette. Monoscene: Monocular 3d semantic scene completion. In *CVPR*, pages 3991–4001, 2022.

[3] Hehe Fan and Yi Yang. Pointrnn: Point recurrent neural network for moving point cloud processing. *arXiv preprint arXiv:1910.08287*, 2019.

[4] Anthony Hu, Zak Murez, Nikhil Mohan, Sofía Dudas, Jeffrey Hawke, Vijay Badrinarayanan, Roberto Cipolla, and Alex Kendall. Fiery: Future instance prediction in bird' s-eye view from surround monocular cameras. In *ICCV*, pages 15273–15282, 2021.

[5] Yuanhui Huang, Wenzhao Zheng, Yunpeng Zhang, Jie Zhou, and Jiwen Lu. Tri-perspective view for vision-based 3d semantic occupancy prediction. In *CVPR*, pages 9223–9232, 2023.

[6] Tarasha Khurana, Peiyun Hu, Achal Dave, Jason Ziglar, David Held, and Deva Ramanan. Differentiable raycasting for self-supervised occupancy forecasting. In *ECCV*, pages 353–369, 2022.

[7] Tarasha Khurana, Peiyun Hu, David Held, and Deva Ramanan. Point cloud forecasting as a proxy for 4d occupancy forecasting. In *CVPR*, pages 1116–1124, 2023.

[8] Peizheng Li, Shuxiao Ding, Xieyuanli Chen, Niklas Hanselmann, Marius Cordts, and Juergen Gall. Powerbev: A powerful yet lightweight framework for instance prediction in bird' s-eye view. In *IJCAI*, pages 1080–1088, 2023.

[9] Fan Lu, Guang Chen, Zhijun Li, Lijun Zhang, Yinlong Liu, Sanqing Qu, and Alois Knoll. Monet: Motion-based point cloud prediction network. *TITS*, 23(8):13794–13804, 2021.

[10] Zhen Luo, Junyi Ma, Zijie Zhou, and Guangming Xiong. Pcpnet: An efficient and semantic-enhanced transformer network for point cloud prediction. *RA-L*, 2023.

[11] Junyi Ma, Xieyuanli Chen, Jiawei Huang, Jingyi Xu, Zhen Luo, Jintao Xu, Weihao Gu, Rui Ai, and Hesheng Wang. Proceedings of the ieee/cvf conference on computer vision and pattern recognition (cvpr), 2024. In *CVPR*, pages 21486–2149, 2024.

[12] Benedikt Mersch, Xieyuanli Chen, Jens Behley, and Cyrill Stachniss. Self-supervised point cloud prediction using 3d spatio-temporal convolutional networks. In *CoRL*, pages 1444–1454, 2022.

[13] Mingjie Pan, Li Liu, Jiaming Liu, Peixiang Huang, Longlong Wang, Shanghang Zhang, Shaoqing Xu, Zhiyi Lai, and Kuiyuan Yang. Uniocc: Unifying vision-centric 3d occupancy prediction with geometric and semantic rendering. *arXiv preprint arXiv:2306.09117*, 2023.

[14] J. Philion and S. Fidler. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In A. Vedaldi, H. Bischof, T. Brox, and JM. Frahm, editors, *Computer Vision – ECCV 2020*, volume 12359 of *Lecture Notes in Computer Science()*. Springer, Cham, 2020.

[15] Xiaoyu Tian, Tao Jiang, Longfei Yun, Yue Wang, Yilun Wang, and Hang Zhao. Occ3d: A large-scale 3d occupancy prediction benchmark for autonomous driving. *arXiv preprint arXiv:2304.14365*, 2023.

[16] Wenwen Tong, Chonghao Sima, Tai Wang, Li Chen, Silei Wu, Hanming Deng, Yi Gu, Lewei Lu, Ping Luo, and Dahua Lin. Scene as occupancy. In *ICCV*, pages 8406–8415, 2023.

[17] Maneekwan Toyungyernsub, Esen Yel, Jiachen Li, and Mykel J Kochendorfer. Dynamics-aware spatiotemporal occupancy prediction in urban environments. In *IROS*, pages 10836–10841, 2022.

[18] Xiaofeng Wang, Zheng Zhu, Wenbo Xu, Yunpeng Zhang, Yi Wei, Xu Chi, Yun Ye, Dalong Du, Jiwen Lu, and Xingang Wang. Openoccupancy: A large scale benchmark for surrounding semantic occupancy perception. In *ICCV*, pages 17850–17859, October 2023.

[19] Yi Wei, Linqing Zhao, Wenzhao Zheng, Zheng Zhu, Jie Zhou, and Jiwen Lu. Surroundocc: Multi-camera 3d occupancy prediction for autonomous driving. In *ICCV*, pages 21729–21740, 2023.

[20] Yunpeng Zhang, Zheng Zhu, Wenzhao Zheng, Junjie Huang, Guan Huang, Jie Zhou, and Jiwen Lu. Beverse: Unified perception and prediction in birds-eye-view for vision-centric autonomous driving. *arXiv preprint arXiv:2205.09743*, 2022.