CRUISE: V2X 场景中使用高斯喷溅的合作重建与编辑

Haoran Xu[1,2,7*], Saining Zhang[1,3*], Peishuo Li[3*], Baijun Ye[4], Xiaoxue Chen[1], Huan-ang Gao[1], Jv Zheng[1], Xiaowei Song[1], Ziqiao Peng[5], Run Miao[6], Jinrang Jia[7], Yifeng Shi[7], Guangqi Yi[7], Hang Zhao[4], Hao Tang[8], Hongyang Li[9], Kaicheng Yu[10], Hao Zhao[1,11†]

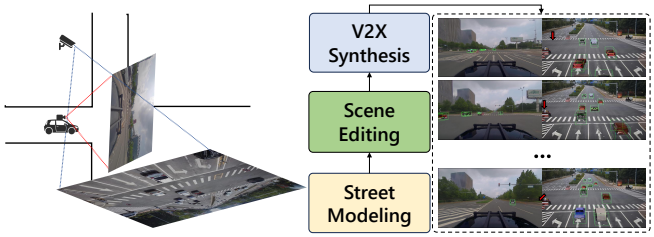

Fig. 1. 左侧是 V2X 设置，显示了来自自车和基础设施的捕获。右侧是 CRUISE 流程，它从街道建模到场景编辑，最终到 V2X 数据合成。红色箭头表示自车的位置。

随着自动驾驶技术的快速发展，通过提高交通安全性和效率来彻底改变交通运输的潜力日益明显。随着端到端自动驾驶模型的不断涌现，能够支持闭环和真实世界评估的可扩展、无领域差距的模拟需求日益迫切。同时，行之有效的模拟框架被认为是为训练可靠的自动驾驶系统以及确保其在各种驾驶条件下的可靠性所必需的 [1]–[7] 。

近年来，神经辐射场（NeRFs）[8]–[11] 和高斯斑点 (GS) [12] 已成为高保真三维场景重建的基础技术。在街景模拟的背景下，这些方法 [13]–[18] 主要侧重于重建静态场景，常常忽略动态元素，如移动的车辆。近期更先进的方法 [19]–[29] 引入了 4D 重建技术，可以高保真地捕捉到动态交通参与者和静态背景。

尽管前述研究在使用自车视角的模拟中取得了显著成果，但未来的端到端自动驾驶可能需要超越以车为中心的方法。最近，结合自车和基础设施传感器数据的 V2X 通信作为增强自动驾驶能力的一种有前途的范式出现了 [30], [31] 。许多工作探索了启用 V2X 的任务，包括检测 [32]–[36] 、跟踪 [37] 、分割 [38] 、定位 [39], [40] 和预测 [37], [41], [42] 。然而，这些工作主要集中于针对特定任务的方法，忽略了支持 V2X 研究和部署所需的闭环数据生态系统。在这项工作中，我们介绍了 CRUISE，这是第一个基于 GS 的 V2X 模拟框架。CRUISE 高保真地重建了真实驾驶环境，并丰富了 V2X 数据集。

为了实现高效的场景编辑和数据生成，CRUISE 首先通过分解的 GS 从 V2X 图像中重建环境，这有效地将动态车辆与静态街景分离。在此重建的基础上，我们引入了一种生成编辑范式，允许通过合成车辆高斯资产和调整场景构图来无缝修改交通场景。编辑完成后，重建的场景可以从自车和基础设施的角度进行渲染，生成用于下游任务的高保真 V2X 数据集，如图 1 所示。

* Equal contribution. [1]Institute for AI Industry Research (AIR), Tsinghua University; [2]Beijing Institute of Technology; [3]Nanyang Technological University; [4]Tsinghua University; [5]Renmin University of China; [6]Beijing University of Technology; [7]Baidu Inc; [8]Peking University; [9]Shanghai AI Lab;[10]Westlake University; [11]Beijing Academy of Artificial Intelligence.
† Corresponding author. zhaohao@air.tsinghua.edu.cn

为了评估我们数据生成框架的有效性，我们使用 V2X-Seq 数据集 [43] 验证 CRUISE 作为多种 3D 检测和协作 3D 跟踪任务的数据增强策略。通过利用我们高保真的 V2X 场景重建和可编辑的高斯表示，CRUISE 增强了自车辆、基础设施和协作视角的感知性能。这些发现强调了 CRUISE 作为一个多功能工具，通过可扩展的、逼真的数据增强来促进 V2X 感知的潜力。总之，这项工作做出了以下关键贡献：

- 我们介绍了 CRUISE，这是第一个专为 V2X 驾驶场景设计的基于 GS 的仿真框架，能够实现高保真重建和多视角合成。
- CRUISE 支持高保真场景重建和灵活编辑，能够从自车和基础设施视角生成多样化且逼真的 V2X 数据集，从而促进更好的模型训练。
- 我们的数据合成框架显著提高了自车、基础设施和协同视图中的 3D 检测以及协同 3D 跟踪。
- CRUISE 进一步促进了极端案例的生成，增强了数据集的多样性，并有助于构建一个数据驱动的闭环车联网驾驶系统。

I. 相关工作

自动驾驶模拟引擎，如 CARLA [44] 和 AirSim [45] ，面临着创建虚拟环境的高昂手动成本和生成数据缺乏真实感的问题。近年来，许多研究致力于从现实世界的自动驾驶数据构建模拟。早期技术 [46]–[51] 通常专注于通过 LiDAR 或多视角图像进行重建，但未能实现高保真新视角合成（NVS）。

现如今，3D 重建技术的快速发展，包括 NeRF [8] 和 3DGS [12] ，在自动驾驶领域引起了极大的关注。对于基于 NeRF 的方法，Block-NeRF [13] 和 Mega-NeRF [52] 通过对分块建模重建了一个大型街道场景。SUDS [20] 和 EmerNeRF [19] 学习了室外场景的分解。此外，其他方法 [23], [24], [53]–[56] 利用神经场将场景建模为移动物体网络和背景网络的组合。然而，这些方法存在高计算成本的问题。

对于 GS，PVG [57] 使用周期性振动 3D 高斯模型来模拟动态城市场景。DrivingGaussian [27] 引入了复合动态高斯图和增量 3D 静态高斯，而 S^3 Gaussian [29] 以自监督的方式区分动态和静态场景，无需额外的注释。虽然这些基于 GS 的方法计算效率高且具有高保真度，但它们无法完美重建车辆和街道。在这项工作中，我们利用了 Street Gaussians [28] ，其优化了动态高斯的跟踪姿态，并引入 4D 球谐变换在帧间改变车辆外观，以在 V2X 仿真中实现车辆和街道的精确变形。

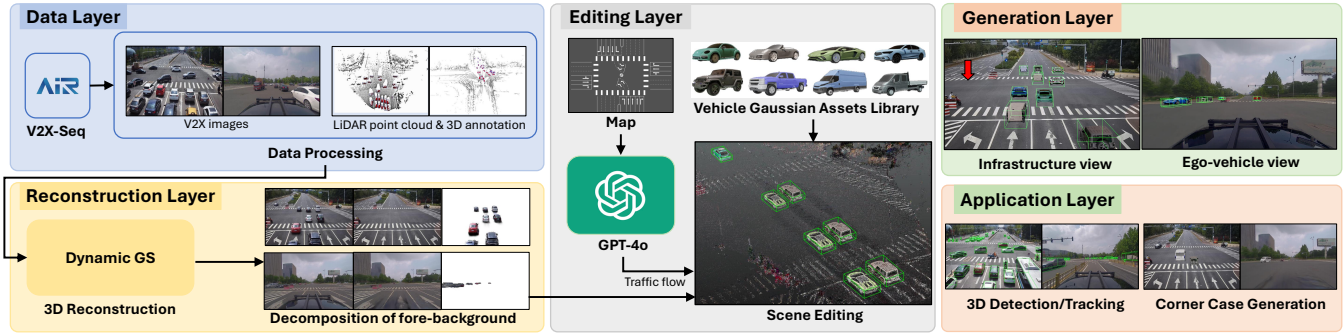

Fig. 2. CRUISE 的工作流程。数据层将 V2X-Seq 数据集处理成适合进一步重构的格式。经过处理的数据用于基于动态 GS 建模在重构层中分解前景与背景场景。在编辑层中，矢量地图和其他信息被输入 GPT-4o 生成可能的交通流量，并用车辆高斯特性编辑场景。生成层随后渲染一个新 V2X 数据集及其相应的注释。最终，新生成的数据集可以应用于 3D 检测/跟踪、极端情况生成和其他进一步的应用。红色箭头指向自车的位置。

A. 车联网合作感知

V2X 协作 [32], [33], [58]–[70] 感知是一个新兴的应用于 V2X 辅助系统中的应用，通过交换互补的感知信息显著增强了自动驾驶感知模块。已经开发出几个先进的平台 [60], [62], [64], [66] 来模拟协作感知场景，提供支持这些系统开发的各种感知标注。特别是，协作感知系统取得了重大进展，CoCa3D [33] 实现了接近完整的感知能力。此外，V2Xverse [71] 引入了一个端到端的协作驾驶系统，旨在促进基于 V2X 的自动驾驶。在这项工作中，我们利用成熟的协作感知基准 V2X-Seq [43] 来评估我们模拟方法的有效性。

II. & 公式预备知识

3D-GS [12] 用一组 3D 高斯来表示一个 3D 场景。高斯的几何由 RSS^TR^T 决定，其中 $R \in \mathbb{R}^{3\times3}$ 是旋转矩阵，$S \in \mathbb{R}^{3\times3}$ 是缩放矩阵。

为了通过平坦化 3D 高斯来有效地建模曲面，我们应用了 [72] 中提出的尺度损失。该损失最小化每个高斯的缩放因子 $\boldsymbol{s} = (s_1, s_2, s_3)^\top \in \mathbb{R}^3$ 的最小分量，将其驱向零：

$$\mathcal{L}_{\text{scale}} = \| \min(s_1, s_2, s_3) \|_1. \tag{1}$$

为了在渲染过程中减少针状伪影，我们建议将 s_1 设为最长的缩放因子，s_2 为第二长的。此外，我们应用一个比例损失来确保高斯近似为圆形：

$$\mathcal{L}_{\text{ratio}} = \max(1, s_1/s_2) - 1. \tag{2}$$

III. 提出的方法

CRUISE 旨在通过动态 GS 模型的重建和渲染技术生成高保真、几乎无限的 V2X 驾驶数据。CRUISE 管道如图 2 所示，并且其层次结构分为五个不同的层。第一个数据层将 V2X-Seq 数据集转换为适合进一步重建的格式。处理后的数据随后用于在重建层中基于动态 GS 建模分解前景-背景场景。在编辑层，矢量地图和其他信息被输入到一个多模态大语言模型中，以生成可能的交通流并使用车辆高斯资产编辑场景。生成层随后渲染一个新的 V2X 数据集并产生相应的注释，便于下游任务。最后，新生成的数据集可以应用于 3D 检测/跟踪任务以及其他未来应用。

数据集的质量对重建和渲染过程的结果有着显著影响。在这项工作中，我们使用经过精心收集的真实世界 V2X 数据集 V2X-Seq [43] 作为我们的数据来源，该数据集包含丰富的 RGB 图像、来自 LiDAR 的点云以及来自自车和基础设施视角的 3D 注释。

在数据处理之前，我们建立了一个基于 GS 的重建基线，这是一种基于点的渲染方法，用 3D 高斯表示场景几何。GS 方法在 3D 重建和新视图合成（NVS）中已实现了最先进的（SoTA）性能，其显式表示非常适合我们的编辑需求。我们采用 Street Gaussians [28] ，这是一种用于街景重建的顶尖方法，因为它能够有效地分解物体和背景，从而促进我们流程中的数据生成。

接下来，我们将数据转换为 Street Gaussians 所需的格式。由于基线方法使用标注框独立训练 4D 对象和 3D 背景，V2X 数据集带来了挑战：基础设施视图标注包含自车的框。在自车视图训练时，自车的边界框通常覆盖了大部分可见场景，导致静态背景元素被重建为动态高斯。这一问题削弱了动静态组件的正确解耦。为了解决这个问题，我们从场景中移除自车框，并引入一个覆盖自车视图中自车区域的自车遮罩，从而实现物体与背景的更精确分离。

为了实现稳健的初始化，我们融合来自自车和基础设施的 LiDAR 点云。为了增强目标的高斯密度，我们使用跟踪框聚合多帧点云。训练过程中，结合了两个视角的 LiDAR 深度监督。此外，我们采用了 GOF [73] 的外观解耦策略，该策略使用轻量级卷积神经网络来模拟真实场景中的不均匀照明，从而减少伪影并提高数据生成中的高保真渲染效果。

GS 训练期间的损失函数设置如下：

$$\begin{aligned}\mathcal{L} = \mathcal{L}_{\text{color}} + \lambda_1\mathcal{L}_{\text{depth}} + \lambda_2\mathcal{L}_{\text{normal}} + \lambda_3\mathcal{L}_{\text{sky}} \\ + \lambda_4\mathcal{L}_{\text{sem}} + \lambda_5\mathcal{L}_{\text{scale}} + \lambda_6\mathcal{L}_{\text{ratio}} + \lambda_7\mathcal{L}_{\text{reg}}.\end{aligned} \tag{3}$$

在方程 (3) 中，$\mathcal{L}_{\text{color}}$ 表示在 GOF [73] 下，渲染图像与观测图像之间的重构损失，其中外观被解耦。$\mathcal{L}_{\text{depth}}$ 和 $\mathcal{L}_{\text{normal}}$ 分别是渲染深度和法线之间的 L1 损失，以及 LiDAR 深度和由 StableNormal 生成的法线之间的损失 [74] 。$\mathcal{L}_{\text{sky}}$ 是一个用于天空监督的二值交叉熵损失，天空掩码由 Grounded SAM 2 生成 [75] 。$\mathcal{L}_{\text{sem}}$ 是渲染语义对数和输入 2D 语义分割预测之间的每像素 softmax 交叉熵损失 [76] 。$\mathcal{L}_{\text{scale}}$ 和 $\mathcal{L}_{\text{ratio}}$ 帮助约束高斯的几何形状为扁平的圆形，以改善表面重建。最后，$\mathcal{L}_{\text{reg}}$ 是一个用于去除悬浮物和增强分解效果的正则化项。

重建后，车辆和街道可以如图 2 所示清晰地分解，这有助于灵活的场景编辑。

在编辑过程中，我们设计了如图 3 所示的基于生成的流程。首先，我们从互联网上收集多视角的车辆图像及

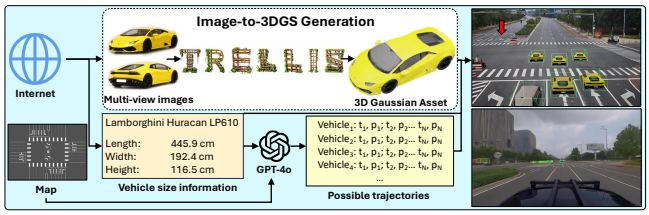

Fig. 3. 场景编辑的流程。通过从互联网收集基本的车辆信息，使用 TRELLIS 处理多视角图像以生成三维高斯车辆资产。同时，车辆尺寸、矢量地图以及自车轨迹被提供给 GPT-4o 作为输入，GPT-4o 输出以帧索引 $\boldsymbol{t}_i$ 和对应姿态 $\boldsymbol{p}_i$ 的形式表示的可能轨迹。最后，使用车辆资产、盒子尺寸和生成的轨迹将车辆放置在场景中，以生成新的 V2X 数据。红色箭头指示了自车的位置。

其尺寸。这些图像随后被送入 TRELLIS [77]，这是一种先进的 3D 生成方法，用于生成具有准确几何形状和外观的高质量 3D 高斯车辆资源。同时，将车辆尺寸、矢量地图和自车轨迹提供给 GPT-4o，以生成可能的场景插入轨迹。具体地，由于所有场景都是交叉路口，我们首先简化矢量地图。基于基础设施的 LiDAR 定位，我们选择最近的标记为 CITY_DRIVING 并带有 is_intersection 标志的车道。然后，我们使用广度优先搜索方法循环添加相邻的车道，直到序列中的所有相关车道都被包括进来，并通过可视化车道中心线验证。接下来，我们提供每条车道中心线的起始位置和方向作为输入给 GPT-4o。此外，我们将每 10 帧采样一次的自车轨迹作为序列提供：$\boldsymbol{t}_1$ ，$\boldsymbol{p}_1$ ；$\boldsymbol{t}_2$ ，$\boldsymbol{p}_2$ ；…，其中 $\boldsymbol{t} \in 0, 10, 20, \ldots, N$ 表示帧索引，而 $\boldsymbol{p} \in \mathbb{R}^3$ 表示位置。随后，GPT-4o 利用其多模态推理能力生成每辆附加车辆的合理轨迹，格式与自车轨迹相同。我们应用插值来估算每帧的位置信息，并与车辆尺寸结合，构建跟踪框用于在场景中插入高斯车辆。

编辑后，我们为每一帧从自车和基础设施视角渲染新的 V2X 图像。对于自车视图，我们将自车区域贴回以更真实地再现原始数据。同时，我们为所有车辆生成 3D 标注框，形成一个完全标注的合成 V2X 数据集。带有标注框的渲染图像如图 4 所示。

A. 下游应用

CRUISE 的下游应用包括多个方面，包括三维检测/跟踪、极端情况生成，以及 V2X 任务的闭环训练和测试。

随着端到端自动驾驶技术的进步，V2X 通信在安全和完全自主方面变得越来越重要。然而，大多数研究侧重于算法的改进，忽视了大型、多样化数据集在有效训练中的关键作用。

CRUISE 重建了真实世界的场景并生成了广泛的 V2X 数据集，包括分布外（OOD）情况，从而使更复杂的模型能够以更高的准确性和泛化能力进行训练。此外，CRUISE 促进了特殊情况生成，帮助自动化系统更好地应对具有挑战性的驾驶情景。凭借其强大的数据合成功能，CRUISE 有可能为未来的 V2X 研究中的闭环训练和评估建立一个基准。

在这项工作中，我们主要评估生成的数据在 V2X 任务上的表现，如 3D 物体检测和跟踪，展示了 CRUISE 框架的有效性。

IV. 实验

A. 数据集

V2X-Seq。V2X-Seq 数据集 [43] 是一个基于 DAIR-V2X-C 数据集 [64] 的基准，用于自动驾驶中的车路协同感知。它包含在 6 个路口的 95 个交通场景中以 10 Hz 捕获的真实世界数据，每个场景持续 10 到 20 秒。该数据集包括从车辆和基础设施单元中获取的高频录音，每个单元配备了 LiDAR 和摄像头，提供交通动态的多模态视图。它还为每个序列中的每个感兴趣对象提供 3D 跟踪注释，其中相同对象共享唯一跟踪 ID，并附带矢量地图。

TABLE I
V2X-Seq 的重构结果。S^3 -GS 是 S^3 高斯，Street-GS 是 Street 高斯。

Method	Box	PSNR ↑	SSIM ↑	LPIPS ↓
3D-GS [12]		24.79	0.902	0.158
PVG [57]		28.51	0.924	0.115
S^3 -GS [29]		27.54	0.931	0.097
HUGS [81]	✓	27.23	0.925	0.101
Street-GS(Ours) [28]	✓	27.97	0.940	0.095

B. 实施细节

所有实验均在 NVIDIA A800 GPU 上进行。为了评估我们框架的效率，我们从 V2X-Seq 数据集的 4 个交叉口中选择了 6 个序列进行场景重建、生成新数据，并评估生成数据在下游任务中的有效性。

对于重建，我们训练 Street Gaussians 进行 50,000 次迭代。对于 3D 检测，我们在基准上选择了 SoTA 方法：MonoLSS [78] 用于仅车辆视图，Bevheight [79] 用于仅基础设施视图，ImVoxelNet [80] 用于协作视图。ImVoxelNet 还可以执行协作 3D 跟踪。MonoLSS 训练 150 个周期，Bevheight 训练 100 个周期，ImVoxelNet 训练 24 个周期。

C. 重建比较

在编辑和生成之前，我们比较了不同方法的重建结果。我们在基于 GS 的基线、3D-GS [12] 、PVG [57] 、S^3 Gaussian [29] 、HUGS [81] 和 Street Gaussians [28] 上测试了相同的 V2X-Seq 的 6 个序列。我们报告了平均 PSNR（峰值信噪比）、SSIM（结构相似性指数）和 LPIPS（学习的感知图像块相似性）。

从表格 I 可以看出，我们合成使用的 Street Gaussians 在场景重建的渲染中表现优于其他方法。此外，通过利用物体框在训练期间区分物体和背景，它有助于后续的场景编辑任务。

D. 重建模块的消融研究

表 II 展示了我们设计模块在 V2X-Seq 上的量化结果。损失函数的有效性。表 II 中展示的结果表明，包含普通损失和几何损失（$\mathcal{L}_{\text{scale}}$ 和 $\mathcal{L}_{\text{ratio}}$ ）有助于改善场景的几何建模，从而实现高保真重建。

自我遮罩的有效性。表格 II 清楚地表明，缺少自我遮罩导致在自我视图中难以将自我车辆部件与背景区分开，从而导致重构和数据生成性能不佳。

外观解耦的有效性。从表格 II 可以看出，包含外观解耦可以改善重建结果，因为它有助于高斯学习一致的几何形状和颜色，而不是为了补偿视角依赖的外观。

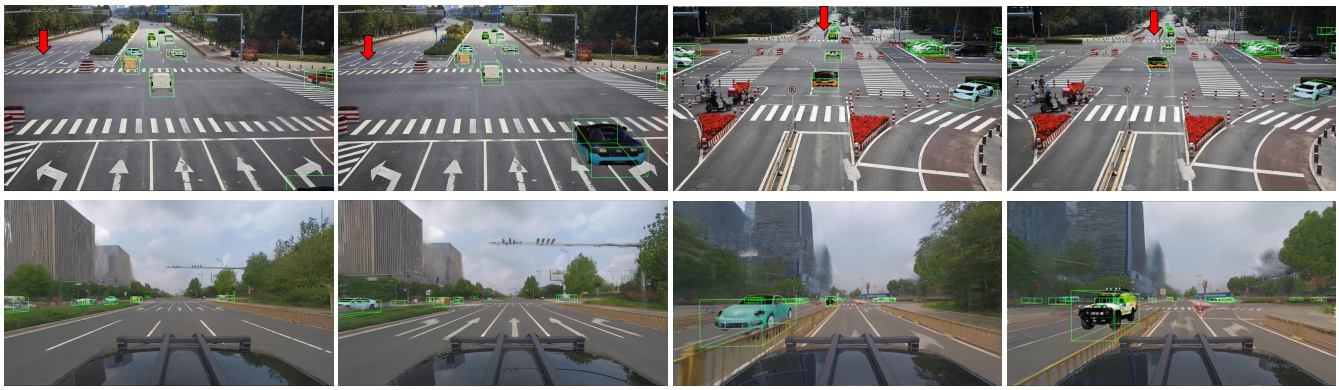

Fig. 4. CRUISE 生成的数据。由上到下：基础设施视图；对应的自车视图。红色箭头指向自车的位置。

Fig. 5. 在原始 V2X-Seq 数据集上合作式 3D 检测的定性结果。上排显示基础设施视图，而下排显示相应的自车视图。从左到右依次为：真实值；真实 + 生成；生成；真实。结果表明，使用增强数据进行训练可以提高检测准确性和车辆识别能力。

TABLE II
重建模块的消融研究

Methods	PSNR ↑	SSIM ↑	LPIPS ↓
Ours w/o Normal loss	27.45	0.933	0.103
Ours w/o Geo. loss	27.37	0.935	0.100
Ours w/o Ego-mask	24.44	0.910	0.125
Ours w/o Appearance	26.89	0.929	0.117
Complete method	27.97	0.940	0.095

TABLE III
V2X-Seq 上车辆视角 3D 检测的定量结果

Train data	$AP_{3D}(IOU=0.7)\uparrow$			$AP_{BEV}(IOU=0.7)\uparrow$		
	Easy	Mod.	Hard	Easy	Mod.	Hard
Real	56.84	37.69	34.14	66.81	46.62	44.21
Gen	54.41	36.77	34.02	65.44	47.13	45.33
Real + Gen	**61.35**	**41.05**	**38.71**	**67.38**	**49.72**	**47.24**

TABLE IV
V2X-Seq 上基础设施视角 3D 检测的定量结果

Train data	$AP_{3D}(IOU=0.7)\uparrow$			$AP_{BEV}(IOU=0.7)\uparrow$		
	Easy	Mod.	Hard	Easy	Mod.	Hard
Real	60.03	50.16	50.13	74.86	62.54	62.53
Gen	63.72	52.22	52.19	75.78	61.65	61.64
Real + Gen	**64.10**	**54.26**	**52.40**	**76.35**	**64.19**	**64.17**

TABLE V
V2X-Seq 上合作视角 3D 检测/跟踪的定量结果

Train data	3D Detection		3D Tracking	
	$AP_{3D}\uparrow$ (IOU=0.5)	$AP_{BEV}\uparrow$ (IOU=0.5)	MOTA ↑	MOTP ↑
Real	14.79	19.75	21.83	56.65
Gen	14.99	20.09	25.03	57.29
Real + Gen	**15.91**	**20.74**	**25.52**	**58.15**

E. 3D 检测/跟踪的结果

对于下游任务，我们在 V2X-Seq 中从 4 个十字路口中构建了一个包含 8 个序列的测试集，其中不包括用于数据生成的部分。对于训练集，我们在重建场景上生成了 1000 帧，并从用于重建的真实序列中选择了相同数量的帧，以进行公平比较。

单视角 3D 检测的结果。如表 III 所示，生成的数据在 3D 检测任务的两端都取得了可与真实数据媲美的结果，这表明生成的数据在真实性和标签准确性方面可以匹敌甚至超过真实数据。为了进一步评估其多样性，我们将生成的数据作为数据增强的一种形式，与真实数据一起进行了一半训练迭代。表 III 显示，这种混合训练策略显著提高了检测模型的性能，表明生成的数据提供了高质量的 OOD 样本，增强了模型的泛化能力。这些结果表明，生成的数据可以在实际应用中作为一种有价值的补充，为单视角 3D 检测任务提供强大而可靠的训练样本。

协作视角下的 3D 检测和跟踪结果。表格 V 中的结果表明，在车-路协同任务中，使用生成数据训练的模型在性能上可以与使用真实数据训练的模型相媲美，甚至在某些情况下表现更佳，特别是在 3D 跟踪中，观察到显著

(a) (b)

Fig. 6. 角落案例的可视化：车辆侧遮挡场景的演示。(a) 基础设施视角；(b) 自主车辆视角。

的性能提升。这种优势可能源于生成的数据基本上不缺失标注，并且是根据车辆轨迹标注了精确的边界框。因此，跟踪标注更为平滑和精确，增强了模型的性能。此外，训练过程中以一半的步数结合真实和生成数据进行训练，其性能提升与单视角的训练相似，这表明生成的数据可以作为一种有效的数据增强策略，以增强模型的鲁棒性。图 5 中的定性结果显示，仅使用真实数据训练的模型在遇到 OOD 场景时表现不佳，导致检测性能不佳。这些发现表明，CRUISE 生成的数据引入了更多样化的序列，增强了模型的泛化能力。这些结果展示了 CRUISE 作为一个全面的 V2X 自动驾驶仿真平台的潜力。

V. 讨论

A. 边界情况生成。

CRUISE 不仅增强了 3D 检测和跟踪能力，还能够在驾驶场景中生成关键极端情况。如图 6 所示，从基础设施视角（a）可以看到两辆车，但在自车视角（b），绿色吉普车被白色货车所遮挡。如果自车向左转弯，V2X 通信可以传递这些隐藏信息，从而使自动驾驶的决策更加明智。模拟这种极端情况对于支持 V2X 的自动驾驶至关重要，它提供了一种具有成本效益且可扩展的方法，以提高系统在复杂现实场景中的安全性和鲁棒性。

B. 同步生成

许多现实世界的 V2X 数据集由于路边和车辆侧传感器之间的时间戳不对齐而受到影响，导致数据同步复杂化并降低了可靠性。CRUISE 是第一个能够重建真实世界街景并从自车和基础设施视角生成同步 V2X 数据的仿真框架——没有任何时间偏移。这种精确对齐提高了数据的准确性，并为更可靠的现实世界 V2X 部署铺平了道路。

C. 精确的 3D 框体。

V2X-Seq 的 3D 边界框中的注释错误可能在现实世界部署中引入模型偏差。CRUISE 通过利用街道高斯混合模型改进了前景和背景的解耦，从而优化 3D 边界框注释。它还通过在数据生成过程中进行受控的场景编辑来确保注释的准确性。因此，CRUISE 不仅生成高保真度的合成数据，还显著减少了标注错误——增强模型的可靠性并改善现实世界的性能。

VI. 局限性

尽管 CRUISE 可以生成高度准确和逼真的 V2X 数据，但它仍有一些限制。当自车和基础设施的视点在鸟瞰图 (BEV) 透视中方向相似时，LiDAR 噪声可能会引入伪影，在渲染时导致自车视图中的道路纹理出现在半空中。未来的工作可以探索过滤技术以缓解这个问题。尽管 CRUISE 能够生成点云，但它并不模拟自车和基础设施视角的真实 LiDAR 点云。未来的工作可以结合来自 Occ3D [82] 的可见性掩码方法用于此目的。此外，本工作中使用的 GS 方法在下雨场景中表现不佳，特别是当水滴遮挡了自车摄像头视图时。未来的工作可以探索训练一个受 DeRainGS [83] 启发的扩散模型，以在恶劣天气条件下提高图像质量。

在这项工作中，我们推出了 CRUISE，这是第一个基于动态 GS 的 V2X 驾驶模拟器，具备强大的编辑和逼真的渲染能力，可以生成 V2X 数据。通过在现实世界中收集的 V2X 数据，我们可以重建环境，并在交通流生成器的指导下灵活地将各种车辆放置到场景中。通过渲染技术，我们可以为下游任务生成新的 V2X 数据集。大量实验表明，CRUISE 可以改进自车、基础设施以及合作视图下的 3D 检测任务，以及合作 3D 跟踪。此外，CRUISE 支持生成多样化的极端案例和高质量样本，有助于为未来 V2X 自动驾驶系统的训练和评估建立闭环数据生态系统。这些能力支持了更加健壮和可扩展的端到端自动驾驶技术的发展。

References

[1] N. Wang, Y. Chen, L. Xiao, W. Xiao, B. Li, Z. Chen, C. Ye, S. Xu, S. Zhang, Z. Yan *et al.*, "Unifying appearance codes and bilateral grids for driving scene gaussian splatting," *arXiv preprint arXiv:2506.05280*, 2025.

[2] W. Xiao, H. Huang, C. Zhong, Y. Lin, N. Wang, X. Chen, Z. Chen, S. Zhang, S. Yang, P. Merriaux *et al.*, "Simulate any radar: Attribute-controllable radar simulation via waveform parameter embedding," *arXiv preprint arXiv:2506.03134*, 2025.

[3] Z. Xu, B. Li, H.-a. Gao, M. Gao, Y. Chen, M. Liu, C. Yan, H. Zhao, S. Feng, and H. Zhao, "Challenger: Affordable adversarial driving video generation," *arXiv preprint arXiv:2505.15880*, 2025.

[4] J. Guo, Y. Ding, X. Chen, S. Chen, B. Li, Y. Zou, X. Lyu, F. Tan, X. Qi, Z. Li *et al.*, "Dist-4d: Disentangled spatiotemporal diffusion with metric depth for 4d driving scene generation," *arXiv preprint arXiv:2503.15208*, 2025.

[5] C. Li, K. Zhou, T. Liu, Y. Wang, M. Zhuang, H.-a. Gao, B. Jin, and H. Zhao, "Avd2: Accident video diffusion for accident video description," *arXiv preprint arXiv:2502.14801*, 2025.

[6] B. Li, J. Guo, H. Liu, Y. Zou, Y. Ding, X. Chen, H. Zhu, F. Tan, C. Zhang, T. Wang *et al.*, "Uniscene: Unified occupancy-centric driving scene generation," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 11 971–11 981.

[7] H.-a. Gao, M. Gao, J. Li, W. Li, R. Zhi, H. Tang, and H. Zhao, "Scp-diff: Spatial-categorical joint prior for diffusion based semantic image synthesis," in *European Conference on Computer Vision*. Springer, 2024, pp. 37–54.

[8] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," in *ECCV*, 2020.

[9] J. T. Barron, B. Mildenhall, M. Tancik, P. Hedman, R. Martin-Brualla, and P. P. Srinivasan, "Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields," in *ICCV*, 2021.

[10] S. Yuan and H. Zhao, "Slimmerf: Slimmable radiance fields," in *2024 International Conference on 3D Vision (3DV)*. IEEE, 2024, pp. 64–74.

[11] J. Liu, W. Hu, Z. Yang, J. Chen, G. Wang, X. Chen, Y. Cai, H.-a. Gao, and H. Zhao, "Rip-nerf: Anti-aliasing radiance fields with ripmap-encoded platonic solids," in *ACM SIGGRAPH 2024 Conference Papers*, 2024, pp. 1–11.

[12] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, "3d gaussian splatting for real-time radiance field rendering," *TOG*, vol. 42, no. 4, July 2023.

[13] M. Tancik, V. Casser, X. Yan, S. Pradhan, B. Mildenhall, P. P. Srinivasan, J. T. Barron, and H. Kretzschmar, "Block-nerf: Scalable large scene neural view synthesis," in *CVPR*, 2022.

[14] R. Martin-Brualla, N. Radwan, M. S. Sajjadi, J. T. Barron, A. Dosovitskiy, and D. Duckworth, "Nerf in the wild: Neural radiance fields for unconstrained photo collections," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 7210–7219.

[15] K. Rematas, A. Liu, P. P. Srinivasan, J. T. Barron, A. Tagliasacchi, T. Funkhouser, and V. Ferrari, "Urban radiance fields," in *CVPR*, 2022.

[16] J. Guo, N. Deng, X. Li, Y. Bai, B. Shi, C. Wang, C. Ding, D. Wang, and Y. Li, "Streetsurf: Extending multi-view implicit surface reconstruction to street views," *arXiv preprint arXiv:2306.04988*, 2023.

[17] G. Yan, J. Pi, J. Guo, Z. Luo, M. Dou, N. Deng, Q. Huang, D. Fu, L. Wen, P. Cai *et al.*, "Oasim: an open and adaptive simulator based on neural rendering for autonomous driving," *arXiv preprint arXiv:2402.03830*, 2024.

[18] S. Zhang, B. Ye, X. Chen, Y. Chen, Z. Zhang, C. Peng, Y. Shi, and H. Zhao, "Drone-assisted road gaussian splatting with cross-view uncertainty," *arXiv preprint arXiv:2408.15242*, 2024.

[19] J. Yang, B. Ivanovic, O. Litany, X. Weng, S. W. Kim, B. Li, T. Che, D. Xu, S. Fidler, M. Pavone, and Y. Wang, "Emernerf: Emergent spatial-temporal scene decomposition via self-supervision," in *ICLR*, 2024.

[20] H. Turki, J. Y. Zhang, F. Ferroni, and D. Ramanan, "Suds: Scalable urban dynamic scenes," in *CVPR*, 2023.

[21] J. Ost, F. Mannan, N. Thuerey, J. Knodt, and F. Heide, "Neural scene graphs for dynamic scenes," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 2856–2865.

[22] A. Kundu, K. Genova, X. Yin, A. Fathi, C. Pantofaru, L. J. Guibas, A. Tagliasacchi, F. Dellaert, and T. Funkhouser, "Panoptic neural fields: A semantic object-aware neural scene representation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 12 871–12 881.

[23] Z. Yang, Y. Chen, J. Wang, S. Manivasagam, W.-C. Ma, A. J. Yang, and R. Urtasun, "Unisim: A neural closed-loop sensor simulator," in *CVPR*, 2023.

[24] Z. Wu, T. Liu, L. Luo, Z. Zhong, J. Chen, H. Xiao, C. Hou, H. Lou, Y. Chen, R. Yang, Y. Huang, X. Ye, Z. Yan, Y. Shi, Y. Liao, and H. Zhao, "Mars: An instance-aware, modular and realistic simulator for autonomous driving," in *CICAI*, 2023.

[25] A. Tonderski, C. Lindström, G. Hess, W. Ljungbergh, L. Svensson, and C. Petersson, "Neurad: Neural rendering for autonomous driving," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 14 895–14 904.

[26] T. Fischer, L. Porzi, S. R. Bulo, M. Pollefeys, and P. Kontschieder, "Multi-level neural scene graphs for dynamic urban environments," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 21 125–21 135.

[27] X. Zhou, Z. Lin, X. Shan, Y. Wang, D. Sun, and M.-H. Yang, "Drivinggaussian: Composite gaussian splatting for surrounding dynamic autonomous driving scenes," *arXiv preprint arXiv:2312.07920*, 2023.

[28] Y. Yan, H. Lin, C. Zhou, W. Wang, H. Sun, K. Zhan, X. Lang, X. Zhou, and S. Peng, "Street gaussians for modeling dynamic urban scenes," *arXiv preprint arXiv:2401.01339*, 2024.

[29] N. Huang, X. Wei, W. Zheng, P. An, M. Lu, W. Zhan, M. Tomizuka, K. Keutzer, and S. Zhang, "S^3 gaussian: Self-supervised street gaussians for autonomous driving," *arXiv preprint arXiv:2405.20323*, 2024.

[30] L. D'Alfonso, F. Giannini, G. Franzè, G. Fedele, F. Pupo, and G. Fortino, "Autonomous vehicle platoons in urban road networks: A joint distributed reinforcement learning and model predictive control approach," *IEEE/CAA Journal of Automatica Sinica*, vol. 11, no. 1, pp. 141–156, 2024.

[31] G. Yuan, J. Cheng, M. Zhou, S. Cheng, S. Gao, C. Jiang, and A. Abusorrah, "An autonomous vehicle group cooperation model in an urban scene," *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 12, pp. 13 852–13 862, 2023.

[32] T.-H. Wang, S. Manivasagam, M. Liang, B. Yang, W. Zeng, and R. Urtasun, "V2vnet: Vehicle-to-vehicle communication for joint perception and prediction," in *European Conference on Computer Vision*. Springer, 2020, pp. 605–621.

[33] Y. Hu, Y. Lu, R. Xu, W. Xie, S. Chen, and Y. Wang, "Collaboration helps camera overtake lidar in 3D detection," *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[34] H. Yu, Y. Tang, E. Xie, J. Mao, P. Luo, and Z. Nie, "Flow-based feature fusion for vehicle-infrastructure cooperative 3D object detection," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[35] W. Tianhang, C. Guang, C. Kai, L. Zhengfa, Z. Bo, K. Alois, and J. Changjun, "Umc: A unified bandwidth-efficient and multi-resolution based collaborative perception framework," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

[36] C. Qiu, S. Yadav, A. Squicciarini, Q. Yang, S. Fu, J. Zhao, and C. Xu, "Distributed data-sharing consensus in cooperative perception of autonomous vehicles," in *2022 IEEE 42nd International Conference on Distributed Computing Systems (ICDCS)*. IEEE, 2022, pp. 1212–1222.

[37] H. Yu, W. Yang, H. Ruan, Z. Yang, Y. Tang, X. Gao, X. Hao, Y. Shi, Y. Pan, N. Sun, J. Song, J. Yuan, P. Luo, and Z. Nie, "V2x-seq: A large-scale sequential dataset for vehicle-infrastructure cooperative perception and forecasting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[38] R. Xu, Z. Tu, H. Xiang, W. Shao, B. Zhou, and J. Ma, "Cobevt: Cooperative bird's eye view semantic segmentation with sparse transformers," *arXiv preprint arXiv:2207.02202*, 2022.

[39] Y. Jiang, E. Javanmard, J. Nakazato, M. Tsukada, and H. Esaki, "Roadside lidar assisted cooperative localization for connected autonomous vehicles," *ACM Intelligent Computing and its Emerging Applications (ICEA)*, 2023.

[40] J. Dong, Q. Chen, D. Qu, H. Lu, A. Ganlath, Q. Yang, S. Chen, and S. Labi, "Lidar-based cooperative relative localization," in *2023 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2023, pp. 1–8.

[41] H. Ruan, H. Yu, W. Yang, S. Fan, Y. Tang, and Z. Nie, "Learning cooperative trajectory representations for motion forecasting," *arXiv preprint arXiv:2311.00371*, 2023.

[42] R. Song, C. Liang, H. Cao, Z. Yan, W. Zimmer, M. Gross, A. Festag, and A. Knoll, "Collaborative semantic occupancy prediction with hybrid feature fusion in connected automated vehicles," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[43] H. Yu, W. Yang, H. Ruan, Z. Yang, Y. Tang, X. Gao, X. Hao, Y. Shi, Y. Pan, N. Sun *et al.*, "V2x-seq: A large-scale sequential dataset for vehicle-infrastructure cooperative perception and forecasting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 5486–5495.

[44] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, "Carla: An open urban driving simulator," in *CoRL*, 2017.

[45] S. Shah, D. Dey, C. Lovett, and A. Kapoor, "Airsim: High-fidelity visual and physical simulation for autonomous vehicles," in *Field and Service Robotics: Results of the 11th International Conference*. Springer, 2018, pp. 621–635.

[46] S. Manivasagam, S. Wang, K. Wong, W. Zeng, M. Sazanovich, S. Tan, B. Yang, W.-C. Ma, and R. Urtasun, "Lidarsim: Realistic lidar simulation by leveraging the real world," in *CVPR*, 2020.

[47] Z. Yang, Y. Chai, D. Anguelov, Y. Zhou, P. Sun, D. Erhan, S. Rafferty, and H. Kretzschmar, "Surfelgan: Synthesizing realistic sensor data for autonomous driving," in *CVPR*, 2020.

[48] J. Fang, D. Zhou, F. Yan, T. Zhao, F. Zhang, Y. Ma, L. Wang, and R. Yang, "Augmented lidar simulator for autonomous driving," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 1931–1938, 2020.

[49] Z. Yang, S. Manivasagam, Y. Chen, J. Wang, R. Hu, and R. Urtasun, "Reconstructing objects in-the-wild for realistic sensor simulation," in *ICRA*, 2023.

[50] Y. Chen, F. Rong, S. Duggal, S. Wang, X. Yan, S. Manivasagam, S. Xue, E. Yumer, and R. Urtasun, "Geosim: Re-

alistic video simulation via geometry-aware composition for self-driving," in *CVPR*, 2021.
[51] J. Wang, S. Manivasagam, Y. Chen, Z. Yang, I. A. Bârsan, A. J. Yang, W.-C. Ma, and R. Urtasun, "Cadsim: Robust and scalable in-the-wild 3d reconstruction for controllable sensor simulation," in *CoRL*, 2022.
[52] H. Turki, D. Ramanan, and M. Satyanarayanan, "Mega-nerf: Scalable construction of large-scale nerfs for virtual fly-throughs," in *CVPR*, 2022.
[53] J. Ost, F. Mannan, N. Thuerey, J. Knodt, and F. Heide, "Neural scene graphs for dynamic scenes," in *CVPR*, 2021.
[54] A. Kundu, K. Genova, X. Yin, A. Fathi, C. Pantofaru, L. Guibas, A. Tagliasacchi, F. Dellaert, and T. Funkhouser, "Panoptic Neural Fields: A Semantic Object-Aware Neural Scene Representation," in *CVPR*, 2022.
[55] Z. Xie, J. Zhang, W. Li, F. Zhang, and L. Zhang, "S-nerf: Neural radiance fields for street views," in *ICLR*, 2023.
[56] A. Tonderski, C. Lindström, G. Hess, W. Ljungbergh, L. Svensson, and C. Petersson, "Neurad: Neural rendering for autonomous driving," in *CVPR*, 2024.
[57] Y. Chen, C. Gu, J. Jiang, X. Zhu, and L. Zhang, "Periodic vibration gaussian: Dynamic urban scene reconstruction and real-time rendering," *arXiv:2311.18561*, 2023.
[58] Y.-C. Liu, J. Tian, C.-Y. Ma, N. Glaser, C.-W. Kuo, and Z. Kira, "Who2com: Collaborative perception via learnable handshake communication," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 6876–6883.
[59] Y.-C. Liu, J. Tian, N. Glaser, and Z. Kira, "When2com: Multi-agent perception via communication graph grouping," in *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, 2020, pp. 4106–4115.
[60] R. Xu, H. Xiang, X. Xia, X. Han, J. Liu, and J. Ma, "Opv2v: An open benchmark dataset and fusion pipeline for perception with vehicle-to-vehicle communication," *2022 International Conference on Robotics and Automation (ICRA)*, pp. 2583–2589, 2021.
[61] Y. Li, S. Ren, P. Wu, S. Chen, C. Feng, and W. Zhang, "Learning distilled collaboration graph for multi-agent perception," *Advances in Neural Information Processing Systems*, vol. 34, pp. 29 541–29 552, 2021.
[62] Y. Li, D. Ma, Z. An, Z. Wang, Y. Zhong, S. Chen, and C. Feng, "V2x-sim: Multi-agent collaborative perception dataset and benchmark for autonomous driving," *IEEE Robotics and Automation Letters*, 2022.
[63] R. Xu, H. Xiang, Z. Tu, X. Xia, M.-H. Yang, and J. Ma, "V2x-vit: Vehicle-to-everything cooperative perception with vision transformer," in *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXIX*. Springer, 2022, pp. 107–124.
[64] H. Yu, Y. Luo, M. Shu, Y. Huo, Z. Yang, Y. Shi, Z. Guo, H. Li, X. Hu, J. Yuan *et al.*, "DAIR-V2X: A large-scale dataset for vehicle-infrastructure cooperative 3d object detection," *In Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition (CVPR)*, 2022.
[65] R. Xu, Z. Tu, H. Xiang, W. Shao, B. Zhou, and J. Ma, "CoBEVT: Cooperative bird's eye view semantic segmentation with sparse transformers," *CoRL*, 2022.
[66] Y. Hu, S. Fang, Z. Lei, Y. Zhong, and S. Chen, "Where2comm: Communication-efficient collaborative perception via spatial confidence maps," *Advances in Neural Information Processing Systems*, 2022.
[67] Y. Lu, Q. Li, B. Liu, M. Dianat, C. Feng, S. Chen, and Y. Wang, "Robust collaborative 3d object detection in presence of pose errors," *IEEE International Conference on Robotics and Automation (ICRA)*, 2023.
[68] Y. Li, J. Zhang, D. Ma, Y. Wang, and C. Feng, "Multi-robot scene completion: Towards task-agnostic collaborative perception," in *Conference on Robot Learning*, 2022.
[69] P. Gao, R. Guo, H. Lu, and H. Zhang, "Regularized graph matching for correspondence identification under uncertainty in collaborative perception," *Robotics: Science and Systems XVI*, 2020.
[70] Y. Hu, X. Pang, X. Qin, Y. C. Eldar, S. Chen, P. Zhang, and W. Zhang, "Pragmatic communication in multi-agent collaborative perception," *arXiv preprint arXiv:2401.12694*, 2024.
[71] G. Liu, Y. Hu, C. Xu, W. Mao, J. Ge, Z. Huang, Y. Lu, Y. Xu, J. Xia, Y. Wang *et al.*, "Towards collaborative autonomous driving: Simulation platform and end-to-end system," *arXiv preprint arXiv:2404.09496*, 2024.
[72] H. Chen, C. Li, and G. H. Lee, "Neusg: Neural implicit surface reconstruction with 3d gaussian splatting guidance," *arXiv preprint arXiv:2312.00846*, 2023.
[73] Z. Yu, T. Sattler, and A. Geiger, "Gaussian opacity fields: Efficient adaptive surface reconstruction in unbounded scenes," *ACM Transactions on Graphics*, 2024.
[74] C. Ye, L. Qiu, X. Gu, Q. Zuo, Y. Wu, Z. Dong, L. Bo, Y. Xiu, and X. Han, "Stablenormal: Reducing diffusion variance for stable and sharp normal," *ACM Transactions on Graphics (TOG)*, 2024.
[75] T. Ren, S. Liu, A. Zeng, J. Lin, K. Li, H. Cao, J. Chen, X. Huang, Y. Chen, F. Yan, Z. Zeng, H. Zhang, F. Li, J. Yang, H. Li, Q. Jiang, and L. Zhang, "Grounded sam: Assembling open-world models for diverse visual tasks," 2024.
[76] X. Li, W. Zhang, J. Pang, K. Chen, G. Cheng, Y. Tong, and C. C. Loy, "Video k-net: A simple, strong, and unified baseline for video segmentation," in *CVPR*, 2022.
[77] J. Xiang, Z. Lv, S. Xu, Y. Deng, R. Wang, B. Zhang, D. Chen, X. Tong, and J. Yang, "Structured 3d latents for scalable and versatile 3d generation," *arXiv preprint arXiv:2412.01506*, 2024.
[78] Z. Li, J. Jia, and Y. Shi, "Monolss: Learnable sample selection for monocular 3d detection," in *2024 International Conference on 3D Vision (3DV)*. IEEE, 2024, pp. 1125–1135.
[79] L. Yang, K. Yu, T. Tang, J. Li, K. Yuan, L. Wang, X. Zhang, and P. Chen, "Bevheight: A robust framework for vision-based roadside 3d object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 21 611–21 620.
[80] D. Rukhovich, A. Vorontsova, and A. Konushin, "Imvoxelnet: Image to voxels projection for monocular and multi-view general-purpose 3d object detection," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 2397–2406.
[81] H. Zhou, J. Shao, L. Xu, D. Bai, W. Qiu, B. Liu, Y. Wang, A. Geiger, and Y. Liao, "Hugs: Holistic urban 3d scene understanding via gaussian splatting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 21 336–21 345.
[82] X. Tian, T. Jiang, L. Yun, Y. Mao, H. Yang, Y. Wang, Y. Wang, and H. Zhao, "Occ3d: A large-scale 3d occupancy prediction benchmark for autonomous driving," *Advances in Neural Information Processing Systems*, vol. 36, pp. 64 318–64 330, 2023.
[83] S. Liu, X. Chen, H. Chen, Q. Xu, and M. Li, "Deraings: Gaussian splatting for enhanced scene reconstruction in rainy," *arXiv e-prints*, pp. arXiv–2408, 2024.