基于扩散的二维地图视觉定位通过 BEV 条件下的 GPS 去噪

Li Gao[1]*, Hongyang Sun*, Liu Liu[1]*, Yunhao Li, Yang Cai[1]

***Abstract*— 精确的视觉定位对于自主驾驶至关重要，但现有方法面临一个基本困境：虽然高清（HD）地图提供了高精度定位参考，但其昂贵的构建和维护成本阻碍了扩展性，这促使研究转向标准定义（SD）地图如 OpenStreetMap。当前基于 SD 地图的方法主要集中在图像与地图之间的鸟瞰图(BEV)匹配上，忽略了普遍存在但有噪声的 GPS 信号。虽然 GPS 易于获取，但在城市环境中它会受到多路径误差的影响。我们提出了 DiffVL，首次将视觉定位重新定义为使用扩散模型进行 GPS 去噪任务的框架。我们的关键见解是，当条件基于视觉 BEV 特征和 SD 地图时，嘈杂的 GPS 轨迹隐式编码了真实姿态分布，这可以通过迭代扩散细化来恢复。与之前的 BEV 匹配方法（如 OrienterNet）或基于变换器的注册方法不同，DiffVL 通过联合建模 GPS、SD 地图和视觉信号学习逆转 GPS 噪声扰动，在不依赖高清地图的情况下实现亚米级精度。在多个数据集上的实验表明，我们的方法相比 BEV 匹配基线达到了最先进的准确性。至关重要的是，我们的工作证明了扩散模型可以通过将嘈杂的 GPS 视为生成先验来实现可扩展定位——这标志着从传统的基于匹配的方法向新范式的转变。代码和模型将开源。**

I. 介绍

视觉定位是自主驾驶 [1], [2]、增强现实和机器人等领域的一项关键技术，其中精确可靠的姿态估计对于安全导航 [3] 和决策 [4] 至关重要。核心任务包括从视觉图像中相对于 2D 地图估算一个 3 自由度的姿态（位置和方向）。为了满足自主系统的严格要求，传统方法 [5] 严重依赖高清地图。然而，创建、标注和频繁更新这些地图的高昂成本极大地限制了它们的可扩展性和广泛应用，成为在全球范围内部署自主技术的重要瓶颈。

作为回应，最近的研究转向了使用低成本且全球可用的标准定义（SD）地图进行定位，例如 OpenStreetMap[6]。代表性工作 [7], [8] 利用深度学习 [9], [10], [11], [12] 通过将从输入图像导出的鸟瞰视图（BEV）表示与 SD 地图对齐来推断 3 自由度姿态。虽然这些方法显示出了希望，但它们容易受到视觉重复区域中的感知混叠等挑战的影响，并且通常忽略了关键而普遍的信息来源：噪声 GPS 数据。这种遗漏本质上限制了其定位精度和鲁棒性的上限，特别是在具有挑战性的情况下。

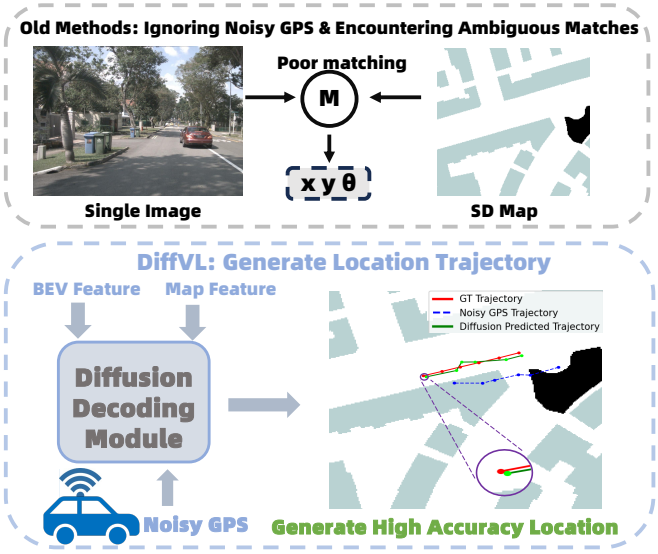

Fig. 1. 提出的 DiffVL 概述。大多数现有的基于 SD 地图的定位方法依赖于从鸟瞰图（BEV）特征和地图元素之间进行详尽的几何匹配来计算姿态。相比之下，我们的方法从根本上将视觉定位重新定义为一个生成建模任务。

近年来，扩散模型 [13], [14], [15], [16], [17] 已经成为强大的生成技术，通过学习逆转渐进的噪声过程实现了突破性的性能。这些技术已经在具身导航 [18] 和自动驾驶 [19] 中成功应用，用于从运动锚点预测未来的轨迹。受这些进展的启发，我们提出了视觉定位的新范式：我们将它重新定义为 GPS 轨迹的条件去噪过程。我们的主要见解是，通常被认为不可靠的噪声 GPS 信号实际上编码了真实的姿态分布，并且可以通过基于扩散的去噪转换成精确的定位。具体来说，我们将传统的图像到地图匹配任务 [20], [21], [22] 转换为一个条件生成问题，在这个问题中，扩散模型学习在视觉观测条件下逆转 GPS 轨迹中的噪声污染。

*Equal Contribution
1Alibaba Amap

我们提出了 DiffVL，一个新颖的基于扩散的框架，该框架协同集成了顺序 GPS 信号和视觉线索。为了赋予扩散模型几何和语义意识，我们引入了双重目标训练策略。主要轨迹细化损失（$\mathcal{L}_{\text{diff}}$）确保去噪轨迹在动力学和时间上的一致性。同时，通过将鸟瞰视图（BEV）视觉特征与地图元素对齐计算的辅助定位先验损失（$\mathcal{L}_{\text{loc}}$），提供了强大的几何正则化。这种双重损失迫使共享特征编码器学习既具有视觉区分力又在地理坐标系统中具有一致性的表示。通过联合优化，我们的模型实现了基于运动预测和基于外观匹配之间的稳健平衡，能够从噪声 GPS 中进行高精度姿态估计，并显著增强定位鲁棒性。

我们的主要贡献总结如下：

- 我们引入了 DiffVL，这是一种新颖的视觉定位范式，据我们所知，它是第一个成功并主要应用扩散模型来去噪嘈杂的 GPS 轨迹以完成此任务的方法。它将定位问题重新定义为条件生成问题。
- 我们重新定义了噪声 GPS 信号在视觉定位中的作用。先前的研究往往忽略了这些信号或对它们进行了过滤，而我们首次将它们视为真实姿态分布的“噪声观测”，为利用生成模型恢复高精度姿态提供了方法论基础。
- 我们设计了一个双目标训练框架，其中轨迹优化损失和定位先验损失协同工作。这确保模型不仅能够恢复连贯的轨迹，还能从视觉和地图数据中学习到强大的、几何一致的特征表示。
- 我们在多个大规模自动驾驶数据集（KITTI[23]、nuScenes[24] 和 MGL[7]）上进行了严格的综合评估。我们的实验结果表明，DiffVL 在所有数据集上的表现显著优于现有方法，达到了最先进的性能水平。我们计划发布我们的代码和模型，以促进未来的研发并为社区做贡献。

II. 相关工作

A. 基于地图的视觉定位

基于地图的视觉定位是机器人学 [3] 和自主系统 [25], [26], [27] 中的一个长期研究任务。传统方法依赖于将传感器数据与精心构建的 3D 地图 [28], [29], [30] 进行匹配。这些地图通常由激光雷达传感器的密集点云组成，或通过多视角图像的运动结构（SfM）重建 [31]。它们通过查询图像和 3D 模型 [32], [33] 之间手工制作或学习到的局部特征对齐进行高精度姿态估计，或者通过点云的直接几何配准，通常使用迭代最近点算法 (ICP) 的变体 [34]。然而，这些高清地图存在显著的实际缺点。它们的创建需要专门的车辆和大量的勘测工作，维护是一个持续且昂贵的努力，并且它们的大内存占用是车载存储和大规模部署的主要障碍。

为了克服这些可扩展性问题，近期的研究转向了标准定义（SD）地图。这些地图通常源自众包数据如 OpenStreetMap (OSM)[6]，轻量级、全球可用且语义丰富，使其成为极具吸引力的替代方案。一项突出的工作涉及从单目相机生成神经鸟瞰图（BEV）[35], [36] 表示形式，然后将其与相应的栅格化地图瓷砖 [7], [8] 对齐。这些方法有效地解决了地面视角和俯视地图之间的跨视图匹配问题。尽管它们表现良好，但这些方法仍面临两个主要挑战。首先，必须弥合渲染地图语义与复杂现实世界视觉特征之间显著的领域差距。其次，几乎普遍忽视了一个重要且易于获取的信号：嘈杂的 GPS 数据。这一关键信息源的缺失，通常是因为难以处理其固有的噪声，从根本上限制了它们的鲁棒性和准确性，特别是在模糊环境中的表现。

B. 扩散模型

近年来，去噪扩散概率模型（DDPMs）[13], [14], [37] 作为生成式 AI 领域的一项颠覆性技术出现，在高保真图像和视频生成 [38], [39], [40]、机器人策略学习 [41] 和运动预测 [42] 等不同领域取得了最先进的成果。其核心思想是学习逆转一个逐步向数据中添加噪声的固定马尔可夫链，使模型能够通过从纯高斯噪声开始逐渐去噪来生成新的样本。它们捕捉复杂多模态分布的能力使它们特别适用于机器人和自主系统。例如，Diffusion Policy[41], [43] 在学习复杂操作任务的多模态动作分布方面表现出极高的有效性。在运动预测中，像 MotionDiffuser[42], [44] 这样的工作利用条件扩散模型生成动力学一致且具备社会意识的多智能体轨迹。

尽管展示了其潜力，扩散模型在视觉定位中的应用仍是一个未开发的领域。受这些开创性工作的启发 [4], [45], [46]，我们首次将扩散模型范式引入到这个任务中。我们的关键见解是重新定义问题：不是将定位视为一个确定性的匹配问题 [7], [8]，而是通过将其表述为一个条件去噪任务来拥抱传感器数据内在的不确定性。在我们的 DiffVL 框架中，扩散模型学习从噪声原始 GPS 序列中恢复出一条动力学平滑的轨迹。关键

的是，这个去噪过程并不是孤立进行的；它是根据从相机图像和 SD 地图 [6] 中提取出来的丰富多模态特征智能条件化的，提供了必要的环境背景。为了确保这些条件特征在几何上是基础且语义上有意义的，我们采用了一种双重目标训练策略。主要轨迹优化损失由一个辅助 BEV 地图匹配损失补充，该损失作为强大的几何正则化器，迫使共享编码器学习具有空间意识的表示。这种方法直接将扩散模型生成能力与视觉-地图匹配的判别任务在特征级别上结合起来，显著增强了定位的鲁棒性和准确性。

III. 方法

A. 问题公式化

给定一个单个前视图像 $\mathcal{I}$ 和一段带有噪声的 GPS 测量历史序列 $\boldsymbol{P}^{\text{gps}} = \{\boldsymbol{p}_t^{\text{gps}}\}_{t=1}^{T} = \{x_t^{\text{gps}}, y_t^{\text{gps}}\}_{t=1}^{T}$，其中 T 是时间范围，而 $\boldsymbol{p}_t^{\text{gps}} = (x_t^{\text{gps}}, y_t^{\text{gps}})$ 表示在时间步长 t 的东-北-上（ENU）坐标，以及一个 SD 地图 M 代表包含轨迹 $\boldsymbol{P}^{\text{gps}}$ 的本地地图瓦片。我们的任务是估计一个 3 自由度的姿态 $\hat{\boldsymbol{p}} = (x, y, \theta) \in \mathbb{R}^3$，其中 (x, y) 表示 ENU 位置，而 $\theta \in (-\pi, \pi]$ 表示围绕垂直轴 $\boldsymbol{z}$ 的航向角。

姿态估计过程被形式化为一个条件扩散模型：

$$\hat{\boldsymbol{p}} = \mathcal{A}_\theta \left(\{\boldsymbol{p}_t^{\text{gps}}\}_{i=1}^{T} \mid \boldsymbol{z} \right) \tag{1}$$

其中：

- $\mathcal{A}_{\boldsymbol{\theta}}(\cdot)$ 表示扩散模型头部
- $\boldsymbol{\theta}$ 表示可训练参数
- $\boldsymbol{z} = f(\mathcal{I}, M)$ 是条件潜向量，编码图像映射特征

B. 概述

DiffVL 框架包含四个核心模块，其架构如图 2所示。

图像编码器: 通过前视图像执行环境感知，提取多尺度视觉特征。利用深度估计分析场景几何结构，并结合视角转换将透视表示转化为几何一致的 BEV（鸟瞰图）特征图。这一过程有效地保留了空间关系和语义内容，同时与真实世界的坐标系统建立了对应关系，从而实现了从头顶视角对周围环境的稳健理解。

地图编码器: 利用 OpenStreetMap[6] 通过基于网格的处理构建栅格化地图表示，编码关于环境的关键先验知识。

扩散引导生成器: 实现环境感知与先验地图知识之间的跨模态融合通过注意力机制。生成条件引导嵌入，将视觉观察与地图语义结合，构建全局上下文表示以指导后续的扩散去噪过程。

扩散头: 通过迭代扩散基础校正实现概率定位细化。将姿态估计表述为一个逐步降噪问题，其中初始位置假设经历由多模态上下文特征引导的多阶段优化。通过联合优化位置和方向参数，它系统地减少了定位不确定性，最终在标准定义（SD）地图上实现了米级精度。

系统架构和多任务损失函数设计的详细参数配置将在后续章节中全面呈现。

C. 图像编码模块

此模块从单个前视图图像 $\mathcal{I} \in \mathbb{R}^{H \times W \times 3}$ 中提取结构化的环境特征，并将其转换为鸟瞰图（BEV）表示。实现由三个关键阶段组成：

多尺度特征提取: 使用 ResNet-101[47] 主干网络来提取透视图 (PV) 中的多尺度特征金字塔：

$$\left\{ \boldsymbol{F}_{\text{pv}}^1, \boldsymbol{F}_{\text{pv}}^2, \boldsymbol{F}_{\text{pv}}^3, \boldsymbol{F}_{\text{pv}}^4 \right\} = \Phi_{\text{img}}(\mathcal{I}) \tag{2}$$

其中，$\boldsymbol{F}_{\text{pv}}^i \in \mathbb{R}^{H_i \times W_i \times C_i}$ 表示第 i 层的特征图。通过跳过连接在不同层之间融合多分辨率信息，增强了模型感知复杂场景的能力。

深度概率分布预测: 为了提高几何建模的准确性，引入了一个并行深度估计分支来预测从特征金字塔中得到的每个像素的深度分布：

$$\mathcal{D} = \Psi_{\text{depth}} \left(\bigoplus_{i=1}^{4} \text{UpSample}(\boldsymbol{F}_{\text{pv}}^i) \right) \tag{3}$$

这里，$\mathcal{D} \in \mathbb{R}^{H \times W \times D}$ 是深度分布张量，其中 D 表示离散深度区间数量，Ψ_{depth} 是深度预测子网络，而 $\bigoplus$ 代表特征拼接。

可微视图变换: 使用类似于 OrienterNet[7] 和 LSS[48] 的极坐标-笛卡尔双重投影：

$$\boldsymbol{F}_{\text{bev}} = \mathcal{P}_{\text{cart}} \left(\mathcal{P}_{\text{polar}} \left(\boldsymbol{F}_{\text{pv}}, \mathcal{D} \right), \mathcal{C} \right) \tag{4}$$

其中：

- $\mathcal{P}_{\text{polar}}$：带有尺度先验的极坐标投影 $\mathcal{D}$
- $\mathcal{P}_{\text{cart}}$：极坐标到直角坐标的变换
- $\mathcal{C}$：相机内外参数

生成 BEV 特征 $\boldsymbol{F}_{\text{bev}} \in \mathbb{R}^{B \times C \times H_b \times W_b}$。

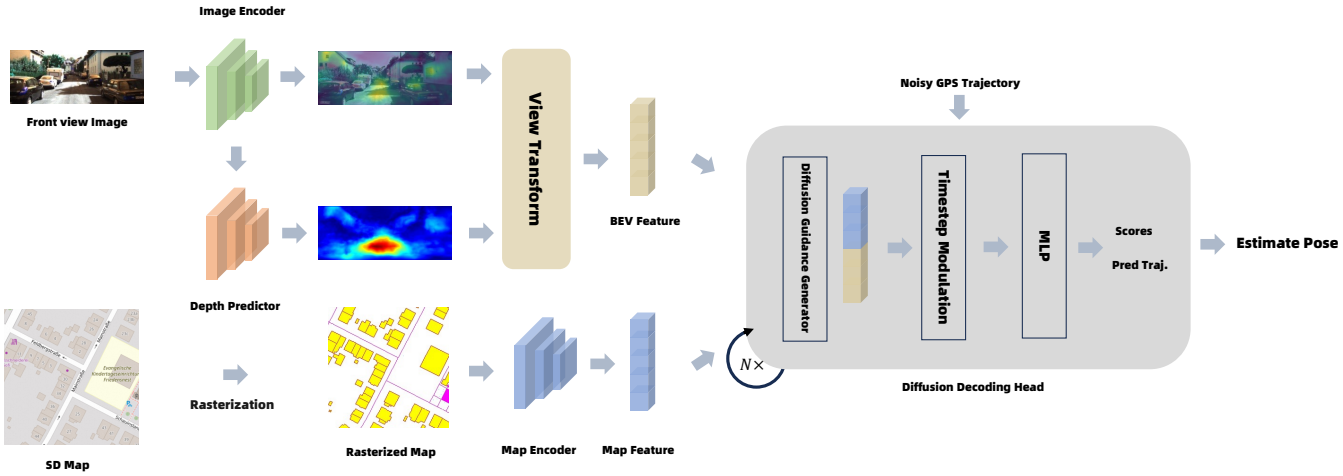

Fig. 2. DiffVL 的架构。作为首个基于扩散模型的视觉定位框架，我们的系统开创了一种从传统匹配方法向生成式表述转变的新范式。该架构接受三个关键输入：(i) 一张单目前视 RGB 图像，捕捉即时场景上下文；(ii) 提供结构先验的标准定义 (SD) 地图数据；以及 (iii) 提供粗略位置线索的噪声 GPS 轨迹。我们的创新核心在于图像编码模块将视角转换为几何一致的鸟瞰图 (BEV) 特征，而地图编码模块则从 SD 地图中提取拓扑表示。这些互补特征经过多模态融合生成用于我们新颖扩散模块的条件特征——该模块是重新定义视觉定位为有条件生成任务的核心组件。通过迭代逆扩散步骤，此模块逐步去噪被污染的 GPS 输入，将不可靠的传感器测量转化为精确的 3 自由度姿态估计。这种生成方法标志着首次成功将扩散模型应用于视觉定位，确立了一种新的轨迹优化范式。

D. 映射编码模块

本模块通过四个处理阶段从开源地图数据构建结构化环境先验。

地图数据采集: 根据历史 GPS 轨迹 $\boldsymbol{P}^{\text{gps}}$ 的空间分布计算出一个边界框，并从 OpenStreetMap (OSM) 中检索该感兴趣区域对应的地图矢量数据。这确保了地图信息与车辆当前位置之间的空间一致性，形成了全局先验构建的基础。

语义光栅化: 向量地图数据被转换为一个三通道的 RGB 图像：通道 1 编码道路网络（包括高速公路、主干道和地方道路），通道 2 代表建筑物轮廓，通道 3 编码自然特征（如植被、水体和地形）。栅格化分辨率与 BEV 特征 (0.5m/pixel) 保持一致，从而得到一张地图图像 $\boldsymbol{M}_{\text{rgb}} \in \mathbb{R}^{X\times Y\times 3}$。这种方法借鉴了之前在 [7] 中的建模技术，旨在提高系统对静态环境结构的理解。

层次特征提取: 使用 VGG16[49] 架构从栅格化地图中提取特征：

$$\boldsymbol{F}_{\text{map}} = \Psi_{\text{MapEnc}}(\boldsymbol{M}_{\text{rgb}}) \tag{5}$$

所得的 $\boldsymbol{F}_{\text{map}}$ 是一个压缩的特征图，捕捉到了诸如道路拓扑和可穿越性约束等关键先验信息，强调了对结构化地图信息的有效表示学习。

E. 扩散指导生成器

该模块实现了视觉感知与地图先验的深度融合，以生成用于扩散模型的多尺度上下文特征，包括多模态特征融合和扩散解码头。

多模态特征融合: BEV 和地图特征进行维度对齐和上下文融合：

$$\boldsymbol{F}_{\text{cond}} = \Gamma\left(\phi_{\text{bev}}(\boldsymbol{F}_{\text{bev}}), \psi_{\text{map}}(\boldsymbol{F}_{\text{map}})\right) \tag{6}$$

其中：

- ψ_{map}：地图特征压缩和空间重构
- ϕ_{bev}：BEV 特征投影和空间对齐
- Γ：跨模态融合算子

生成一个综合视觉感知和语义先验的统一表示 $\boldsymbol{F}_{\text{cond}} \in \mathbb{R}^{B\times C_f\times H\times W}$。

扩散解码头: 此模块通过扩散建模实现条件轨迹生成和姿态优化。首先，历史轨迹通过线性扩散过程进行归一化和注入噪声。我们将历史 GPS 轨迹 $\boldsymbol{P}^{\text{gps}} \in \mathbb{R}^{T\times 3}$ 归一化到 $[-1, 1]$ 范围内：

$$\boldsymbol{P}_{\text{norm}} = \text{norm}_{\text{odo}}(\boldsymbol{P}^{\text{gps}}) \tag{7}$$

然后注入高斯噪声：

$$\boldsymbol{P}_t = \sqrt{\bar{\alpha}_t}\boldsymbol{P}_{\text{norm}} + \sqrt{1-\bar{\alpha}_t}\epsilon, \quad \epsilon \sim \mathcal{N}(0, \boldsymbol{I}) \tag{8}$$

Fig. 3. 我们在 KITTI 数据集上的定位结果。在这些可视化中，红色轨迹代表数据集中的地面实况（GT）GPS 轨迹，而蓝色轨迹是我们合成生成的带噪声 GPS 轨迹。给定带噪声的蓝色轨迹和单个图像作为输入，我们的方法产生精炼后的绿色“生成位置”轨迹。

受 DiffusionDrive[50] 的启发，在训练过程中，扩散解码器以 N_{anchor} 噪声轨迹 $\{\boldsymbol{p}_t^{\text{gps}}\}_{k=1}^{\text{anchor}}$ 为输入，并预测分类分数 $\{\hat{s}_k\}_{k=1}^{N_{\text{anchor}}}$ 和去噪轨迹 $\hat{\boldsymbol{p}}_{k=1}^{N_{\text{anchor}}}$：

$$\{\hat{s}_k, \hat{\boldsymbol{p}}_k\}_{k=1}^{N_{\text{anchor}}} = f_\theta\left(\{\boldsymbol{p}^k\}_{k=1}^{N_{\text{anchor}}}, \boldsymbol{F}_{\text{cond}}\right) \tag{9}$$

其中 $\boldsymbol{F}_{\text{cond}}$ 表示条件信息。我们将最接近锚点的嘈杂轨迹分配给真实轨迹 τ_{gt} 作为正样本（$y_k = 1$），其余作为负样本（$y_k = 0$）。

F. 损失函数的构建

总体训练目标结合了轨迹细化和定位损失：

$$\mathcal{L}_{\text{total}} = \underbrace{\mathcal{L}_{\text{diff}}}_{\text{Trajectory Refinement}} + \alpha \underbrace{\mathcal{L}_{\text{loc}}}_{\text{Localization Prior}} \tag{10}$$

其中 α 平衡了定位损失的贡献。

1) 轨迹优化损失： 通过多模态优化引导基于扩散的轨迹生成：

$$\mathcal{L}_{\text{diff}} = \sum_{k=1}^{N_{\text{anchor}}} \left[y_k \left\| \hat{\boldsymbol{p}}_k - \tau_{\text{gt}} \right\|_1 + \lambda \mathcal{L}_{\text{BCE}}(\hat{s}_k, y_k) \right] \tag{11}$$

与 $y_k = \mathbb{I}\left[k = \underset{j}{\operatorname{argmin}} \left\|\boldsymbol{p}_k - \tau_{\mathrm{gt}}\right\|_2\right]$ 结合，强制模式选择围绕真实锚点。

2) 局部化损失: 通过 BEV-map 匹配提供空间正则化：

$$\mathcal{L}_{\mathrm{loc}} = -\log \boldsymbol{P}(\tau_{\mathrm{gt}} \mid \boldsymbol{S}) \tag{12}$$

$$\boldsymbol{S} = \mathrm{Match}(\boldsymbol{F}_{\mathrm{bev}}, \boldsymbol{F}_{\mathrm{map}}) \tag{13}$$

实现带标签平滑的负对数似然，以考虑标注不确定性。这种多目标设计使精确轨迹生成通过 $\mathcal{L}_{\mathrm{diff}}$ 和几何一致性通过 $\mathcal{L}_{\mathrm{loc}}$ 实现。

IV. 实验

A. 实现细节

输入表示。我们的方法遵循先前工作的设置以进行公平比较。我们使用单个正视图图像作为视觉输入。地图输入是从栅格化导航图中提取的 128 米 ×128 米瓷砖，中心位于自我车辆嘈杂的 GPS 位置。该地图瓷砖的分辨率为每像素 0.5 米（mpp）。

训练设置。为了模拟现实世界中的 GPS 误差并训练一个稳健的模型，我们对每个训练样本的真实姿态应用随机扰动。这些扰动是从旋转范围 $\theta \in [-30°, 30°]$ 和平移范围 $t \in [-30\,\mathrm{m}, 30\,\mathrm{m}]$ 中均匀采样的。我们的模型使用 AdamW 优化器以学习率 1×10^{-4} 和权重衰减 1×10^{-2} 进行端到端训练。实现基于 PyTorch。我们在单个 NVIDIA RTX 2080 GPU 上训练 DiffVL 模型。

B. 数据集

我们在三个大规模且多样化的数据集上进行了广泛的实验：KITTI[23]、MGL[7] 和 nuScenes[24]。下面，我们简要介绍每个数据集，并突出它们所带来的具体挑战。

KITTI. KITTI 数据集 [23] 是自动驾驶领域广泛使用的权威基准。该数据集收集于德国卡尔斯鲁厄及其周边地区，涵盖了多样的真实世界交通场景，从密集的市区街道和乡村道路到高速公路。其包含动态物体和各种光照条件下的具有挑战性的序列使其成为评估定位稳健性理想的工具。KITTI 提供了精确同步和校准过的多模态传感器数据，并伴有由高精度 GPS/IMU 系统生成的真实姿态数据。我们的实验严格遵循官方的训练和测试分割。

MGL. MGL 数据集由 OrienterNet[7] 引入，以促进大规模视觉地理定位的研究。该数据集从 Mapillary 平台收集，包含来自欧洲和美国 12 座城市的超过 760,000 张图像。由于其极大的多样性，此数据集特别具有挑战性；图像由各种设备（手持、安装在汽车和自行车上）在广泛的天气和光照条件下拍摄。这种多样性考验了模型的泛化能力。所有图像都配有地面实况 (GT) 姿态和 OpenStreetMap (OSM) 数据。截至撰写本文时，来自两个城市（阿姆斯特丹和维尔纽斯）的数据已不再可访问。因此，我们利用剩余 10 座城市的数据显示用于训练和评估。

nuScenes. nuScenes[24] 数据集是一个广泛应用的大规模自动驾驶数据集，以其全面的传感器套件和复杂的都市环境而著称。它包含在波士顿和新加坡捕捉到的 1,000 个驾驶场景，每个场景持续 20 秒。该数据集的特点是密集的交通、复杂的交叉路口以及大量的行人活动，使其成为评估安全关键情况性能的理想测试平台。我们遵循官方的数据划分，使用 750 个序列（约 28,000 帧）作为训练集和 150 个序列（约 6,000 帧）作为验证集进行实验。

C. 定位结果

基蒂数据集的定量分析。我们首先在 KITTI 数据集上评估我们的方法，并将其与包括 OrienterNet[7]、DSM[51]、VIGOR[52] 和 BeyondRetrieval[29] 在内的最新方法进行比较。遵循标准评估协议，我们使用横向召回率@Xm、纵向召回率@Xm 和方向召回率@X° 作为主要指标。如表 1 所示，DiffVL 在所有度量上都显著优于所有基线方法。

大规模 MGL 和 nuScenes 数据集上的性能。为了展示我们方法的可扩展性和泛化能力，我们在 MGL 和 nuScenes 数据集上进行了进一步比较。评价指标是在 1, 2, 5, 10 米距离阈值下的召回准确率 (RA) 以及在 1, 2, 5, 10 度阈值下的方向召回准确率。实验结果汇总于表 2 中，证实了我们方法的优越性。在高度多样化的 MGL 数据集上，DiffVL 的一贯领先表明我们的模型学习到了一种可泛化的表示形式，而不是针对特定城市或相机类型过拟合。在复杂的城区 nuScenes 数据集上，我们方法的强大表现突显了其在密集交通和感知挑战场景中的鲁棒性。

定位结果的可视化。图 3可视化了我们在 KITTI 数据集上的定位结果。在这些可视化中，红色轨迹代

TABLE I

KITTI 数据集上的定量比较。所有指标均为召回率（%），数值越高越好。每个单元格都进行了着色，以指示每列中的最佳性能。

方法	侧向回忆（%）			纵向召回率（%）			方向召回率（%）		
	1m	3m	5m	1m	3m	5m	1°	3°	5°
DSM	10.77	31.37	48.24	3.87	11.73	19.50	3.53	14.09	23.95
VIGOR	17.38	48.20	70.79	4.07	12.52	20.14	-	-	-
BeyondRetrieval	27.82	59.79	72.89	5.75	16.36	26.48	18.42	49.72	71.00
OrienterNet	51.26	84.77	91.81	22.39	46.79	57.81	20.41	52.24	73.53
我们的	**65.95**	**90.08**	**94.67**	**23.51**	**51.72**	**63.03**	**26.00**	**66.07**	**84.27**

TABLE II

MGL 和 nuScenes 数据集上的定量比较。所有指标均为召回准确率（%），数值越高越好。每个数据集组中每列的最佳性能已被突出显示。

数据集	方法	位置召回率（%）				方向回忆率（%）			
		1m	2m	5m	10m	1°	2°	5°	10°
MGL	OrienterNet	10.78	29.88	54.72	67.25	18.98	35.15	63.03	76.63
	我们的	**11.07**	**31.46**	**57.23**	**69.30**	**19.57**	**35.74**	**64.79**	**77.91**
nuScenes	OrienterNet	2.89	6.01	18.57	38.49	9.30	16.86	35.40	55.81
	我们的	**15.70**	**28.83**	**56.74**	**79.20**	**19.32**	**37.46**	**70.41**	**86.91**

TABLE III

轨迹优化模块在 KITTI 数据集上的消融研究。度量是位置召回准确率，数值越高越好。最佳结果已突出显示。

方法	位置召回率			
	1m	2m	5m	10m
w/o Trajectory Refinement	0.0978	0.2871	0.5325	0.6591
我们的	**0.1554**	**0.3530**	**0.5935**	**0.7075**

表数据集中的地面真实（GT）GPS 轨迹，而蓝色轨迹是我们合成生成的带噪声 GPS 轨迹。给定带噪声的蓝色轨迹和单张图像作为输入，我们的方法产生了优化后的绿色“生成位置”轨迹。

如图所示，我们方法输出的轨迹与紫色区域（对应输入图像的位置）的真实轨迹紧密一致，实现了高质量的定位。值得注意的是，对于紫色区域之外的 GPS 点，生成轨迹的准确性略有下降。这是可以预料的，因为我们的推理过程仅利用了来自紫色区域的单个图像；其他 GPS 点对应的视觉信息未知，因此无法对这些点进行准确的视觉定位。然而，这表明在有可用视觉信息的区域内，我们的方法能够实现高质量的视觉定位性能。

D. 消融研究

为了隔离并验证我们核心轨迹优化模块的贡献，我们在 KITTI 数据集上进行了有针对性的消融研究。我们配置了一个基线变体，标记为无轨迹优化，通过移除扩散头及其相关损失（$\mathcal{L}_{\text{diff}}$）。该变体仅依赖于 BEV 地图匹配机制进行定位，类似于传统方法。

结果如表 3 所示，清楚地展示了该模块的关键作用。移除轨迹细化会导致所有位置召回指标的性能显著下降。扩散模型降噪时间 GPS 序列并施加动力学一致先验的能力对于实现高精度结果至关重要。这一消融实验提供了有力证据，证明我们提出的条件去噪范式是 DiffVL 达到业界领先性能的关键驱动因素。

V. 结论

本文介绍了 DiffVL，这是一个开创性的框架，通过将任务重新定义为条件 GPS 去噪问题来引领视觉定位的新范式。我们的核心贡献是将嘈杂的 GPS 信号重新构架为有价值的生成先验，为利用扩散模型恢复高精度姿态提供了方法论基础。我们通过一种双目标训练策略实现这一点，这种策略协同引导模型产生动力学一致的轨迹，同时通过视图到地图的对齐学习几何上可靠的表现形式。在包括 KITTI、MGL 和 nuScenes

在内的大规模数据集上的广泛实验验证了我们的方法，证明 DiffVL 始终达到最先进的性能。

References

[1] I. Yaqoob, L. U. Khan, S. A. Kazmi, M. Imran, N. Guizani, and C. S. Hong, "Autonomous driving cars in smart cities: Recent advances, requirements, and challenges," *IEEE Network*, vol. 34, no. 1, pp. 174–181, 2019.

[2] J. Levinson, J. Askeland, J. Becker, J. Dolson, D. Held, S. Kammel, J. Z. Kolter, D. Langer, O. Pink, V. Pratt, *et al.*, "Towards fully autonomous driving: Systems and algorithms," in *2011 IEEE intelligent vehicles symposium (IV)*. IEEE, 2011, pp. 163–168.

[3] W. Cai, J. Peng, Y. Yang, Y. Zhang, M. Wei, H. Wang, Y. Chen, T. Wang, and J. Pang, "Navdp: Learning sim-to-real navigation diffusion policy with privileged information guidance," *arXiv preprint arXiv:2505.08712*, 2025.

[4] B. Liao, S. Chen, H. Yin, B. Jiang, C. Wang, S. Yan, X. Zhang, X. Li, Y. Zhang, Q. Zhang, *et al.*, "Diffusiondrive: Truncated diffusion model for end-to-end autonomous driving," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 12 037–12 047.

[5] I. A. Barsan, S. Wang, A. Pokrovsky, and R. Urtasun, "Learning to localize using a lidar intensity map," *arXiv preprint arXiv:2012.10902*, 2020.

[6] M. Haklay and P. Weber, "Openstreetmap: User-generated street maps," *IEEE Pervasive computing*, vol. 7, no. 4, pp. 12–18, 2008.

[7] P.-E. Sarlin, D. DeTone, T.-Y. Yang, A. Avetisyan, J. Straub, T. Malisiewicz, S. R. Bulo, R. Newcombe, P. Kontschieder, and V. Balntas, "OrienterNet: Visual Localization in 2D Public Maps with Neural Matching," in *CVPR*, 2023.

[8] Z. Zhou, Z. Qi, L. Cheng, and G. Xiong, "Seglocnet: Multimodal localization network for autonomous driving via bird's-eye-view segmentation," *arXiv preprint arXiv:2502.20077*, 2025.

[9] L. Gao, J. Zhang, L. Zhang, and D. Tao, "Dsp: Dual soft-paste for unsupervised domain adaptive semantic segmentation," in *Proceedings of the 29th ACM international conference on multimedia*, 2021, pp. 2825–2833.

[10] L. Gao, L. Zhang, and Q. Zhang, "Addressing domain gap via content invariant representation for semantic segmentation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 9, 2021, pp. 7528–7536.

[11] L. Gao, D. Nie, B. Li, and X. Ren, "Doubly-fused vit: Fuse information from vision transformer doubly with local representation," in *European Conference on Computer Vision*. Springer, 2022, pp. 744–761.

[12] C. Xing, G. Li, and L. Zhang, "Bsam: Bidirectional scene-aware mixup for unsupervised domain adaptation in semantic segmentation," in *CAAI International Conference on Artificial Intelligence*. Springer, 2022, pp. 54–66.

[13] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[14] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," *arXiv preprint arXiv:2010.02502*, 2020.

[15] M. Jiang, Y. Bai, A. Cornman, C. Davis, X. Huang, H. Jeon, S. Kulshrestha, J. Lambert, S. Li, X. Zhou, *et al.*, "Scenediffuser: Efficient and controllable driving simulation initialization and rollout," *Advances in Neural Information Processing Systems*, vol. 37, pp. 55 729–55 760, 2024.

[16] F.-A. Croitoru, V. Hondru, R. T. Ionescu, and M. Shah, "Diffusion models in vision: A survey," *IEEE transactions on pattern analysis and machine intelligence*, vol. 45, no. 9, pp. 10 850–10 869, 2023.

[17] J. Li, M. Zhang, N. Li, D. Weyns, Z. Jin, and K. Tei, "Generative ai for self-adaptive systems: State of the art and research roadmap," *ACM Transactions on Autonomous and Adaptive Systems*, vol. 19, no. 3, pp. 1–60, 2024.

[18] S. Wang, J. Zhang, M. Li, J. Liu, A. Li, K. Wu, F. Zhong, J. Yu, Z. Zhang, and H. Wang, "Trackvla: Embodied visual tracking in the wild," *arXiv preprint arXiv:2505.23189*, 2025.

[19] X. Jiang, Y. Ma, P. Li, L. Xu, X. Wen, K. Zhan, Z. Xia, P. Jia, X. Lang, and S. Sun, "Transdiffuser: End-to-end trajectory generation with decorrelated multi-modal representation for autonomous driving," *arXiv preprint arXiv:2505.09315*, 2025.

[20] M. A. Quddus, W. Y. Ochieng, and R. B. Noland, "Current map-matching algorithms for transport applications: State-of-the art and future research directions," *Transportation research part c: Emerging technologies*, vol. 15, no. 5, pp. 312–328, 2007.

[21] O. Pink, "Visual map matching and localization using a global feature map," in *2008 IEEE computer society conference on computer vision and pattern recognition workshops*. IEEE, 2008, pp. 1–7.

[22] Z. Huang, S. Qiao, N. Han, C.-a. Yuan, X. Song, and Y. Xiao, "Survey on vehicle map matching techniques," *CAAI Transactions on Intelligence Technology*, vol. 6, no. 1, pp. 55–71, 2021.

[23] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," *The international journal of robotics research*, vol. 32, no. 11, pp. 1231–1237, 2013.

[24] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuscenes: A multimodal dataset for autonomous driving," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11 621–11 631.

[25] W. Li, C. Pan, R. Zhang, J. Ren, Y. Ma, J. Fang, F. Yan, Q. Geng, X. Huang, H. Gong, *et al.*, "Aads: Augmented autonomous driving simulation using data-driven algorithms," *Science robotics*, vol. 4, no. 28, p. eaaw0863, 2019.

[26] Y. Hu, J. Yang, L. Chen, K. Li, C. Sima, X. Zhu, S. Chai, S. Du, T. Lin, W. Wang, *et al.*, "Planning-oriented autonomous driving," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 17 853–17 862.

[27] Y. Yan, H. Lin, C. Zhou, W. Wang, H. Sun, K. Zhan, X. Lang, X. Zhou, and S. Peng, "Street gaussians: Modeling dynamic urban scenes with gaussian splatting," in *ECCV*, 2024.

[28] Y. Li, N. Snavely, D. Huttenlocher, and P. Fua, "Worldwide pose estimation using 3d point clouds," in *European conference on computer vision*. Springer, 2012, pp. 15–29.

[29] Y. Shi and H. Li, "Beyond cross-view image retrieval: Highly accurate vehicle localization using satellite image," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 17 010–17 020.

[30] Y. Li, R. Bu, M. Sun, W. Wu, X. Di, and B. Chen, "Pointcnn: Convolution on x-transformed points," in *Advances in Neural Information Processing Systems*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds., vol. 31. Curran Associates, Inc., 2018.

[31] S. Agarwal, Y. Furukawa, N. Snavely, I. Simon, B. Curless, S. M. Seitz, and R. Szeliski, "Building rome in a day," *Communications of the ACM*, vol. 54, no. 10, pp. 105–112, 2011.

[32] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[33] H. Bay, T. Tuytelaars, and L. Van Gool, "Surf: Speeded up robust features," in *European conference on computer vision*. Springer, 2006, pp. 404–417.

[34] J. Zhang, Y. Yao, and B. Deng, "Fast and robust iterative closest point," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 7, pp. 3450–3466, 2021.

[35] Z. Zhang, M. Xu, W. Zhou, T. Peng, L. Li, and S. Poslad, "Bevlocator: An end-to-end visual semantic localization network using multi-view images," *Science China Information Sciences*, vol. 68, no. 2, p. 122106, 2025.

[36] Z. Li, W. Wang, H. Li, E. Xie, C. Sima, T. Lu, Q. Yu, and J. Dai, "Bevformer: learning bird's-eye-view representation from lidar-camera via spatiotemporal transformers," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[37] Y. Lipman, R. T. Chen, H. Ben-Hamu, M. Nickel, and M. Le, "Flow matching for generative modeling," *arXiv preprint arXiv:2210.02747*, 2022.

[38] C. Zhang, C. Zhang, M. Zhang, and I. S. Kweon, "Text-to-image diffusion models in generative ai: A survey," *arXiv preprint arXiv:2303.07909*, 2023.

[39] S. Gao, J. Yang, L. Chen, K. Chitta, Y. Qiu, A. Geiger, J. Zhang, and H. Li, "Vista: A generalizable driving world model with high fidelity and versatile controllability," *Advances in Neural Information Processing Systems*, vol. 37, pp. 91 560–91 596, 2024.

[40] Y. Yan, Z. Xu, H. Lin, H. Jin, H. Guo, Y. Wang, K. Zhan, X. Lang, H. Bao, X. Zhou, *et al.*, "Streetcrafter: Street view synthesis with controllable video diffusion models," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 822–832.

[41] C. Chi, Z. Xu, S. Feng, E. Cousineau, Y. Du, B. Burchfiel, R. Tedrake, and S. Song, "Diffusion policy: Visuomotor policy learning via action diffusion," *The International Journal of Robotics Research*, p. 02783649241273668, 2023.

[42] C. Jiang, A. Cornman, C. Park, B. Sapp, Y. Zhou, D. Anguelov, *et al.*, "Motiondiffuser: Controllable multi-agent motion prediction using diffusion," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 9644–9653.

[43] C. R. Shipan and C. Volden, "The mechanisms of policy diffusion," *American journal of political science*, vol. 52, no. 4, pp. 840–857, 2008.

[44] G. Barquero, S. Escalera, and C. Palmero, "Belfusion: Latent diffusion for behavior-driven human motion prediction," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 2317–2327.

[45] T. Wang, C. Zhang, X. Qu, K. Li, W. Liu, and C. Huang, "Diffad: A unified diffusion modeling approach for autonomous driving," *arXiv preprint arXiv:2503.12170*, 2025.

[46] R. Zhao, Y. Fan, Z. Chen, F. Gao, and Z. Gao, "Diffe2e: Rethinking end-to-end driving with a hybrid action diffusion and supervised policy," *arXiv preprint arXiv:2505.19516*, 2025.

[47] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[48] J. Philion and S. Fidler, "Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d," in *European conference on computer vision*. Springer, 2020, pp. 194–210.

[49] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[50] B. Liao, S. Chen, H. Yin, B. Jiang, C. Wang, S. Yan, X. Zhang, X. Li, Y. Zhang, Q. Zhang, *et al.*, "Diffusiondrive: Truncated diffusion model for end-to-end autonomous driving," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 12 037–12 047.

[51] Y. Shi, X. Yu, D. Campbell, and H. Li, "Where am i looking at? joint location and orientation estimation by cross-view matching," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 4064–4072.

[52] S. Zhu, T. Yang, and C. Chen, "Vigor: Cross-view image geo-localization beyond one-to-one retrieval," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 3640–3649.