

深度学习为何能改进视觉 SLAM?

Giovanni Cioffi, Davide Scaramuzza

摘要—视觉同时定位与建图 (Visual Simultaneous Localization and Mapping, V-SLAM) 是一项成熟技术, 已广泛应用于各类现实场景。然而, 在低纹理、严重运动模糊和弱光照等具有挑战性的视觉条件下, 其性能仍会下降。基于深度学习 (Deep Learning, DL) 的系统将学习式二维数据关联和不确定性与可微几何优化相结合, 并采用循环架构, 其性能优于经典几何方法并达到当前最佳水平。但究竟哪些组件从根本上促成了这一成功, 仍不明确。本文提出如下问题: 深度学习系统的优越性能主要源于学习式二维数据关联、学习式二维数据关联与不确定性的结合, 还是循环架构本身? 本文通过受控研究对这一问题进行实证分析。研究结果表明, 基于 DL 的 V-SLAM 系统之所以成功, 关键在于学习式二维数据关联和不确定性, 而非循环架构; 这凸显了采用学习范式设计这些组件的必要性。

关键词—SLAM, 机器人与自动化中的深度学习, 定位, 建图。

补充材料

视频: <https://youtu.be/EiuZ7MT0iVc>

代码: https://github.com/uzh-rpg/why_does_deep_learning_improve_visual_slam

I. 引言

视觉同时定位与建图 (Visual Simultaneous Localization and Mapping, V-SLAM) 在重建周围环境地图的同时估计相机运动。它是机器人导航 [1]、自动驾驶 [2]、增强现实与虚拟现实 (Augmented Reality/Virtual Reality, AR/VR) [3]、视障人士辅助系统 [4] 以及医疗操作 [5] 的关键支撑技术。

近三十年来, V-SLAM 系统一直沿用以几何计算机视觉为基础的前端-后端架构 [6]。在基于特征的方法中, 前端通过手工设计的关键点和描述子检测并跟踪视觉观测, 后端则通过光束法平差 (Bundle Adjustment, BA) 最小化重投影误差, 进而估计相机位姿与三维结构。直接法绕过显式特征提取, 直接处理像素强度, 并由后端最小化光度误差。尽管这些系统在许多场景中精度很高 [6], 面对具有挑战性的真实环境时却存在内在脆弱性。具体而言, 基于特征的方法依赖可重复检测的特征与描述子, 直接法则假设光度一致; 在极低纹理场景 [7]、严重运动模糊 [8]、高动态范围场景 [9] 以及光照不足 [10] 或存在动态物体 [11] 的环境中, 这些依赖和假设往往导致系统失效。

深度学习 (Deep Learning, DL) 改变了 V-SLAM 的性能格局。早期端到端方法直接从图像序列回归相机运动 [12, 13], 但通常难以泛化到训练期间未见的视觉条件和运动分布。相比之下, 性能最好的系统采用混合范式: 在循环架构中将学习式视觉前端与几何优化后端相结合 [14, 15, 16]。典型代表是 DROID-SLAM [14], 它将基于 RAFT 的光流网络 [17] 与几何光束法平差相结合。光流网

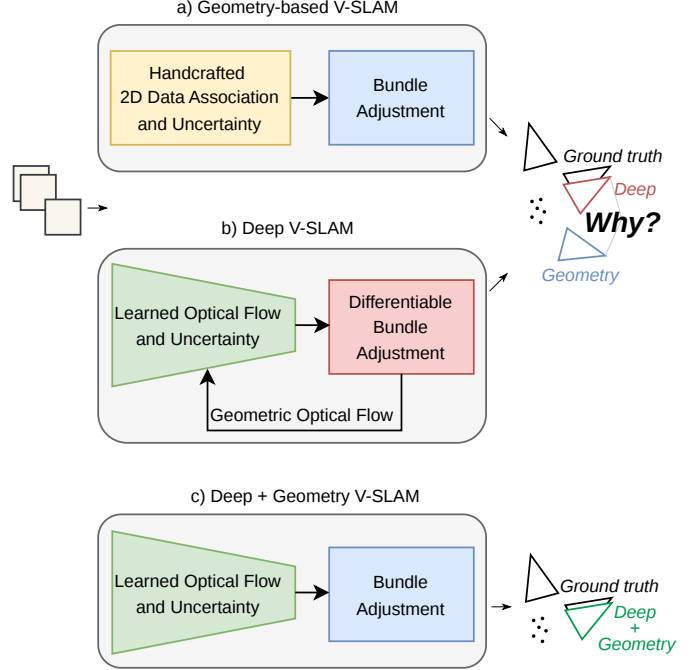


图 1: 揭示深度学习为何能提升视觉同时定位与建图 (Visual SLAM, V-SLAM) 性能。(a) 经典几何 V-SLAM 依赖手工设计的特征匹配与几何优化。(b) 现代深度学习 SLAM 系统具有更强的鲁棒性, 但尚不清楚哪些学习式组件发挥了关键作用。(c) 将学习式二维数据关联与学习式不确定性集成到经典几何流程中, 可分离并评估二者的影响; 实验表明, 这两个组件使经典系统达到了当前最佳性能。

络预测稠密对应关系与置信度估计, 并据此定义光束法平差的重投影约束。光束法平差随后估计相机位姿和三维地图点深度, 用它们计算更新后的几何光流, 再将其反馈至网络以指导下一次迭代。这一循环架构是 DROID-SLAM 性能优异的关键创新之一。DROID-SLAM 是一种跟踪所有像素的稠密方法, 但采用了显著降低的图像分辨率。沿用这一范式的后续工作 (如 DPVO 与 DPV-SLAM) 表明, 由随机选取图像块组成的稀疏表示也能达到相近精度。尽管结构各异, 这些系统都依赖学习式光流、不确定性估计和循环架构。DROID-SLAM、DPVO 和 DPV-SLAM 代表了轨迹估计的当前最佳水平, 因而被选为本文的研究对象。较新的系统利用新型场景表示 [18, 19, 20, 21] 与三维基础模型 [22, 23, 24], 在室内环境稠密重建中取得很高精度, 但在非结构化室外环境与相机高速运动条件下仍缺乏鲁棒性。

基于 DL 的 V-SLAM 取得成功, 引出了该领域的一个关键问题: 深度学习系统的优越性能主要源于学习式二维数据关联、学习式二维数据关联与不确定性的结合, 还是循环架构本身? 回答这一问题至关重要: 它决定了未来的 V-SLAM 系统是否应彻底放弃传统的手工设计视觉前端, 仅在后端优化中应用几何原理, 并保留前馈架构。

本文通过受控实验回答这一问题。本文以成熟且高

作者来自瑞士苏黎世大学信息学系机器人与感知课题组, <https://rpg.ifi.uzh.ch>。

本研究获得欧盟“地平线欧洲”研究与创新计划 (资助协议编号 101120732, AUTOASSESS) 以及欧洲研究理事会 (ERC, 资助协议编号 864042, AGILEFLIGHT) 的支持。

度优化的几何 V-SLAM 系统ORB-SLAM3为基础, 用DROID-SLAM 采用的光流网络替换其基于描述子的数据关联模块。由此得到采用学习式光流的ORB-SLAM3 (ORB-SLAM3-OF) : 该系统保留ORB-SLAM3的经典SLAM 骨干, 以学习式对应关系取代手工设计的描述子匹配。进一步地, 在采用学习式光流与不确定性的ORB-SLAM3 (ORB-SLAM3-OF-U) 中, 使用光流网络预测的不确定性估计为光束法平差中的残差加权。这样即可分离并量化学习式不确定性的作用; DROID-SLAM、DPVO 和 DPV-SLAM 均使用它为几何残差加权, 近期研究也将其用于改善挑战性条件下视觉里程计的立体匹配 [25] 与多传感器融合 [26]。本文方法仅改变同一 SLAM 流程中的数据关联与不确定性机制。因此, 可以直接衡量这两个组件在多大程度上解释了几何方法与学习方法之间的 V-SLAM 性能差距。

本文将标准ORB-SLAM3流程、ORB-SLAM3-OF和ORB-SLAM3-OF-U与 DROID-SLAM、DPVO 及 DPV-SLAM 等领先学习式系统进行比较。实验在 TartanAir [7] 与 UZH-FPV [8] 上开展。这两个极具挑战性的基准将 V-SLAM 系统置于多种运动模式和不利感知条件下, 包括低纹理、弱光照与严重运动模糊。结果表明, 以学习式光流替代手工设计的描述子匹配可显著提升ORB-SLAM3性能。引入学习式不确定性后, 在视觉退化场景中的鲁棒性进一步提升; 这类场景难以估计对应关系, 统一的残差权重也不再可靠。最终得到的ORB-SLAM3-OF-U达到当前最佳性能: 在分布内数据 (TartanAir) 上可与现代深度学习 V-SLAM 系统媲美, 在分布外序列 (UZH-FPV) 上则显著优于它们。

研究结果表明, 配备学习式二维数据关联与学习式不确定性的经典前馈 SLAM 系统可获得最佳性能。这揭示了一个关键结论: 基于 DL 的 V-SLAM 系统之所以成功, 关键在于学习式二维数据关联和学习式不确定性, 而非其循环架构, 说明这些组件必须采用学习范式进行设计。本文的研究发现有望指导下一代 V-SLAM 系统的设计。

综上, 如图 1所示, 本文的主要贡献如下:

- 提出一个研究 V-SLAM 学习式数据关联的受控框架。在保持经典ORB-SLAM3流程不变的前提下, 使用源自 DROID-SLAM 的学习式光流对应估计器替换ORB-SLAM3中基于手工描述子的匹配模块, 从而分离学习式数据关联的作用。
- 系统评估学习式不确定性对光束法平差残差加权的作用。将光流网络预测的不确定性估计引入ORB-SLAM3的光束法平差, 并量化其对位姿估计的影响。
- 在极具挑战性的 V-SLAM 基准上开展广泛的实证研究。在 TartanAir 和 UZH-FPV 上的实验表明, 学习式对应关系与学习式不确定性显著提升了经典 V-SLAM 系统, 使ORB-SLAM3-OF-U达到当前最佳性能。

II. 相关工作

V-SLAM 领域经历了从早期概率建模到现代空间智能系统的深刻演进。近期出版的《SLAM 手册》[6] 全面梳理了这一演进过程。该书指出, SLAM 不只是状态估计问题, 还涉及因子图 [27]、李群 [28] 和鲁棒优化 [29] 等多学科挑战。当前 V-SLAM 研究的前沿方向是经典几何方法与深度学习方法的协同。

A. 经典几何 SLAM 系统

按照前端方法, 经典几何 V-SLAM 系统通常分为基于特征的方法和直接法。基于特征的方法 (如 ORB-SLAM 系列 [30, 31, 32]) 依赖手工设计的关键点 [33] 与描述子 [34] 的提取和匹配, 并通过最小化重投影误差估计相机运动与场景结构。相比之下, DSO [35] 等直接法直接处理像素强度并最小化光度误差。直接法能够生成更稠密的地图, 在低纹理环境中表现更好, 但对光照变化、滚动快门伪影和非朗伯表面明显更加敏感。此外还有 SVO [36] 等混合方法, 它们提取手工设计的关键点并最小化光度误差。

经典系统的后端架构大致分为基于滤波和基于优化的框架。MonoSLAM [37] 等早期系统以及计算高效的 MSCKF [38] 使用扩展卡尔曼滤波器 (Extended Kalman Filter, EKF) 维护状态向量。然而, 正如文献 [6] 所述, 该领域已大体转向因子图上的非线性最小二乘优化。OKVIS [39] 与 VINS-Mono [40] 等基于优化的系统可通过光束法平差重新处理历史状态, 因而具有更高的一致性与精度 [41]。

经典几何 SLAM 的主要优势在于数学可解释性, 以及在资源受限硬件上的高效率。但在剧烈运动、严重运动模糊或极端光照场景中, 这类系统往往表现不佳。关于这些经典失效模式的全面综述, 读者可参阅文献 [42]。

B. 基于深度学习的 SLAM 系统

深度学习的发展为克服经典几何方法的局限带来了新范式。DeepVO [12] 与 TartanVO [13] 等早期端到端系统尝试直接从图像序列回归相机位姿。这类方法虽然对视觉伪影具有一定鲁棒性, 却难以泛化到未见环境。

一种成功的范式采用混合架构, 将学习式前端与可微几何后端相结合 [14, 15, 16]。D3VO [43] 是最早在保留几何优化的同时利用深度学习实现鲁棒数据关联的系统之一, 但并未采用可微后端。DROID-SLAM [14] 率先将循环光流网络与可微 BA 层结合, 开创了这一混合方法。该系统利用光束法平差模块施加几何一致性, 同时受益于光流网络的鲁棒数据关联能力。DPVO [15] 与 DPV-SLAM [16] 等后续工作通过采用基于稀疏图像块的表示进一步改进了这一方法, 在保持当前最佳精度的同时显著降低计算开销。

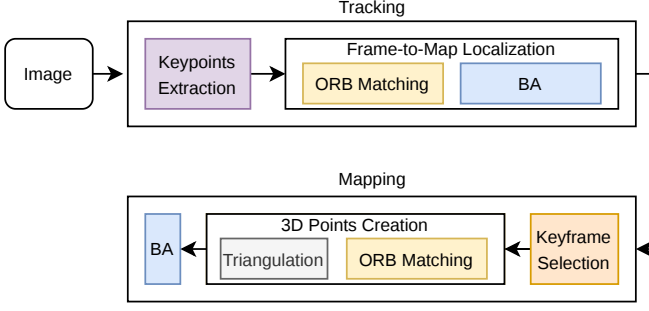
该领域最新的发展是三维基础模型和神经表示的兴起。iMAP[44] 与 NICE-SLAM [18, 19] 等系统利用神经辐射场 (Neural Radiance Fields, NeRF) [45] 进行联合跟踪与稠密建图。近来, 高斯泼溅 SLAM (Gaussian Splatting SLAM, GS-SLAM) [20, 21] 因能提供高保真实时渲染而受到关注。MASt3r-SLAM [22] 与 VGGT-SLAM [23, 24] 等系统则建立在 MASt3r [46] 和 VGGT [47] 等前馈三维几何基础模型之上; 这些模型只需一次前向计算, 便可从未经标定的图像对完成稠密三维重建。这些模型利用大规模 Transformer 预训练消解投影歧义, 即使视觉重叠极低也能建立稠密对应关系。

基于 DL 的系统优势在于面对感知挑战时的鲁棒性。本文连接了这两类方法, 研究将基于 DL 的系统所具备的鲁棒数据关联与不确定性估计集成到可靠的经典几何 V-SLAM 框架后, 将如何影响系统性能。

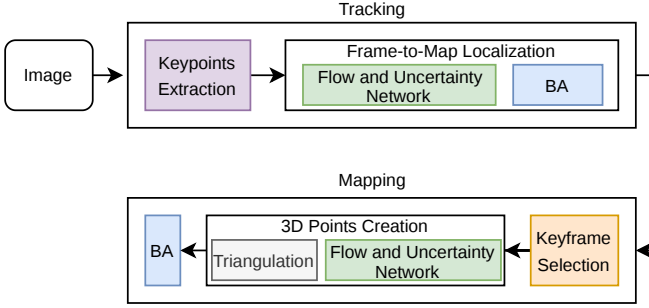
III. 方法

本节介绍用于评估学习式二维数据关联与不确定性对经典几何式 V-SLAM 架构影响的研究框架。为回答本文的研究问题, 我们以基于深度学习 (Deep Learning, DL)

ORB-SLAM3



ORB-SLAM3-OF-U



DROID-SLAM

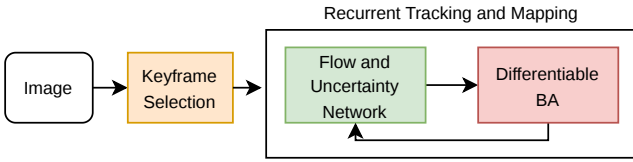


图 2: 经典 ORB-SLAM3 架构 (顶部)、集成学习式光流与不确定性的 ORB-SLAM3 架构 (中部)、以及 DROID-SLAM、DPVO 和 DPV-SLAM 的共用架构 (底部)。

的光流估计器系统性替换当前最佳经典基线中的数据关联模块。该模块化方法在保留几何光束法平差 (Bundle Adjustment, BA) 后端进行位姿优化的同时, 可直接比较两种范式。本节其余内容安排如下: 首先说明研究所用 V-SLAM 系统的选择依据; 随后介绍这些系统的核心架构组件; 最后详述如何将学习式二维数据关联与不确定性模块集成至经典 V-SLAM 流程。

A. V-SLAM 系统的选择

选择具有代表性的系统对于本实证研究的有效性至关重要。尽管许多基于几何和 DL 的 V-SLAM 系统在特定环境或传感器配置下表现出色, 但本文优先选用在多样场景中始终达到当前最佳性能的系统。所选框架均拥有成熟、稳定的代码库, 也是机器人领域公认的基准。

几何式系统采用 ORB-SLAM3 (具体为 ORB-SLAM3 实现)。ORB-SLAM3 被广泛认为是最稳健的几何式 V-SLAM 系统, 其完整流程涵盖跟踪、局部建图和全局闭环检测。其二维数据关联依赖描述子匹配, 代表了传统 SLAM 范式。以学习式方案替换描述子匹配, 可以相对于高度优化的经典方法量化基于 DL 的数据关联所带来的收益。

基于 DL 的系统则评估 DROID-SLAM、DPVO 和 DPV-SLAM。DROID-SLAM 提出一种将学习式前端与可微光束法平差层结合的循环架构, 相比端到端方法具有更高的精度和泛化能力。后续工作 DPVO 和 DPV-SLAM 采用基于稀疏图像块的表示扩展了这一思路, 以提高计算效率。这些系统代表了当前轨迹估计精度的最高水平。尽管利用三维基础模型和神经表示的新兴方法在室内场景中展现出较高精度和稠密重建能力, 但面对高速运动或非结构化室外环境时往往不够稳健。

B. 学习式 V-SLAM 系统

DROID-SLAM 在帧图上运行, 其中的边表示图像之间的共视关系。该架构为每条边提取特征图并构建相关体, 以评估视觉相似度。DROID-SLAM 的光流网络是基于卷积门控循环单元 (Convolutional Gated Recurrent Unit, ConvGRU) 的更新算子, 可迭代预测光流修正量及相应置信度权重, 再将其传入可微稠密 BA 层以优化相机位姿和逐像素深度。优化后的相机位姿和深度用于计算新的光流预测, 并将其作为光流网络的输入之一。DROID-SLAM 的循环结构是其取得成功的原因之一。DROID-SLAM 最后通过光束法平差优化完整的关键帧轨迹和三维地图点。实验还评估了禁用该光束法平差的版本 DROID-VO。

DROID-SLAM 对稠密光流图的依赖导致较高的计算与内存开销。为解决这一效率瓶颈, DPVO 被提出作为 DROID-SLAM 的稀疏对应方案。DPVO 不再预测每个像素的稠密光流, 而是跨时间跟踪一组稀疏图像块。这些图像块采用随机采样; 与传统关键点检测器相比, 这一策略取得了出人意料的良好效果。DPVO 使用二部图像块图表示图像块与帧之间的关系。其循环更新算子包含时序卷积和全局消息传递层, 用于在图像块图的边之间传播信息。DPVO 在保持有竞争力精度的同时, 运行速度快于 DROID-SLAM, 且内存需求更低。

DPV-SLAM 通过引入高效的闭环检测与全局优化, 将 DPVO 扩展为完整的 SLAM 框架。深度 SLAM 面临的一项主要挑战是 CUDA 操作的顺序执行特性, 这往往会在后端优化期间导致帧率下降。DPV-SLAM 为此维护一个共享的图像块图, 同时纳入局部里程计因子和长距离闭环因子。该系统采用两种闭环检测机制: 邻近闭环和经典闭环。邻近闭环利用预测光流和相机位姿检测闭环, 并向图像块图中插入长距离邻近因子, 以进行全局光束法平差。经典闭环利用图像检索 (如 BoW2 [48]) 以及在 CPU 上运行的位姿图优化, 校正显著的尺度漂移。本文启用邻近闭环, 但禁用经典闭环模块。DPVO 和 DPV-SLAM 使用与 DROID-SLAM 相同的循环光流网络架构。DROID-SLAM、DPVO 和 DPV-SLAM 的共用架构如图 2 底部所示。

C. 在 ORB-SLAM3 中集成学习式数据关联

ORB-SLAM3 是一种基于特征的 V-SLAM 系统, 面向多样室内外环境中的实时运行。该系统采用 FAST 角点作为关键点, 并使用 ORB 描述子。系统从每个输入帧中提取关键点及其对应描述子。系统架构包含三个主要并行线程: 跟踪、局部建图和闭环检测。所有任务 (跟踪、建图、重定位和闭环检测) 均使用 ORB 特征, 从而保证效率。跟踪线程通过将 ORB 描述子与局部地图匹配, 并使用仅运动 BA 优化相机位姿, 对每一帧进行相机定位。局部建图线程处理新关键帧, 并通过局部 BA 优化局部重建。探索

过程中会较为宽松地生成关键帧，以保持跟踪稳健性，随后再采用严格的筛选策略。该策略移除冗余信息，得到紧凑且可跟踪的地图，使系统能够长期运行。闭环线程会为每个新关键帧搜索闭环。检测到闭环后，系统计算七自由度相似变换以校正累积尺度漂移，并在本质图上执行位姿图优化。该稀疏子图既能维持全局一致性，又能在大规模环境中保持较高计算效率。对于单目相机，ORB-SLAM3 采用基于模型选择的自动初始化过程。系统并行计算两个模型，即适用于平面场景的单应矩阵和适用于非平面场景的基础矩阵，再通过启发式规则选择最合适的模型，确保仅在视差充足时创建可靠的初始地图。原始版本仅支持单目相机，后续版本直至 ORB-SLAM3 已扩展为支持双目相机、RGB-D 传感器和惯性测量单元 (Inertial Measurement Unit, IMU)。

为单独研究学习式数据关联的贡献，集成过程中保持 ORB-SLAM3 的基本系统设计（跟踪、建图和闭环线程）不变。本文仅以 DROID-SLAM 的光流网络替换二维对应关系引擎。ORB-SLAM3 的系统架构如图 2 顶部所示。本文使用单目模式。

1) 光流网络推理：取自 DROID-SLAM 的光流网络在 TartanAir 数据集上预训练。推理时，将输入图像转换为灰度图并下采样至 80×60 分辨率（原始 TartanAir 分辨率的 $1/8$ ）。特征网络计算特征图，并据此构建四维相关体。上下文网络同时提取上下文特征。推理过程的核心是一个 3×3 卷积 GRU (ConvGRU)，用于迭代优化光流估计。

不同于标准 DROID-SLAM 实现，本文使用零光流先验初始化，以消除对运动模型先验的依赖。因此，用于从相关体提取特征的对应该为恒等映射。ConvGRU 的输入包括相关特征、诱导光流（当前对应关系与初始零光流先验之差）以及上下文特征。其输出为光流修正量，即对当前对应场的修正以及置信度图。ConvGRU 持续迭代，直至平均光流修正量低于 0.0625（在原始图像分辨率下约等于 0.5 个像素）或达到最多 20 次迭代。最终光流估计通过双线性插值上采样至原始图像分辨率。

2) 特征匹配流程：在标准 ORB-SLAM3 跟踪线程中，系统利用恒速运动模型，将与上一帧（或关键帧）关键点关联的三维地图点重投影至当前帧。随后在重投影点周围的局部搜索窗口内比较 ORB 描述子，以寻找匹配。这些匹配作为仅运动 BA 的输入，用于优化当前相机帧的位姿。若帧间跟踪失败，即 BA 中的内点数低于阈值（实验中设为 7），系统将尝试帧到关键帧定位。若该定位同样失败，系统会进入丢失状态，并尝试使用 BoW2 在全局地图中定位当前相机。

在本文提出的 ORB-SLAM3-OF 和 ORB-SLAM3-OF-U 变体中，描述子匹配模块被替换。给定与源帧二维关键点 (u_s, v_s) 关联的三维地图点，利用网络预测的光流 (f_u, f_v) 确定其在目标帧中的预测位置 (u_t, v_t) ： $(u_t, v_t) = (u_s + f_u, v_s + f_v)$ 。随后与 (u_t, v_t) 周围 2 像素半径内最近的特征建立对应关系。若该半径内不存在特征，则将此地图点标记为在当前帧中不可见。该过程同时用于帧间跟踪和帧到关键帧定位。

局部建图线程还利用学习式光流网络建立关键帧间的特征对应关系，前提是关键帧对位于 5 帧的时间窗口内。若时间间隔超过该阈值，系统恢复使用标准的基于 BoW 的匹配方法。其底层匹配机制与帧间跟踪及帧到关键帧跟踪所用流程相同。在 ORB-SLAM3-OF 变体中，若匹配满足前述跟踪模块的几何和距离准则，则予以接受。在 ORB-SLAM3-OF-U 变体中还施加额外的有效性要求：仅

当预测光流置信度超过 0.3 的阈值时，才将候选匹配视为有效。随后利用已建立的特征匹配对新三维地图点进行三角化。ORB-SLAM3-OF 和 ORB-SLAM3-OF-U 的系统架构如图 2 中部所示。

3) 光束法平差中的学习式不确定性：在标准 ORB-SLAM3 系统中，仅运动 BA 的重投影误差权重由特征提取所在的图像金字塔层级决定（原始图像分辨率对应 1.0，分辨率较低的更高金字塔层级采用更低权重）。ORB-SLAM3 使用八层图像金字塔，每升高一层，图像分辨率按 1.2 的因子降低。相比之下，ORB-SLAM3-OF 和 ORB-SLAM3-OF-U 仅在原始图像分辨率下提取 FAST 角点。ORB-SLAM3 在光束法平差期间采用基于 χ^2 检验的外点剔除策略，显著性水平设为 0.05。优化器每迭代 10 次便执行一次 χ^2 检验，共执行 4 次。未通过检验的点会被标记为外点，并从后续优化迭代中移除。在 ORB-SLAM3-OF 中，所有重投影误差项统一赋予 1.0 的权重。在 ORB-SLAM3-OF-U 中，利用光流网络预测的置信度图对 BA 中的重投影误差加权。这使后端能够降低网络不确定区域内对应关系的权重。所有版本的 BA 均采用 Huber 损失函数和 χ^2 检验。

由于网络联合估计光流及其对应不确定性，不确定性形式与学习式数据关联空间天然绑定。因此，将基于经典 ORB 的匹配与该学习式不确定性结合，在方法上并不一致。

IV. 实验

我们评估了 ORB-SLAM3、ORB-SLAM3-OF、ORB-SLAM3-OF-U、DROID-VO、DROID-SLAM、DPVO 和 DPV-SLAM 的性能，具体方法见第 III 节。这些方法在 TartanAir 和 UZH-FPV 数据集上进行测试。由于涵盖剧烈运动、大幅光流及多变光照条件，这两个数据集已成为 V-SLAM 的标准基准。评估的主要目标是比较预测相机轨迹的准确性。为此，以平移 (ATE_T [m]) 和旋转 (ATE_R [deg]) 绝对轨迹误差 (Absolute Trajectory Error, ATE) 作为主要评估指标 [50]。首先使用 Umeyama 方法 [49] 求得的 7 自由度相似变换，将估计轨迹与真值对齐，而后计算上述指标。针对每个数据集分别调节各 V-SLAM 系统的参数。每个 V-SLAM 系统运行三次，并报告平均 ATE。除非另有说明，每个系统均处理序列中的每一帧，所报告的轨迹误差为所有帧上的平均值。实验在运行 Ubuntu 22.04 的计算机上进行，配备 Intel Core i9 2.3GHz CPU 和 Nvidia A100 80GB GPU。

A. TartanAir 数据集

1) 实验设置：TartanAir 数据集 [7] 包含多种模拟环境，涵盖城市、乡村、住宅、公共区域和科幻场景等室内外环境。每个序列均通过定义真值轨迹并渲染相应的 RGB 帧与位姿生成，同时提供沿轨迹采样的深度、光流、事件、LiDAR 和 IMU 测量等多模态数据。TartanAir 包含数十万帧。得益于这一规模，它被用于训练本文采用的当前最佳 DL V-SLAM 系统 (DROID-SLAM、DPVO、DPV-SLAM)，也已成为训练 DL 端到端三维几何重建方法的标准数据集 [46, 51]。本文使用相关工作中常用的 15 个测试序列对 V-SLAM 系统进行基准评估¹。这些测试序列按难度分为简单 (ME) 和困难 (MH) 两类，具有很高挑战性。

¹由于存在表面缺失等渲染伪影，评估中排除了序列 MH004。

表 I: TartanAir 数据集上的平移 (ATE_T) 和旋转 (ATE_R) 绝对轨迹误差。平移指标单位为米, 旋转指标单位为度。所有方法均在相同序列集合上比较。结果按序列逐项列出。数值越低表示性能越好, **粗体**表示平移和旋转的最佳结果。符号“-”表示 SLAM 系统未能在整个序列中持续跟踪。学习式 ORB-SLAM3 版本优于标准 ORB-SLAM3 版本。采用与不采用学习式不确定性的两个变体总体性能相当。学习式 ORB-SLAM3 版本的性能与 DL 系统相当。

序列	评估指标: ATE_T [m] / ATE_R [deg]						
	ORB-SLAM3	ORB-SLAM3-OF	ORB-SLAM3-OF-U	DROID-VO	DROID-SLAM	DPVO	DPV-SLAM
ME000	0.17 / 0.85	0.27 / 0.49	0.20 / 0.62	0.14 / 0.47	0.10 / 0.20	0.13 / 0.36	4.98 / 20.08
ME001	0.07 / 0.24	0.06 / 0.24	0.07 / 0.23	0.09 / 0.18	0.05 / 0.10	0.04 / 0.14	0.06 / 0.12
ME002	1.84 / 1.65	0.15 / 0.35	0.24 / 0.41	0.14 / 1.31	0.13 / 0.32	0.11 / 0.15	1.49 / 1.51
ME003	2.42 / 4.45	1.11 / 0.41	0.69 / 0.62	1.17 / 1.37	1.15 / 1.28	0.41 / 0.45	1.31 / 5.44
ME004	-	1.79 / 3.95	1.20 / 3.39	0.84 / 0.76	0.90 / 0.90	0.42 / 0.59	0.52 / 0.88
ME005	0.10 / 0.91	0.10 / 0.91	0.09 / 1.08	0.09 / 0.66	0.10 / 0.59	0.13 / 0.33	0.21 / 0.95
ME006	-	0.61 / 1.14	0.57 / 0.64	0.57 / 0.92	0.45 / 1.10	1.00 / 0.58	9.22 / 2.50
ME007	0.05 / 0.28	0.10 / 0.43	0.08 / 0.36	0.24 / 0.79	0.08 / 0.25	0.19 / 0.66	1.06 / 5.69
MH000	1.17 / 2.82	0.64 / 0.96	0.51 / 1.28	0.31 / 0.62	0.07 / 0.26	0.52 / 1.41	0.53 / 1.62
MH001	0.09 / 1.30	0.06 / 0.69	0.04 / 0.46	0.44 / 1.01	0.39 / 0.75	0.04 / 0.49	0.04 / 0.42
MH002	-	0.09 / 0.40	0.06 / 0.33	0.07 / 0.11	0.02 / 0.06	0.03 / 0.11	1.34 / 4.66
MH003	-	0.08 / 0.82	0.11 / 0.70	0.05 / 0.50	0.01 / 0.20	0.08 / 0.25	0.04 / 0.28
MH005	-	0.98 / 1.39	0.45 / 0.89	0.53 / 0.79	0.52 / 0.77	0.16 / 0.26	0.65 / 1.96
MH006	-	0.28 / 1.35	0.24 / 1.08	0.16 / 0.40	0.17 / 0.51	0.08 / 0.29	0.48 / 3.82
MH007	-	0.86 / 1.64	0.61 / 0.96	0.83 / 0.33	0.09 / 0.16	0.25 / 0.28	4.51 / 5.79
平均值	-	0.48 / 1.01	0.34 / 0.87	0.38 / 0.61	0.28 / 0.50	0.24 / 0.42	1.76 / 3.72

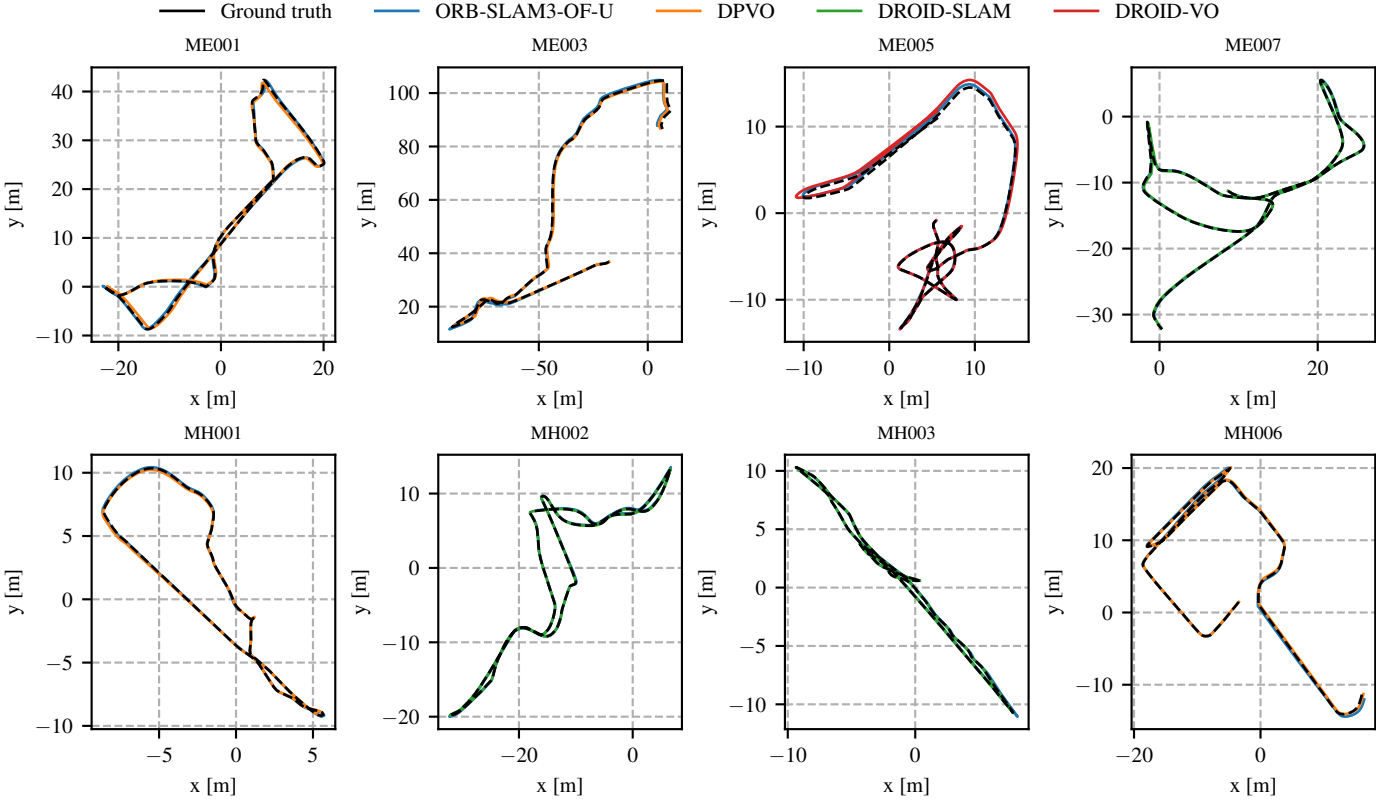


图 3: TartanAir 数据集中八个序列的俯视轨迹。将真实轨迹与 ORB-SLAM3-OF-U 及按 ATE_T 衡量表现最佳的基于 DL 的 SLAM 方法所得轨迹进行比较。所有预测轨迹均基于每个序列的前 100 帧, 采用文献 [49] 的方法与真实轨迹对齐。

挑战包括大幅光流位移 (主要出现在 *MH* 序列中, 如图 4 顶图所示)、动态物体 (*ME004*)、光照突变 (*ME007*) 以及近乎全黑的环境 (*MH000*, 如图 5 中间一行所示)。

2) 结果: 表 I 汇总了所有测试序列的绝对轨迹误差。此外, 图 3 从俯视角度对比了真值轨迹、ORB-SLAM3-OF-U 的估计轨迹以及按 ATE_T 衡量表现最佳的 DL SLAM 方法的估计轨迹。

融合学习式光流的 ORB-SLAM3 变体 (无论是否使用不

确定性) 均显著优于标准 ORB-SLAM3。这一性能差距在困难 (*MH*) 序列中最为明显, 标准 ORB-SLAM3 仅能完成七个序列中的两个。这些序列具有较大的帧间光流位移和不利光照条件, 二者都会降低基于 ORB 描述子的匹配可靠性。在标准 ORB-SLAM3 算法中, 相邻帧间的特征匹配依赖恒速运动模型, 对当前帧中的特征位置作初始预测。为保证计算效率, 描述子匹配的搜索范围被限制在预测坐

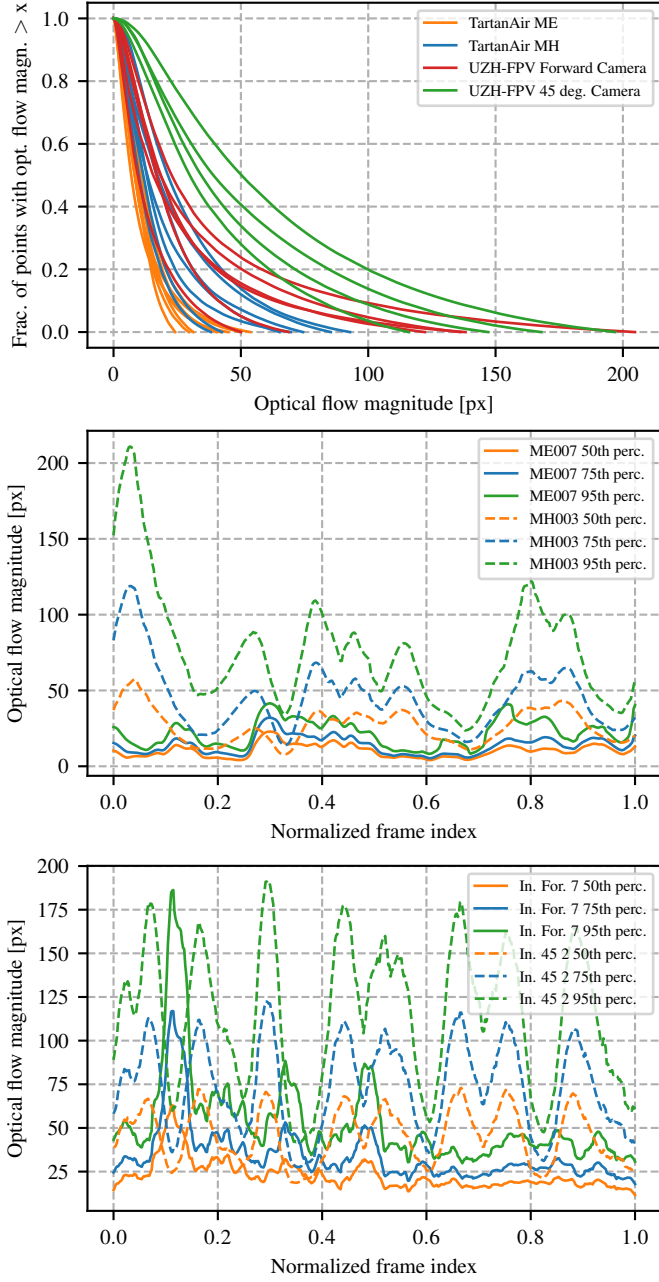


图 4: 相邻帧间光流幅值分析。UZH-FPV 序列的光流幅值高于 TartanAir 序列。上: 光流幅值超过阈值 x 的像素比例。中: TartanAir 数据集中代表性简单序列 (*ME007*) 与困难序列 (*MH003*) 的第 50、75 和 95 百分位数对比。下: UZH-FPV 数据集中代表性前视相机序列 (*Indoor Forward 7*) 与 45 deg 相机序列 (*Indoor 45 3*) 的相同百分位数对比。

标周围的局部区域内。在 V-SLAM 系统中, 对当前帧的所有描述子进行穷举搜索在计算上不可行。然而, *MH* 序列中的大加速度及其导致的大幅光流位移经常违背恒速假设。这会产生不准确的先验, 使搜索窗口错过真实特征对应关系, 最终导致跟踪失败。图 4 的顶图印证了这一点: *MH* 序列 (蓝色) 的平均光流幅值始终高于 *ME* 序列 (橙色)。图 4 中间的 *ME007* (简单) 与 *MH003* (困难) 对比进一步说明了这一点。标准 ORB-SLAM3 在 *ME007* 上的表现与学习式版本相当, 但在 *MH003* 上完全丢失跟踪。困难序列还具有极具挑战性的光照条件。图 5 展示了一组

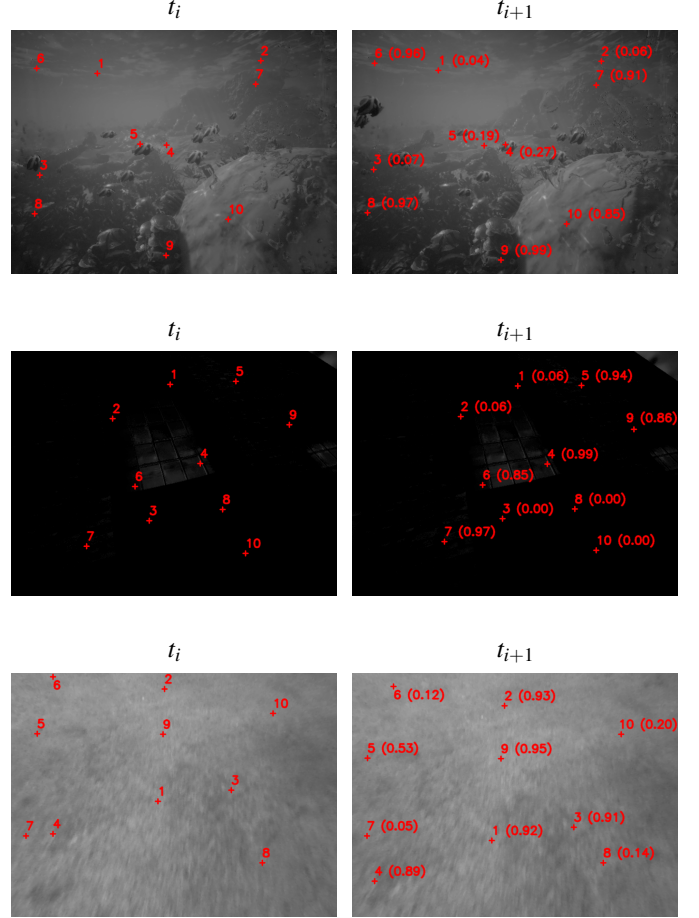


图 5: 光流及相应置信度分数 (括号内) 的可视化。上、中、下三行分别展示 TartanAir 数据集序列 *ME004*、*MH000* 以及 UZH-FPV 数据集序列 *In. 45 14* 中的两幅相邻图像。每组均可视化五个高置信度预测和五个低置信度预测。即使在视觉跟踪困难的低纹理区域, 模型仍能以高置信度预测光流。

基于光流的特征匹配及其预测置信度, 分别来自光照充足的环境 (顶图中的序列 *ME004*) 和弱光、低纹理区域 (中间图中的序列 *MH001*)。即使在传统描述子匹配失效的低纹理或黑暗区域, 光流网络仍能保持高置信度预测。因此, 在序列 *MH001* 上, ORB-SLAM3-OF-U 的 ATE_T 和 ATE_R 均比标准 ORB-SLAM3 提升 55%。

ORB-SLAM3-OF 与 ORB-SLAM3-OF-U 的总体性能相当, 但引入光流不确定性可在特定场景中带来可观增益。例如, 序列 *ME004* 是包含鱼类和移动植被等动态对象的水下环境, ORB-SLAM3-OF-U 的 ATE_T 相比 ORB-SLAM3-OF 提升 33%。类似地, *MH005* 中密集植被产生重复纹理, 光流不确定性显著升高 (见图 6)。在此情况下, 考虑不确定性使 ATE_T 提升 55%。

总体而言, 融合学习式光流后, ORB-SLAM3 能够成功处理所有序列, 性能接近当前最佳的 DL 系统。ORB-SLAM3-OF-U 与 DPVO 之间最大的性能差距出现在序列 *ME004* 上 (ATE_T 分别为 1.20 和 0.42)。尽管 ORB-SLAM3 和 DPVO 均未专门针对动态环境设计, 但结果表明 DPVO 在此类场景中具有更强的鲁棒性。此外, DPV-SLAM 在这些环境中的性能下降主要源于错误的回环检测。表 III 第一行给出了处理单帧所需的平均时间, 计时从图像加载开始, 到位姿与地图点预测结束。

表 II: UZH-FPV 数据集上的平移 (ATE_T) 和旋转 (ATE_R) 绝对轨迹误差。结果按序列逐项列出。数值越低表示性能越好, **粗体**表示平移和旋转的最佳结果。符号 “-” 表示 SLAM 系统未能在整个序列上保持跟踪, 符号 “*” 表示误差指标并非基于序列中的全部图像计算。**完成所有序列必须使用学习式不确定性。融合学习式光流与不确定性的ORB-SLAM3版本优于 DL 系统。**

序列	评估指标: $ATE_T[m]$ / $ATE_R[deg]$						
	ORB-SLAM3	ORB-SLAM3-OF	ORB-SLAM3-OF-U	DROID-VO	DROID-SLAM	DPVO	DPV-SLAM
In. Forw. 3	0.09 / 0.45	0.12 / 0.44	0.07 / 0.42	1.55 / 0.88	0.79 / 2.25	0.40 / 1.11	0.02 / 0.37
In. Forw. 6	0.12 / 2.24	0.05 / 0.56	0.07 / 0.59	1.32 / 0.98	0.43 / 2.32	0.41 / 0.58	0.03 / 0.50
In. Forw. 7	0.27 / 3.40	0.54 / 1.11	0.20 / 2.60	5.11 / 21.67	5.11 / 22.64	0.63 / 2.31	0.65 / 1.86
In. 45 2	-	-	0.24 / 2.13	1.09 / 6.93	0.06 / 1.61	1.06 / 7.42	0.92 / 6.67
In. 45 9*	-	-	0.32 / 6.11	0.90 / 4.39	0.15 / 2.12	0.81 / 4.45	0.71 / 4.48
In. 45 14*	-	-	0.64 / 3.90	5.27 / 22.41	5.26 / 22.05	2.79 / 9.24	3.46 / 10.09
Out. Forw. 1	1.50 / 1.69	0.31 / 1.00	0.78 / 1.67	3.23 / 0.54	4.29 / 2.27	1.04 / 0.63	0.29 / 0.41
Out. Forw. 3	-	2.01 / 2.17	2.32 / 2.03	8.47 / 0.87	8.35 / 1.54	3.94 / 1.23	2.85 / 0.94
Out. Forw. 5	-	-	0.72 / 1.43	2.28 / 2.59	2.34 / 2.63	0.99 / 0.60	0.77 / 0.66
Out. 45 1	-	-	0.98 / 2.91	1.91 / 5.17	1.91 / 5.26	1.69 / 3.27	1.06 / 3.44
平均值	-	-	0.63 / 2.38	3.11 / 6.64	2.87 / 6.47	1.38 / 3.08	1.08 / 2.94

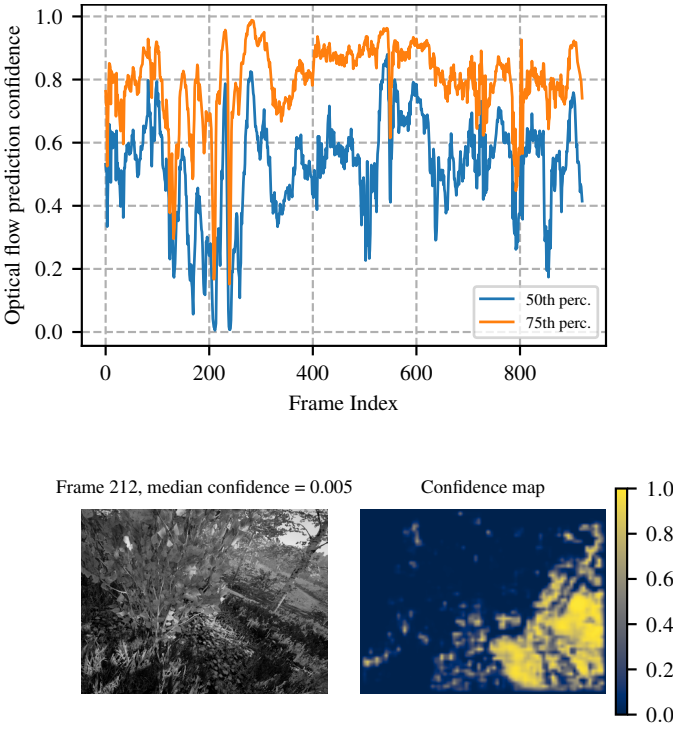


图 6: TartanAir 序列 MH005 的光流预测置信度分析。(上) 相邻帧间置信度的第 50 和 75 百分位数。(下) 置信度中位数最低帧的图像及相应置信度图。

B. UZH-FPV 数据集

1) 实验设置: UZH-FPV 数据集 [8] 由竞速无人机执行剧烈机动时采集的真实室内外序列组成。无人机配备同步相机和 IMU, 并由专业飞手操控。相机图像以 30 Hz 频率记录。序列按难度 (简单、中等和困难) 及相机朝向 (前视或向下倾斜 45 deg) 分类。难度分类主要依据光流幅值。UZH-FPV 被认为是最具挑战性的 V-SLAM 基准之一。其剧烈相机运动产生的光流幅值显著超过 TUM-RGBD [52]、EuRoC [53] 和 TartanAir 等成熟数据集。如图 4 顶图所示, UZH-FPV 序列的光流幅值显著高于 TartanAir 数据集。实验从不同环境与相机配置中选取具有代表性的序列子集。具体而言, 对于室内 (45 度和前视) 及室外 (前视) 配置, 分别评估一个简单序列 (*In. 45 2*、*In. Forw.*

3、*Out. Forw. 1*)、一个中等序列 (*In. 45 9*、*In. Forw. 6*、*Out. Forw. 3*) 和一个困难序列 (*In. 45 14*、*In. Forw. 7*、*Out. Forw. 5*)。此外还包括 *Out. 45 1*, 这是室外 45 度配置中唯一具有公开真值的序列。该数据集的真值通过联合优化框架 [54] 计算, 该框架最小化视觉、惯性和三维位置残差。三维位置约束由机器人全站仪跟踪无人机上安装的棱镜获得, 从而确保高速飞行时仍能得到高精度参考轨迹。

2) 结果: 表 II 汇总了实验所评估序列的绝对轨迹误差。相应地, 图 7 从俯视角度展示了真值轨迹、ORB-SLAM3-OF-U 的估计轨迹, 以及按 ATE_T 衡量表现最佳的 DL SLAM 方法的估计轨迹。

与 TartanAir 数据集上的发现一致, 融合学习式光流与不确定性后, 性能较标准 ORB-SLAM3 显著提升。这一差距在相机向下倾斜 45 deg 安装的配置中最为明显。在这些序列中, 标准 ORB-SLAM3 无法完成任何一条轨迹。相比之下, 在前视相机采集的室内序列中, ORB-SLAM3-OF-U 的性能增益较小。45 deg 配置尤其具有挑战性: 帧间光流位移较大 (如图 4 顶图和底图所示), 且相机正对地面时会出现低纹理区域 (如图 5 底图所示)。

此处 ORB-SLAM3-OF-U 与 ORB-SLAM3-OF 的性能差距明显大于 TartanAir 实验中的差距。具体而言, ORB-SLAM3-OF 无法完成任何 45 deg 序列, 也无法完成 *Out. Forw. 5* (室外困难) 序列。在 45 deg 朝向下, 无人机飞行速度较高且靠近地面, 导致光流置信度很低 (如图 8 所示)。快速运动还会使大部分场景在相邻帧间移出视场, 大幅降低视觉重叠。

在序列 *In. 45 9* 和 *In. 45 14* 中, 硬件问题导致所有 ORB-SLAM3 版本均发生跟踪失败: 记录缺失两帧, 产生了 100 ms 的时间间隔。由此造成的视觉重叠不足使系统丢失当前帧的跟踪。因此, 在这两个序列上, 所有系统的误差指标分别仅基于前 76% 和 84% 的帧计算。DL 方法通过恒速运动模型初始化帧位姿, 从而保持跟踪连续性。当缺少视觉测量来约束局部光束法平差 (BA) 时, 这些方法仅传播先验运动模型, 而基于 ORB-SLAM3 的系统会触发跟踪丢失状态。总体而言, 融合学习式光流与不确定性后, ORB-SLAM3 在这一高难度数据集上的性能超过了当前最佳 DL 系统。表 III 第二行给出了处理单帧所需的平均时间。

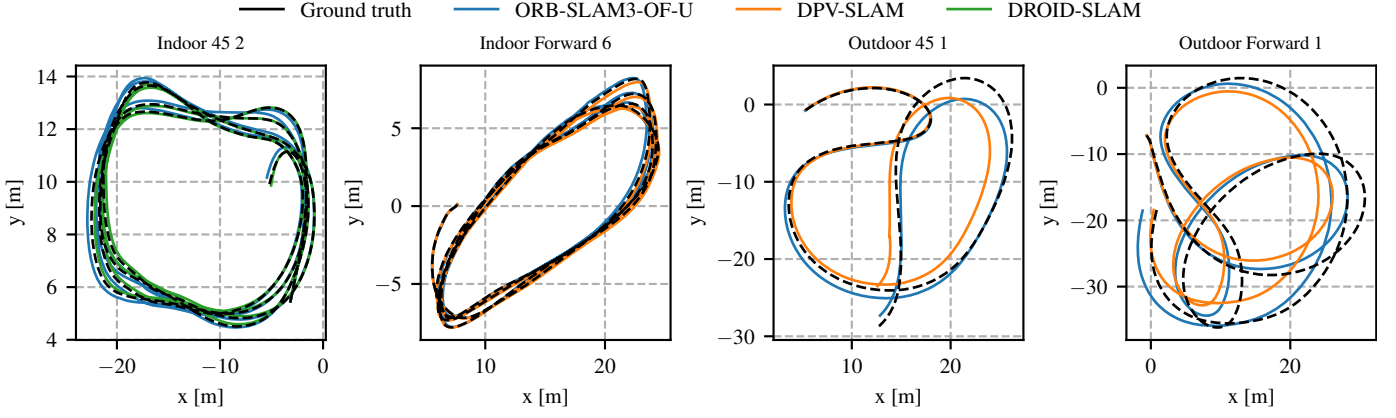


图 7: UZH-FPV 数据集中四个序列的俯视轨迹。对比真值轨迹、ORB-SLAM3-OF-U 的轨迹, 以及按 ATE_T 衡量表现最佳的 DL SLAM 方法的轨迹。基于每个序列的前 100 帧, 使用文献 [49] 的方法将所有预测轨迹与真值对齐。

表 III: 在 TartanAir 和 UZH-FPV 数据集的两个选定序列上进行运行时间评估 (单位: 秒)。所报告数值为处理单帧所需的平均时间, 计时从图像加载开始, 到位姿与地图点预测结束。

序列	每帧平均时间 [sec]						
	ORB-SLAM3	ORB-SLAM3-OF	ORB-SLAM3-OF-U	DROID-VO	DROID-SLAM	DPVO	DPV-SLAM
TartanAir ME007	0.024	0.050	0.050	0.101	0.201	0.043	0.048
UZH-FPV Ind. Forw. 3	0.025	0.051	0.051	0.159	0.360	0.038	0.091

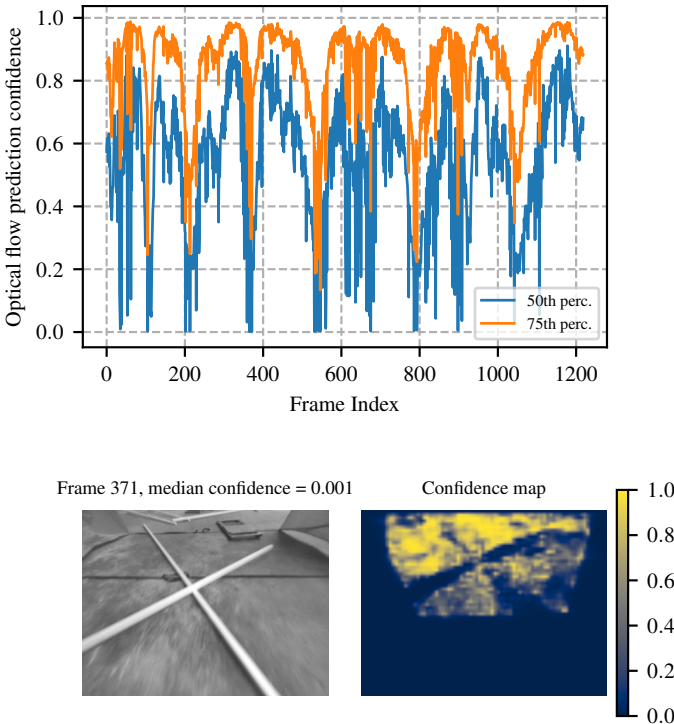


图 8: UZH-FPV 序列 *In. 45 2* 的光流预测置信度分析。(上) 相邻帧间置信度的第 50 和 75 百分位数。(下) 置信度中位数最低帧的图像及相应置信度图。

V. 讨论

1) 学习式二维数据关联是否是深度学习系统获得优越性能的关键因素? 研究结果为学习式二维数据关联确实是现代基于 DL 的 V-SLAM 系统性能更优的主要原因提供了有力实证依据。尽管基线 ORB-SLAM3 的性能始终低于基于 DL 的方法, 但集成学习式光流后, 无论底层架构有何差异, ORB-SLAM3 的性能均达到与当前最佳 DL 系统相当的水平。

学习式二维数据关联在挑战性场景中的优势最为显著, 包括剧烈运动导致的大光流位移 (见 UZH-FPV 数据集), 以及不利光照或低纹理环境 (见 TartanAir 数据集的 *MH* 序列)。在这些情况下, 经典描述子匹配算法通常会失败, 或需要针对具体环境手动调整参数。本研究表明, 学习式前端能够有效应对这些挑战, 说明下一代 V-SLAM 系统应优先采用学习式数据关联模块。

此外, 基于 DL 的前端减轻了手动调参的负担。经典几何流程依赖众多启发式参数, 例如特征提取的像素强度阈值和描述子匹配的汉明距离上限。寻找一组能泛化至不同环境的通用参数极具挑战, 通常还需要专家参与。相比之下, DL 技术将挑战从手动调参转移到了多样且有代表性的训练数据收集上。只要训练数据能够反映预期部署条件, 学习式前端就能实现更稳健的泛化。近期工作 [55, 56] 也在探索这一方向, 为提升 V-SLAM 在视觉退化场景中的稳健性提供了可行路径。

2) 关键是否在于将学习式二维数据关联与学习式不确定性相结合? 评估结果表明, 学习式不确定性估计对于稳健定位至关重要, 尤其是在底层二维数据关联过程噪声较大时。这一点在 UZH-FPV 数据集上最为明显: 从仿真 TartanAir 训练数据到真实高速无人机竞速图像的域偏移, 增大了光流预测的不确定性。在这些高噪声条件下,

ORB-SLAM3 所采用的经典外点剔除策略, 如 χ^2 统计检验和稳健代价函数 (例如 Huber 范数), 通常无法在光束法平差期间充分过滤错误对应关系或降低其权重。本研究从实证角度表明, 引入学习式不确定性后, 光束法平差能够考虑二维数据关联的内在随机性。根据网络置信度对重投影误差加权, 可使系统达到固定统计阈值难以实现的稳健性水平。因此, 若要使 V-SLAM 系统在具有挑战性的视觉环境中稳健运行, 就必须将学习式数据关联与学习式不确定性结合。

3) 还是说, 循环架构才是整合这些组件并超越经典方法的关键机制? 评估得到的一项关键认识是, 要达到当前最佳轨迹估计精度, 并非必须采用 DROID-SLAM、DPVO 和 DPV-SLAM 所使用的循环架构。在 UZH-FPV 数据集上, 本文的学习式 ORB-SLAM3 变体将经典 ORB-SLAM3 前馈架构与学习式二维数据关联和不确定性估计相结合, 取得了最佳性能。这一优势源于经典 ORB-SLAM3 系统是一个复杂的多阶段流程, 涵盖帧到地图配准、局部建图、复杂的关键帧选择和闭环检测, 各模块均由许多相互依赖的参数控制。更关键的是, 该架构包含结构性冗余, 即使存在大量外点测量 (例如有噪声的二维数据关联), 仍能准确估计轨迹。这说明, 现代基于深度学习的 V-SLAM 系统取得高性能的根本驱动力, 主要来自学习式二维数据关联及相应不确定性的稳健性。由此可以明确下一代 SLAM 的设计原则: 将学习式前端与可解释且计算高效的几何后端结合, 即可实现当前最佳的精度与稳健性。

VI. 结论

本文系统研究了基于 DL 的 V-SLAM 系统相较经典几何框架性能更优的架构驱动因素。通过以 ORB-SLAM3 作为经典基线并集成 DROID-SLAM 的光流网络, 本文回答了一个根本性研究问题: 深度学习系统的优越性能主要源于学习式二维数据关联、学习式二维数据关联与不确定性的结合, 还是循环架构本身? 在 TartanAir 和 UZH-FPV 数据集上的实证评估表明, 学习式二维数据关联确实是基于 DL 的系统超越经典几何流程的首要因素。学习式光流的集成使 ORB-SLAM3 能够克服帧间大位移、运动模糊和挑战性光照条件等传统失效模式; 这些情况通常会导致定位失败。结果表明, 配备学习式数据关联后, ORB-SLAM3 等经典 SLAM 架构也能达到当前最佳性能。此外, 本研究证实学习式不确定性估计对于稳健定位不可或缺。经典外点剔除方法虽然在许多情况下效果良好, 但在高噪声场景中往往不够充分。研究发现, 与仅依赖经典外点剔除方法相比, 利用学习式置信度对重投影误差加权, 可显著提高视觉退化条件下的轨迹精度。综上, 下一代 V-SLAM 架构应在经典前馈架构中, 将稳健的学习式数据关联与不确定性估计前端同几何光束法平差相结合。

参考文献

- [1] Kamak Ebadi, Lukas Bernreiter, Harel Biggie, et al. Present and future of SLAM in extreme environments: The DARPA SubT challenge. *IEEE Trans. Robot.*, 2024.
- [2] Joel Janai, Fatma Güney, Aseem Behl, and Andreas Geiger. Computer vision for autonomous vehicles: Problems, datasets and state of the art. *Foundations and Trends in Computer Graphics and Vision*, 2020.
- [3] Jakob Engel, Kiran Somasundaram, Michael Goesele, et al. Project aria: A new tool for egocentric multi-modal ai research. *arXiv preprint arXiv:2308.13561*, 2023. Accessed 2026-04-20.
- [4] Patrick Pfreundschuh, Giovanni Cioffi, Cornelius von Einem, et al. Sight guide demonstrates robotics-inspired vision assistance at the cybathlon. *Science Robotics*, 2025.
- [5] Emma Most, Jonas Hein, Frédéric Giraud, et al. Acquiring submillimeter-accurate multi-task vision datasets for computer-assisted orthopedic surgery. *Int. Journal of Computer Assisted Radiology and Surgery*, 2025.
- [6] Luca Carlone, Ayoung Kim, Timothy Barfoot, Daniel Cremers, and Frank Dellaert, editors. *SLAM Handbook. From Localization and Mapping to Spatial Intelligence*. Cambridge University Press, 2026.
- [7] Wenshan Wang, DeLong Zhu, Xiangwei Wang, Yaoyu Hu, Yuheng Qiu, Chen Wang, Yafei Hu, Ashish Kapoor, and Sebastian Scherer. Tartanair: A dataset to push the limits of visual slam. In *IEEE/RSJ Int. Conf. Intell. Robot. Syst. (IROS)*, 2020.
- [8] Jeffrey Delmerico, Titus Cieslewski, Henri Rebecq, Matthias Faessler, and Davide Scaramuzza. Are we ready for autonomous drone racing? the uzh-fpv drone racing dataset. In *IEEE Int. Conf. Robot. Autom. (ICRA)*, 2019.
- [9] Mathias Gehrig, Willem Aarents, Daniel Gehrig, and Davide Scaramuzza. Dsec: A stereo event camera dataset for driving scenarios. *IEEE Robot. Autom. Lett.*, 2021.
- [10] Michael Helmberger, Kristian Morin, Beda Berner, Nitish Kumar, Giovanni Cioffi, and Davide Scaramuzza. The hilti slam challenge dataset. *IEEE Robot. Autom. Lett.*, 2022.
- [11] E. Palazzolo, J. Behley, P. Lottes, P. Giguère, and C. Stachniss. ReFusion: 3D Reconstruction in Dynamic Environments for RGB-D Cameras Exploiting Residuals. In *IEEE/RSJ Int. Conf. Intell. Robot. Syst. (IROS)*, 2019.
- [12] Sen Wang, Ronald Clark, Hongkai Wen, and Niki Trigoni. Deepvo: Towards end-to-end visual odometry with deep recurrent convolutional neural networks. In *IEEE Int. Conf. Robot. Autom. (ICRA)*, 2017.
- [13] Wenshan Wang, Yaoyu Hu, and Sebastian Scherer. Tartanvo: A generalizable learning-based vo. In *Conf. on Robot. Learning (CoRL)*, 2021.
- [14] Zachary Teed and Jia Deng. Droid-slam: Deep visual slam for monocular, stereo, and rgb-d cameras. *Advances in Neural Information Processing Systems*, 34: 16558–16569, 2021.
- [15] Zachary Teed, Lahav Lipson, and Jia Deng. Deep patch visual odometry. *arXiv preprint arXiv:2208.04726*, 2022. URL <https://arxiv.org/abs/2208.04726>. Retrieved on Jan. 4th, 2023.
- [16] Lahav Lipson, Zachary Teed, and Jia Deng. Deep patch visual slam. In *Eur. Conf. Comput. Vis. (ECCV)*, 2024.

- [17] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Eur. Conf. Comput. Vis. (ECCV)*, 2020.
- [18] Zihan Zhu, Songyou Peng, Viktor Larsson, Weiwei Xu, Hujun Bao, Zhaopeng Cui, Martin R Oswald, and Marc Pollefeys. Nice-slam: Neural implicit scalable encoding for slam. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2022.
- [19] Zihan Zhu, Songyou Peng, Viktor Larsson, Zhaopeng Cui, Martin R Oswald, Andreas Geiger, and Marc Pollefeys. Nicer-slam: Neural implicit scene encoding for rgb slam. In *Int. Conf. on 3D Vision (3DV)*, 2024.
- [20] Hidenobu Matsuki, Riku Murai, Paul HJ Kelly, and Andrew J Davison. Gaussian splatting slam. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2024.
- [21] Vladimir Yugay, Yue Li, Theo Gevers, and Martin R Oswald. Gaussian-slam: Photo-realistic dense slam with gaussian splatting. *arXiv preprint arXiv:2312.10070*, 2023. Accessed 2026-05-06.
- [22] Riku Murai, Eric Dexheimer, and Andrew J Davison. Mast3r-slam: Real-time dense slam with 3d reconstruction priors. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2025.
- [23] Dominic Maggio, Hyungtae Lim, and Luca Carlone. Vggt-slam: Dense rgb slam optimized on the sl (4) manifold. *Conf. Neural Inf. Process. Syst. (NIPS)*, 2025.
- [24] Dominic Maggio and Luca Carlone. Vggt-slam 2.0: Real-time dense feed-forward scene reconstruction. *arXiv preprint arXiv:2601.19887*, 2026. Accessed 2026-04-20.
- [25] Yuheng Qiu, Yutian Chen, Zihao Zhang, Wenshan Wang, and Sebastian Scherer. Mac-vo: Metrics-aware covariance for learning-based stereo visual odometry. In *IEEE Int. Conf. Robot. Autom. (ICRA)*, 2025.
- [26] Shibo Zhao, Sifan Zhou, Yuchen Zhang, Ji Zhang, Chen Wang, Wenshan Wang, and Sebastian Scherer. Resilient odometry via hierarchical adaptation. *Science Robotics*, 2025.
- [27] Frank Dellaert. Factor graphs and gtsam: A hands-on introduction. *Georgia Institute of Technology, Tech. Rep.*, 2012.
- [28] Joan Solà, Jeremie Deray, and Dinesh Atchuthan. A micro Lie theory for state estimation in robotics. *arXiv preprint arXiv:1812.01537*, 2018. Accessed 2026-04-20.
- [29] Giorgio Grisetti, Rainer Kümmerle, Cyrill Stachniss, and Wolfram Burgard. A tutorial on graph-based slam. *IEEE Intel. Transportation Sys. Mag.*, 2011.
- [30] Raul Mur-Artal, Jose Maria Martinez Montiel, and Juan D Tardos. Orb-slam: A versatile and accurate monocular slam system. *IEEE Trans. Robot.*, 2015.
- [31] Raul Mur-Artal and Juan D Tardós. Orb-slam2: An open-source slam system for monocular, stereo, and rgb-d cameras. *IEEE Trans. Robot.*, 2017.
- [32] Carlos Campos, Richard Elvira, Juan J Gómez Rodríguez, José MM Montiel, and Juan D Tardós. Orb-slam3: An accurate open-source library for visual, visual-inertial, and multimap slam. *IEEE Trans. Robot.*, 2021.
- [33] Edward Rosten and Tom Drummond. Machine learning for high-speed corner detection. In *Eur. Conf. Comput. Vis. (ECCV)*, 2006.
- [34] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. Orb: An efficient alternative to sift or surf. In *Int. Conf. Comput. Vis. (ICCV)*, 2011.
- [35] Jakob Engel, Vladlen Koltun, and Daniel Cremers. Direct sparse odometry. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2017.
- [36] Christian Forster, Zichao Zhang, Michael Gassner, Manuel Werlberger, and Davide Scaramuzza. Svo: Semidirect visual odometry for monocular and multicamera systems. *IEEE Trans. Robot.*, 2016.
- [37] Andrew J Davison, Ian D Reid, Nicholas D Molton, and Olivier Stasse. Monoslam: Real-time single camera slam. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2007.
- [38] Anastasios I Mourikis and Stergios I Roumeliotis. A multi-state constraint kalman filter for vision-aided inertial navigation. In *IEEE Int. Conf. Robot. Autom. (ICRA)*, 2007.
- [39] Stefan Leutenegger, Simon Lynen, Michael Bosse, Roland Siegwart, and Paul Furgale. Keyframe-based visual-inertial odometry using nonlinear optimization. *Int. J. Robot. Research*, 2015.
- [40] Tong Qin, Peiliang Li, and Shaojie Shen. Vins-mono: A robust and versatile monocular visual-inertial state estimator. *IEEE Trans. Robot.*, 2018.
- [41] Hauke Strasdat, José MM Montiel, and Andrew J Davison. Visual slam: why filter? *Image and Vision Computing*, 2012.
- [42] Jakob Engel, Juan D. Tardós, Javier Civera, Margarita Chli, Stefan Leutenegger, Frank Dellaert, and Daniel Cremers. Visual SLAM. In Luca Carlone, Ayoung Kim, Timothy Barfoot, Daniel Cremers, and Frank Dellaert, editors, *SLAM Handbook. From Localization and Mapping to Spatial Intelligence*. Cambridge University Press, 2026.
- [43] Nan Yang, Lukas von Stumberg, Rui Wang, and Daniel Cremers. D3vo: Deep depth, deep pose and deep uncertainty for monocular visual odometry. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2020.
- [44] Edgar Sucar, Shikun Liu, Joseph Ortiz, and Andrew J Davison. imap: Implicit mapping and positioning in real-time. In *Int. Conf. Comput. Vis. (ICCV)*, 2021.
- [45] Ben Mildenhall, Pratul P Srinivasan, Matthew Tatik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 2021.
- [46] Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r. In *Eur. Conf. Comput. Vis. (ECCV)*, 2024.
- [47] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David

- Novotny. Vggt: Visual geometry grounded transformer. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2025.
- [48] Dorian Gálvez-López and Juan D Tardos. Bags of binary words for fast place recognition in image sequences. *IEEE Trans. Robot.*, 2012.
- [49] Shinji Umeyama. Least-squares estimation of transformation parameters between two point patterns. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2002.
- [50] Zichao Zhang and Davide Scaramuzza. A tutorial on quantitative trajectory evaluation for visual(-inertial) odometry. In *IEEE/RSJ Int. Conf. Intell. Robot. Syst. (IROS)*, 2018.
- [51] Nikhil Keetha, Norman Müller, Johannes Schönberger, Lorenzo Porzi, Yuchen Zhang, Tobias Fischer, Arno Knapitsch, Duncan Zauss, Ethan Weber, Nelson Antunes, et al. Mapanything: Universal feed-forward metric 3d reconstruction. In *Int. Conf. on 3D Vision (3DV)*, 2026.
- [52] Jürgen Sturm, Nikolas Engelhard, Felix Endres, Wolfram Burgard, and Daniel Cremers. A benchmark for the evaluation of rgb-d slam systems. In *IEEE/RSJ Int. Conf. Intell. Robot. Syst. (IROS)*, 2012.
- [53] Michael Burri, Janosch Nikolic, Pascal Gohl, Thomas Schneider, Joern Rehder, Sammy Omari, Markus W Achtelik, and Roland Siegwart. The euroc micro aerial vehicle datasets. *Int. J. Robot. Research*, 2016.
- [54] Giovanni Cioffi, Titus Cieslewski, and Davide Scaramuzza. Continuous-time vs. discrete-time vision-based slam: A comparative study. *IEEE Robot. Autom. Lett.*, 2022.
- [55] Nico Messikommer, Giovanni Cioffi, Mathias Gehrig, and Davide Scaramuzza. Reinforcement learning meets visual odometry. In *Eur. Conf. Comput. Vis. (ECCV)*, 2024.
- [56] Simone Nascivera, Leonard Bauersfeld, Jeff Delaune, and Davide Scaramuzza. Image-conditioned adaptive parameter tuning for visual odometry frontends. *arXiv preprint arXiv:2603.21785*, 2026. Accessed 2026-04-15.