doi:10.19306/j.cnki.2095-8110.2024.06.002

基于深度学习的图像深度感知 SLAM 综述

渠海榕,杨钊龙,张　海,任　章

(北京航空航天大学自动化科学与电气工程学院,北京 100191)

摘　要:同步定位与地图构建(SLAM)技术在精度和建图方面取得了显著进展,并广泛应用于家用机器人和自动驾驶等领域。随着深度学习和神经网络的快速发展,现阶段的神经网络已经具备从大量数据中学习普适规律的能力,而且还能作为一种新型三维表示方法。基于此,将深度学习与 SLAM 技术相结合的方法成为了研究热点。概述了 SLAM 技术与基于深度学习的图像深度感知技术结合的最新进展,对最新的方法进行了总结,并提出了一种可行的框架构建 SLAM 系统。其中的深度感知技术包括深度估计网络、神经辐射场(NeRF)和三维高斯喷溅(3DGS)技术,详细分析了这 3 种深度感知技术之间的联系以及它们在 SLAM 中的潜在应用,为 SLAM 技术的未来发展提供了一个新的视角,并为进一步的研究提供了参考。

关键词:同步定位与地图构建;深度学习;图像深度估计;里程计;智能定位技术;神经辐射场;三维高斯喷溅

中图分类号:V249.32　　**文献标志码**:A　　**文章编号**:2095-8110(2024)06-0011-17

A review of SLAM based on deep learning image depth perception

QU Hairong, YANG Zhaolong, ZHANG Hai, REN Zhang

(School of Automation Science and Engineering, Beihang University, Beijing 100191, China)

Abstract: Simultaneous localization and mapping (SLAM) technology has made significant progress in accuracy and mapping, and has been widely used in areas such as home robotics and autonomous driving. With the rapid development of deep learning and neural networks, neural networks at this stage have the ability to learn universal laws from a huge amount of data, and can also be used as a new 3D representation method. Based on this, the method of combining deep learning with SLAM technology has become a research hotspot. Recent advances in combining the SLAM techniques with deep learning-based image depth perception techniques is outlined, the latest approaches are summarized, and a feasible framework for building SLAM systems is proposed, where the depth perception techniques inside include depth estimation networks, neural radiance field (NeRF) and 3D Gaussian splatting (3DGS). The relationship between these three depth perception techniques and their potential applications in SLAM are analysed in detail, offering a new perspective for the future development of SLAM and providing references for future studies.

Key words: Simultaneous localization and mapping (SLAM); Deep learning; Image depth estimation; Odometer; Intelligent positioning technology; Neural radiance field (NeRF); 3D Gaussian splatting(3DGS)

收稿日期: 2024-09-04;**修订日期**:2024-10-29

基金项目:国家自然科学基金(62373031);贵州省科技计划项目(2023-341)

作者简介:渠海榕(2001—),男,博士研究生,主要从事视觉语言导航、机器人导航与控制等方面的研究。

通信作者:张海(1970—),男,博士,副教授,主要从事最优估计理论及组合导航方面的研究。

0 引言

同步定位与地图构建(simultaneous localization and mapping, SLAM)技术是机器人自主导航中的关键技术之一。SLAM 技术在计算机视觉和机器人学中扮演着关键角色,它使机器能够在未知环境中进行自主导航并构建地图。随着机器人、计算机视觉、传感器技术以及人工智能的快速发展,SLAM 技术在各领域中的应用需求日益增长。SLAM 技术的核心在于,机器人能够在未知环境中实时构建环境地图,并准确确定自身在地图中的位置。该技术广泛应用于无人驾驶汽车、无人机、移动机器人、增强现实(augmented reality, AR)和虚拟现实(virtual reality, VR)等领域。随着人工智能和传感器技术的快速发展,SLAM 技术的研究不断取得新的突破,具有重要的理论意义和实际应用价值。

SLAM 技术的起源可以追溯到 20 世纪 90 年代初期,当时的研究主要集中在利用扩展卡尔曼滤波器(extended Kalman filter, EKF)解决机器人在未知环境中的定位和地图构建问题。EKF-SLAM 利用卡尔曼滤波器对机器人的位置和环境特征进行联合估计,尽管这一方法提供了理论框架,但由于计算复杂度高和精度限制,其应用范围相对有限。

21 世纪初,SLAM 技术迎来了重要的发展阶段,其中包括视觉 SLAM 的崛起。视觉 SLAM 通过相机捕获的图像数据实现环境的建图和机器人定位。ORB-SLAM[1]是这一时期的重要进展之一,它通过特征点匹配实现了高效的实时 SLAM。ORB-SLAM 的成功不仅提高了视觉 SLAM 的准确性和实时性,也推动了相关技术的广泛普及与应用。随后,Mur-Artal 等改进了该系统,推出了 ORB-SLAM2,支持单目、立体和 RGB-D 相机 。

与此同时,激光 SLAM 技术的出现进一步丰富了 SLAM 的应用场景。激光 SLAM 系统如 Cartographer[2]利用激光雷达提供的高分辨率距离信息,显著提高了地图构建的精度和效率。激光 SLAM 在处理大型和复杂环境时表现出色,特别适用于无人驾驶汽车和仓储机器人等需要高精度定位的应用场景。

随着计算能力和算法的发展,SLAM 技术在 21 世纪第 2 个十年进入了新的阶段。深度学习的迅猛发展为 SLAM 技术带来了新的机遇。深度学习在图像处理和特征提取方面展现出显著优势,从而使 SLAM 系统在复杂和动态环境中能够取得更好的性能。深度学习技术通过深度神经网络对图像进行深度估计,形成了直接深度估计网络;同时,也有研究将深度网络融入 SLAM 系统的多个模块,形成了间接深度估计网络,以实现更全面的优化。

近年来,神经辐射场(neural radiance field, NeRF)[3]和三维高斯喷溅(3D Gaussian splatting, 3DGS)[4]等新兴技术在 SLAM 领域得到了广泛关注。这些技术通过优化位姿实现实时稠密建图,显著提高了定位精度和地图细节。NeRF 通过学习场景的三维表示,能够生成高质量的稠密地图;而 3DGS 则利用高斯分布的特性进行高效的图像特征提取和匹配。这些进展不仅提升了 SLAM 系统在动态环境中的表现,也为其在复杂场景中的应用提供了新的可能性。

当前,SLAM 技术的发展主要聚焦于 3 个方向:1)实时性和精度:尽管视觉 SLAM 和激光雷达 SLAM 在实时性和精度上不断提升,但在动态环境中仍面临挑战。2)鲁棒性和稳定性:通过引入深度学习和应用多传感器融合技术,提高了 SLAM 系统的鲁棒性和稳定性,但在极端环境下的性能仍需进一步验证与优化。3)不同下游任务的需求对建图提出了不同的要求,例如需要避障或者按照视觉线索进行导航[5]。

本文将详细分析 SLAM 技术的发展历程,并重点探讨深度感知与 SLAM 结合的最新进展,从 2 个主要方面进行综述:深度感知对 SLAM 跟踪部分的精度的提升,以及深度感知在建图层面的改变。通过对现有方法的分析,本文将总结其优势和面临的挑战,并展望未来的发展方向。希望本文能够为 SLAM 技术的进一步研究和实际应用提供新的理论支持和技术指导。

1 深度感知方法与 SLAM 综述

首先说明 NeRF、3DGS、三维场景重建和深度估计之间的联系。

如图 1 所示,NeRF 和 3DGS 都可以视作三维重建的方法,其中 NeRF 于 2020 年被提出,意为神经辐射场,是一种可微的、自动生成的且连续的三维隐式表达方式,最初用于解决新视角合成的任务;3DGS 于 2023 年被提出,意为三维高斯喷溅,是一种显式的三维表达方式,可以直接生成显式的场景的三维结构。有了三维重建的场景,就可以得到 NeRF 和 3DGS 的表示方法,即图 1 中的关系①和③。

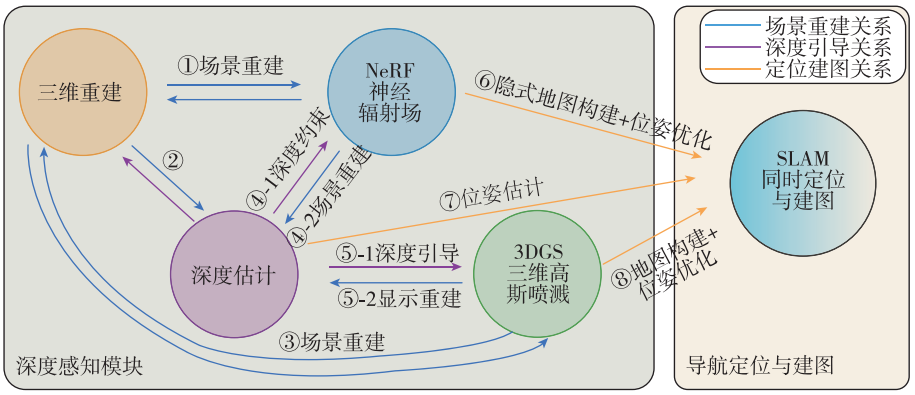


图 1　NeRF,3DGS 和深度估计与导航定位关系

Fig. 1　Relation between NeRF, 3DGS, depth estimation and localization

深度估计,即对一张单目 RGB 图像的每一个像素的深度进行估计,得到深度图。深度图可以分为相对深度和绝对深度,经过训练的绝对深度网络往往只在训练场景下泛化[6],而得到相对深度的网络则泛化性更强。高质量的深度估计在虚拟遮挡处理、路径规划和物体规避等方面均表现出极高的有效性,现有离线的方法虽然可以获取每帧的深度,但是不适用于交互式应用场景。已有研究[7]在假设短期或长期的场景几何快照是获取高质量深度信息来源的前提下,通过引入截断符号距离函数(truncated signed distance function, TSDF)维护 3D 几何的低成本全局表示,利用网络提取这种几何提示,并将其输入给深度估计器以获取实时绝对深度,即一个先验的场景有利于提高深度获取的效果和实时性,对应图 1 中的②。

如图 2 所示, NeRF 的输入是多视角的二维图像及其对应的相机位姿,输出是一个隐式的 3D 场景表示,通过一个神经网络预测每个点的颜色和密度,从而实现高质量的新视角合成。因此,对其体密度场进行渲染可以得到新视角下的深度图,也可以根据真实的深度对 NeRF 添加深度损失,进而更好地优化 NeRF[8],即图 1 中的④。然而,由于数据噪声等影响,渲染的图像中会出现漂浮物,这些漂浮物可被视作空间密度的噪声集群,目前已经有研究[9-10]对这方面进行改进。针对 NeRF 训练和渲染速度较慢的问题,已有许多研究[11-14]致力于加速 NeRF,但至今仍难以找到一种能够在消费级图形处理器(graphics processing unit, GPU)上快速训练(≤1 h),并在普通设备(如手机和笔记本电脑)上以约 30 帧/s 的交互帧率渲染 3D 场景的稳健方法。因此,通过 NeRF 可以得到某个场景下的深度图,是一种具有深度感知能力的技术。

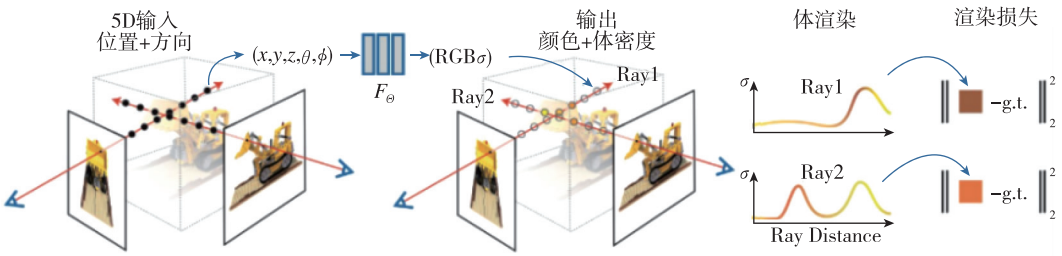


图 2　NeRF 流程图,来自 NeRF 论文[3]

Fig. 2　Flow chart of NeRF, from NeRF paper[3]

相比之下,如图 3 所示,3DGS 的输入也是多视角的二维图像及其相机位姿,但输出的是一组高斯椭球体,这些椭球体的属性(位置、旋转、缩放及透明度等)通过优化可直接渲染逼真的 3D 场景。如果有真实的深度图,可以指导初始的椭球体生成,即图 1 的⑤。3DGS 提供的显式几何表示使得能够在不同视点之间灵活重投影,从而缓解视点之间的错位,相较于NeRF的方法可获得更好的重建效果。

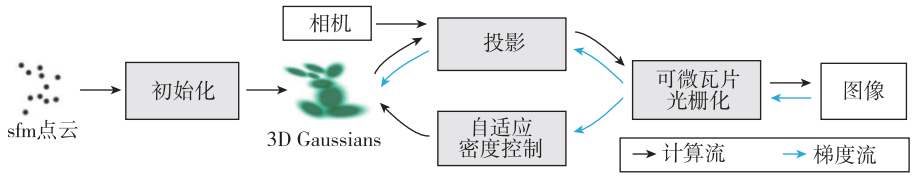


图 3 3DGS 优化流程图，来自 3DGS 原论文[4]

Fig. 3 3DGS optimization flow chart, from 3DGS original paper[4]

此外，3DGS 的实时渲染特性也增强了基于神经网络的 SLAM 方法的实用性，因为其在时间需求上比基于 NeRF 的方法更少。

SLAM 包含跟踪和建图两部分，由于传统的基于特征点法或者直接法建立的地图多为点云地图，主要用于回环检测和后端优化，而建立出的点云地图对于后续任务的指导有限。因此，经典的 SLAM 算法更侧重于跟踪，通过各种约束提升跟踪精度，而对建图效果的要求相对较低。将 NeRF 和 3DGS 与 SLAM 相结合的研究展示了新型地图表示方式，提供了更优的建图效果。然而，目前提出的融合方法在精度上并未显著超越经典 SLAM 算法。因此，本文后续章节将详细地阐述深度网络在 SLAM 跟踪部分的改进，以及 NeRF 和 3DGS 应用在跟踪部分对整体框架的改变及其对建图层面的改进。

2 单目图像深度估计与 SLAM

2.1 单目图像深度估计的深度学习方法概述

1）单目图像深度估计的深度学习方法分类

传统的深度估计方法依赖于各种假设、限制以及优化算法提供密集的深度估计。然而，基于传感器的方法，如利用 RGB-D 相机或激光雷达，面临着测量范围有限、对室外照明条件敏感、标定要求高和能耗高等问题[15]。在高度复杂的深度计算任务中，深度学习方法将深度计算问题转化为一个学习问题，通过对每个像素的深度值进行回归，在单目图像的密集深度估计方面取得了显著进展。具体的进步路线是从最初对卷积神经网络（convolutional neural networks，CNN）的改进，到后来生成式方法的应用，期间还伴随着一些图像增强、语义识别以及时序分析等新方法的产生和应用。单目图像深度估计的深度学习方法可以根据其学习范式分为监督学习、自监督学习和半监督学习。

监督学习方法依赖于标注的训练数据，其中训练图像的每个像素都包含深度标注，使得网络能够学习到输入图像与其对应的像素深度值之间的映射关系。监督学习方法的优点在于，当在标注丰富的数据集上训练时，能够实现很高的精度。然而，大规模的像素级深度标注数据集的制作既耗时又成本高昂[16]。

自监督学习方法利用未标记数据中固有的结构和信息学习深度估计。这些方法通过利用已知相机位姿或场景几何形状的图像对或图像序列，自动生成训练信号。通过将任务表述为重建问题或相对深度排序问题，自监督学习方法能够在没有明确深度标注的情况下学习深度的估计。这种方法减轻了对大量手动标注数据的依赖，但仍然需要仔细地选择数据。例如，在选择相机位姿时，可优先考虑相互间隔较小且视角变化适中的图像对，以有效捕捉场景结构信息；而相机位姿变化过大的图像对则可能引入噪声，导致重建困难。

半监督学习方法旨在平衡监督学习和自监督学习方法的优点。在训练过程中结合了标记数据和未标记数据，通过利用少量标记数据和大量未标记数据，半监督学习方法可以在减少标注工作量的同时，达到有竞争力的性能水平。然而，在联合学习框架中有效利用标记和未标记数据是一项具有挑战性的任务，需要仔细地设计和考虑。

近年来，各种网络架构已被应用于单目深度估计，包括 CNN、全卷积神经网络（fully convolutional networks，FCN）、编码器-解码器（encoder-decoder，ED）、自动编码器（autoencoder，AE）、生成对抗网络（generative adversarial networks，GAN）、Transformer 架构以及扩散模型（diffusion model）。每种架构在准确性、效率和计算要求方面都展现出独特的优势和权衡。在以下部分中，将更详细地描述每种类型方法的学习模式。

2）单目图像深度估计的深度学习方法评估指标

以下是一些常用的误差函数和评估深度估计准确性的指标，一个成功的方法应该具有更高的准确性和更低的误差。对于总像素个数为 N 的像素，每个像素 i 的真实深度为 d_i，预测的深度值为 $\hat{d}_i$，使用的指标如表 1 所示。

表 1　深度估计方法的评估指标

Tab. 1　Estimation metrics for assessing depth estimation methods

指标名称	描述	计算公式
绝对相对误差 AbsRel (absolute relative error)	用于衡量预测深度与真实深度之间的差异	$\mathrm{AbsRel}=\frac{1}{N}\sum_{i=1}^{N}\frac{\lvert d_i-\hat{d}_i\rvert}{d_i}$
平方相对误差 SqRel (squared relative error)	用于衡量预测深度与真实深度之间的差异	$\mathrm{SqRel}=\frac{1}{N}\sum_{i=1}^{N}\frac{(d_i-\hat{d}_i)^2}{d_i^2}$
均方根误差 RMSE (root mean squared error)	用于衡量预测值与真实值之间的差异	$\mathrm{RMSE}=\sqrt{\frac{1}{N}\sum_{i=1}^{N}(d_i-\hat{d}_i)^2}$
对数均方根误差 log-RMSE	RMSE 的变体，首先对深度取对数，然后计算 RMSE	$\mathrm{RMSE(log)}=\sqrt{\frac{1}{N}\sum_{i=1}^{N}(\log(d_i)-\log(\hat{d}_i))^2}$
尺度不变均方根误差	RMSE 的变体，引入一个尺度对齐值以进一步减小尺度变化的影响	$\mathrm{sRMSE(log)}=\sqrt{\frac{1}{N}\sum_{i=1}^{N}(\log(d_i)-\log(\hat{d}_i)+a(d_i,\hat{d}_i))^2}$ 其中，$a(d_i,\hat{d}_i)=\frac{1}{N}\sum_i(\log\hat{d}_i-\log d_i)$
阈值精度 (δ_x, accuracy with threshold)	用于衡量在特定阈值下的预测深度的准确性	$\mathrm{Accuracy}(\delta_x)=\frac{\mathrm{sum}\left[\max\left(\frac{d_i}{\hat{d}_i},\frac{\hat{d}_i}{d_i}\right)<\mathrm{thr}\right]}{N}$ 其中，$\mathrm{thr}=1.25(x=1),1.25^2(x=2),1.25^3(x=3)$

3)单目图像深度估计的深度学习数据集

主要受到当前自动驾驶汽车研究趋势的推动，深度估计领域创建并公开了一些有限但广泛使用的基准数据集。深度估计基准数据集的一个共同要求是，场景图像应配备密集的深度数据，形成所谓的真值(ground truth)。深度数据可以用作最小化深度误差估计的参考点。然而，大多数数据集在这方面并不完备，通常仅提供稀疏的深度值。在其他情况下，例如使用立体图像或成对的视频帧进行自监督学习时，深度图可能完全缺失。研究者通过数据增强等手段在受限的基准数据集上训练他们的新方法，并评估新方法的泛化能力(例如，在数据集 A 上进行训练，然后在数据集 B 上进行测试)。无论采用哪种方法，使用多个数据集的情况并不少见。

表 2 列出了最常用的数据集，并报告了每个数据集中的数据量、数据特征、数据集的创建方式及其最初的应用目的(例如，用于室内还是室外场景)。除了表 2 中列出的数据集外，还开发了许多其他用于单目深度估计的数据集，包括 ETH3D[17]，DIW[18]，TUM-RGBD[19]和 Sintel[20]。这些数据集为研究人员提供了多样化的训练和评估资源，从而推动单目深度估计方法的发展。

2.2　单目图像深度估计的深度学习方法总结

1)单目深度估计网络有监督学习方法总结

监督学习方法利用标记图像进行训练，通常需要成对的图像和对应的深度图。真实深度值 (D) 与对应的估计深度图($\hat{D}$) 之间的差异是监督学习方法的主要特征。作为一个回归问题，这些深度神经网络的目标就是从单个深度图像中学习深度预测规律，真实深度信息 d_i 通常基于传感器的方法获取。在监督学习中，设置一些简单的损失函数，如 L_2 损失，类似于上述的平方相对误差，并在优化过程中减小该损失，如式(1)所示。

$$L_2(D,\hat{D})=\frac{1}{N}\sum_{i=1}^{N}(d_i-\hat{d}_i)^2 \tag{1}$$

在早期的研究中，监督学习方法将单目深度估计问题视为一个典型的判别式任务。Eigen 等[34]引入了一种使用单张图像进行深度预测的技术，首次将 CNN 结构应用于这一领域，设计了一个端到端的有监督单目图像深度估计网络，结合了全局和局部的细化，重点处理尺度不变的损失。通过引入表面法线并将深度图扩展到原始输入分辨率，从而提高网络性能。

作为有监督学习方法，训练过程需要大量标注数据。然而，通过激光雷达或 RGB-D 获取的数据存在噪声且相对稀疏，激光与照相机的投影中心不重合，以及不同训练集的不同标注方式都使得模型泛化更为困难。随着深度学习的发展，研究者们逐渐意识到单一数据集的局限性和不同标注方式带来的泛化问题。Ranftl 等[35]介绍了实现多样化数据集融合的方法MiDaS，用于全面的单目深度估计。

表 2 深度估计研究最常用的数据集

Tab. 2 Datasets most commonly used for the depth estimation research

数据集名称	数据特征	数据来源	场景类型	官方网站
KITTI[21]	图像:389 对立体图像和光流图(1 242×375),深度测量占图像像素的 4.1% 里程计:39.2 km 长度的立体视觉里程计序列 标注数据:超过 200 k 个 3D 物体标注	传感器组	室外	https://www.cvlibs.net/datasets/kitti/
NYU-v2[22]	图像:1 449 张 RGBD 图像(640×480)	深度相机	室内	https://cs.nyu.edu/~silberman/datasets/nyu_depth_v2.html
DIODE[23]	图像:由高分辨率 RGB-D 图像构成 训练集由 8 574 个室内样本和 16 884 个室外样本组成 验证集包含 325 个室内样本和 446 个室外样本	传感器组	室内、室外	https://diode-dataset.org/
Make3D[24-26]	图像:534 张单目图像(2 272×1 704)和对应深度(55×305)	激光雷达	室外	http://make3d.cs.cornell.edu/data.html
Cityspaces[27]	图像:立体视频序列(2 048×1 024) 标注数据:5 000 个具有高质量像素级标注,20 000个具有粗标注	立体相机	室外	https://www.cityscapes-dataset.com/
HRWSI[28]	图像:来自 Web 立体图像的大约 21 k 不同的高分辨率 RGBD 图像对 提供天空分割模板、实例分割模板以及无效像素蒙版	网络照片	室内、室外	https://github.com/KexianHust/Structure-Guided-Ranking-Loss
BlendedMVS[29]	图像:包含 17 000 张高分辨率图像,覆盖各种场景	Mesh 模型渲染	室内、室外	https://github.com/YoYo000/BlendedMVS
MegaDepth[30]	图像:130 000 张图像-深度图对 196 个利用 COLMAP 重建的 3D 位置	网络照片	室外	https://www.cs.cornell.edu/projects/megadepth/
DDAD[31]	图像:包含单目视频,150 个场景,3 080 张图像 具有 360°视角,最远可达 250 m,精度为亚厘米级 包括 200 张图像的全景标签	激光雷达	室外	https://github.com/TRI-ML/DDAD
SYNS[32]	图像:包含对齐的图像和来自 92 个场景的 LiDAR 全景图 密集 LiDAR 地图,覆盖率为 78.3%,每张图像包含 365 000 个点	激光雷达	室内、室外	https://syns.soton.ac.uk/
SYNS-Patches[33]	SYNS 数据集的子集	激光雷达	室内、室外	https://github.com/jspenmar/monodepth benchmark

他们提倡有原则的多目标学习,从而促进深度范围和尺度变化的稳健性。该方法的重点在于对编码器进行额外任务的预训练,利用 5 个训练数据集和 3D 电影数据,并通过整合另外 5 个数据集进一步扩展他们的方法[36]。该架构名为“密集预测变换器”,是一个新型的 Transformer 架构,并与全卷积模型进行了对比。在此基础上,作者改进了原有架构,提出了新的基于视觉 Transformer 的方法 DPT (dense prediction transformer)[37],相较于 MiDaS,DPT 在性能和泛化性上都得到了提升。

Bhat 等[38]提出了一个两阶段的框架 ZoeDepth,使用一个通用的编码-解码器架构进行相对深度估计的预训练,在第二阶段添加绝对深度估计的轻量级检测头,并使用绝对深度数据集进行微调。ZoeDepth 是第一个结合了相对深度和绝对深度的方法,有效弥补了相对深度估计和绝对深度估计之间的差距,通过在更多的数据集上定义更细化的域,并在更多的绝对深度数据集进行微调以进一步改进网络性能。

Feldmann 等[39]提出了使用 NeRF 进行数据增

强。单目深度估算(monocular depth estimation, MDE)模型的功能受限于是否能获得充足且多样化的数据集。对于应用于自动驾驶领域的 MDE 模型,这一问题因捕获数据轨迹的线性而更加严重。因此,设计了基于 NeRF 的数据增强管道,将具有更多样化观察方向的合成数据引入训练数据集,并展示了其在模型性能和鲁棒性方面的提升。

在生成式深度学习框架的推动下,单目深度估计开始向生成式模型转变。Ke 等[40]提出了将单目深度估计任务定义为一个条件去噪扩散生成任务。具体地,开发了 Marigold 模型,这是一个基于潜在扩散模型的框架,用于建模条件分布 $D(d\mid x)$,其中 d 代表深度,x 是给定的 RGB 图像。通过逐步添加和移除噪声,模型能够从噪声数据中恢复出清晰的深度信息。在训练过程中,模型通过对抗性学习调整参数,以最小化去噪扩散目标函数,从而有效地从输入的 RGB 图像中预测出对应的深度图。这种生成式框架不仅提升了深度估计的准确性,也增强了模型对新领域数据的泛化能力。

Fu 等[41]进一步扩展了生成式框架的应用,提出了基于扩散模型的深度估计网络 GeoWizard。在微调过程中,首先利用稳定扩散 VAE 编码器,将输入图像及其对应的真实深度图、真实法线图映射到潜在空间,获得各自的潜在编码。然后,将两种几何潜在编码与图像潜在编码连接,形成两个组。这些组被输入到 U-Net 中,在几何切换器的引导下生成深度或法线域中的输出。此外,引入场景提示,用于生成具有室内、室外或对象 3 种场景布局的结果。在推理过程中,给定图像和场景提示,通过初始深度噪声和法线噪声,GeoWizard 可以联合生成高质量的深度和法线。

单目深度估计的有监督学习方法经历了从判别式方法到生成式框架的演变。生成式模型通过建模数据分布、多任务学习、数据增强和不确定性处理,显著提升了单目深度估计的泛化能力。与传统判别式方法相比,生成式模型不仅可以更好地理解数据多样性,还能生成类似训练数据的样本,从而扩展数据集多样性。此外,其强大的上下文理解和处理复杂场景的能力,使模型能够在不同环境下实现更稳健的深度预测。

2)单目深度估计网络自监督学习方法总结

由于图像的深度信息并不总能获得,因此很多时候监督学习方法并不能使用。自监督学习方法就是在损失函数的指导下,在图像之间学习图像重建规则。自监督深度估计方法主要通过双目视差或时序约束构建损失函数,避免了对深度真值的依赖。然而,在无纹理区域的深度预测仍具挑战性。

首先介绍了利用双目图像左右视差约束的方法。Garg 等[42]提出了基于 AE 的方法,其中编码器(CNN)负责预测深度,解码器通过逆扭曲重建图像。该方法通过最小化图像重建误差学习深度,无需深度标注。但在无纹理区域表现欠佳,推动了后续研究致力于解决此问题,如探索额外的几何约束。

Aleotti 等[43]介绍了利用 GAN 的方法,通过使用立体图像对进行深度估计。这个 GAN 包括一个生成新图像的生成器和一个经过训练以区分真实的和生成的(扭曲的)图像的判别器。

Wang 等[44]提出了统一处理光流和立体深度估计的方法,使用 3 个轻量级 CNN 分别处理深度、相机运动和光流。引入刚性感知直接视觉里程计模块,改善相机运动估计。但该方法面临多任务平衡和动态场景处理的挑战。

Godard 等[45]提出了一种 ED 网络 MonoDepth,通过左右图像的相互计算推导视差。该方法在损失函数中引入了左右一致性项、平滑项和重建项。之后,Godard 等[46]进一步改进了从视频或立体数据中进行单目深度估计的方法,包括处理遮挡的每像素最小光度误差、用于检测静止像素的自动掩码以及减少伪影的全分辨率多尺度上采样技术。

另一种可利用的自监督方式是利用单目图像序列或者双目图像序列提取帧间的几何约束。设 p_t 是目标视图 I_t 中像素的齐次坐标,则 p_t 可以由源视图 I_s 中的点 p_s 投影,如式(2)所示。

$$p_s \sim \boldsymbol{K}\boldsymbol{T}_{t\to s}D_t(p_t)\boldsymbol{K}^{-1}p_t \tag{2}$$

其中,$\boldsymbol{K}$ 为相机内参矩阵;$D_t(p_t)$ 为 p_t 的深度值;$\boldsymbol{T}_{t\to s}$ 为帧 I_t 到帧 I_s 的齐次变换矩阵。通常情况下,$D_t(p_t)$ 和 $\boldsymbol{T}_{t\to s}$ 都是未知的,需要通过计算得到,如式(3)所示。

$$p_s \sim \boldsymbol{K}\hat{\boldsymbol{T}}_{t\to s}\hat{D}_t(p_t)\boldsymbol{K}^{-1}p_t \tag{3}$$

Kumar 等[47]介绍了一种基于 GAN 的方法,通过使用视频序列进行深度和姿态估计。生成器包括深度和姿态网络,预测连续帧的深度图和旋转/平移参数。从当前图像预测深度,并从先前、当前和下一帧的三元组中估计姿势。该算法将真实的图像与生成的图像区分开来,从而产生了一个包含图像一致性、对抗性和深度项的三项损失函数,用

于网络训练。

3)单目深度估计网络半监督学习方法总结

另一种可以利用的方法是半监督学习方法,通过利用一小部分标注数据和大部分未标注的数据训练模型。因此,进行数据增强与辅助监督非常重要。

Zhao 等[48]通过图像风格转换和深度估计模块,将合成数据融入到他们的单目深度估计方法 GASDA 中。基于 CycleGAN[49]的图像风格可以在真实图像和合成图像之间进行转换。这种合成数据的集成为无监督方法提供了原本无法捕获的深度信息。

Depth Anything[50]使用少量已标注深度的图像数据训练教师模型,然后利用数据引擎收集的大规模无标注图像数据,通过教师模型得到伪深度标签。将伪深度作为无标注图像的优化目标,引入无标注图像到伪深度的学生模型。在这个过程中,学生模型通过数据增强和语义分割的预训练编码器学习到更丰富的先验知识,提升了模型的学习能力。作者给出了 3 种不同大小的模型,可以在边缘设备上运行。

Gui 等[51]提出了 DepthFM,是一种用于单目深度估计的流匹配方法,通过计算回归图像和深度分布之间的向量场,实现高效推理。仅在 63k 纯合成样本上训练的 DepthFM,在室内外数据集上实现了 Zero-shot 深度估计,并通过辅助表面法线损失提升了深度估计的准确性。这种方法在节省计算资源的同时,强调模型的适应性和训练效率,可提供细粒度的深度图和可靠的置信度估计。

表 3 总结了今年以前效果较好的 3 种类型的单目深度估计网络,并展示了它们的性能表现("—"表示原论文中未提到)。

表 3 深度学习网络深度估计精度总结

Tab. 3 Summary of deep learning network depth estimation accuracy

年份	方法	训练样本		NYUv2		KITTI		DIODE	
		真实数据	合成数据	AbsRel	δ_1/% ↑	AbsRel	δ_1/% ↑	AbsRel	δ_1/% ↑
2014	Eigen 等[34]监督学习	2 M	0	0.215	61.1	0.190	69.2	—	—
2017	MonoDepth[45]自监督学习	29 k	0	—	—	0.148	80.3	—	—
2019	GASDA[48]半监督学习	22.6 k	21.26 k	—	—	0.149	82.4	—	—
2019	MonoDepth2[46]自监督学习	40 k	0	—	—	0.106	87.6	—	—
2020	MiDaS[35]监督学习	2 M	0	0.111	88.5	0.236	63.0	0.332	71.5
2021	Omnidata[52]监督学习	11.9 M	301 k	0.074	94.5	0.149	83.5	0.339	74.2
2021	DPT[37]监督学习	1.2 M	188 k	0.098	90.3	0.100	90.1	0.182	75.8
2022	HDN[53]半监督学习	300 k	0	0.069	94.8	0.115	86.7	0.246	78.0
2023	ZoeDepth[38]监督学习	200 k	—	—	95.5	—	—	—	—
2024	Marigold[40]监督学习	0	74 k	0.061	94.9	0.099	91.6	0.275	78.5
2024	Depth Anything[50]半监督学习	1.5 M labeled	(small)	0.053	97.2	0.080	93.6	—	—
			(base)	0.046	97.9	0.080	93.9	—	—
		62 M unlabeled	(big)	0.043	98.1	0.076	94.7	0.277	75.9
2024	GeoWizard[41]监督学习	0.28 M	0	0.052	96.6	0.097	92.1	0.297	79.2
2024	DepthFM[51]半监督学习	0	63 k	0.065	95.6	0.083	93.4	0.225	80.0

2.3 SLAM 评估常用指标

视觉里程计(visual odometry, VO)/视觉惯性里程计(visual-inertial odometry, VIO)是机器人领域感知模块的重要组成部分,也是自主导航与定位中不可或缺的一环。近期,该领域取得了新的进展,涌现出深度学习混合 VO 的方法,这些方法结合了单目深度估计网络和传统的姿态估计方法。根据两者结合的紧密程度,可以分为单目深度网络提供的信息辅助位姿估计方法和单目深度网络与位姿估计方法相互监督的方法。其中,单目深度估计网络与位姿估计方法相互监督的方法是对深度估计的自监督学习方法,正如 2.2 节所述,这种方法通过引入来自位姿估计的约束,从而提升里程计的精度;而在单目深度网络提供的信息辅助位姿估计方法中,会通过进行一些筛选从而提高定位精度。

常用的 SLAM 系统评估指标包括三维重建、二

维深度估计以及轨迹估计等指标，能够评估与真实值之间的差异。

姿态估计的评估指标包括：绝对轨迹误差(absolute trajectory error，ATE)，用于衡量 SLAM 系统估计的轨迹与实际路径之间的偏差。通过计算估计位置和真实位置的欧氏距离以量化定位精度，低 ATE 值表明系统能更准确地估计自身运动。相对姿态误差(relative pose error，RPE)则用于评估相邻帧之间的相对运动估计误差，通过比较估计的相对运动与真实相对运动之间的差异衡量局部精度，常用于评估短时间内的累积误差，特别适用于动态环境。

2.4　图像深度估计与 SLAM 结合方法总结

上述有关图像深度的自监督学习中也提到了利用相邻帧之间的帧间位姿约束优化深度，这种方法的重点在于深度估计，在训练完成之后位姿估计网络就不再使用，即不再关注定位问题。以下介绍的都是深度估计与 SLAM 相结合的方法，以及其对定位精度或深度预测精度的影响。

Liu 等[54]提出了一种新型的 VO 方法，使用比真实深度图更容易获得的真实姿态监督网络，从而获得密集的深度图。从神经网络获得的深度图用于将当前图像扭曲到参考帧中，并通过最小化对扭曲图像和参考图像之间的相似性进行编码的成本函数以获得最佳位姿估计。

Yang 等[55]提出了一种新颖的框架 D3VO，用于单目 VO，在 3 个层面上利用深度网络：深度估计、位姿估计和不确定性估计。具体而言，作者设计了一个自监督的单目深度估计网络，通过立体视频进行训练，并采用亮度变换参数对齐训练图像，从而提高了深度估计的准确性。D3VO 将预测的深度、位姿和光度不确定性紧密集成，显著提升了前端跟踪和后端非线性优化的性能。此外，作者还通过光度不确定性建模，优化了图像匹配过程，提高了在光照变化和动态场景中的鲁棒性。实验结果表明，D3VO 在 KITTI 和 EuRoC MAV 数据集上表现出色，达到了与最先进的立体/激光雷达方法相当的效果，并在 VIO 任务中的表现也接近最先进水平。

Almalioglu 等[56]提出了 SelfVIO，将深度估计与 VIO 有效结合。SelfVIO 利用自监督学习方法，从未标记的 RGB 图像序列和惯性测量单元(inertial measurement unit，IMU)读数中联合估计 6 自由度(6-DoF)相机运动和场景深度图。该方法采用 GAN 进行对抗训练，通过对比生成的图像和真实图像以优化深度图和相机位姿的预测。此外，SelfVIO 引入了自适应传感器融合技术，将 IMU 数据和 RGB 图像数据相结合，提升了运动估计和深度预测的准确性。与传统方法不同，SelfVIO 不需要 IMU 与相机之间的精确校准，而是通过自适应调整传感器数据解决校准问题。通过这种迭代优化过程，SelfVIO 实现了深度估计和 VIO 任务的相互促进，显著提升了整体性能。

Sun 等[57]将单目深度估计融入 VO 系统，提出了一个结合几何基础 VO 和学习基础单目深度估计网络的框架。该框架包含两种工作模式：第一种模式是利用单张 RGB 图像输出相对深度图以提高定位准确性；第二种模式则结合 RGB 图像和稀疏深度图生成尺度一致的稠密深度图，用于实现高精度的稠密映射。在 VO 系统中，作者通过最小化重投影误差获得相机姿态，并引入了近远一致性约束以剔除错误深度点，增强了定位的准确性。尺度恢复通过将稀疏点投影到关键帧的图像平面，并使用中位数尺度以简化尺度调整。最终，通过学习获得的深度被用于稠密深度映射，通过像素级的强度和深度一致性检查以提升映射质量，使得该系统能够在多样化的场景中表现出更好的鲁棒性和准确性。

Xing 等[58]提出了一种新方法，将经典视觉 SLAM 与基于 CNN 的单目深度估计相结合，以提升三维重建和位姿估计的性能。该方法通过利用 CNN 预测的深度图改善伪 RGB-D SLAM 的 VO 和地图构建的准确性，同时使用 SLAM 获得的 3D 场景结构对预训练的深度估计网络进行微调，解决了单目深度估计在强光照条件下不准确的问题。实验结果表明，所提出的方法在 KITTI 和 TUM-RGBD 数据集上的深度预测和位姿估计任务中表现出优越的性能，相较于传统方法有了显著提升。该方法通过在深度估计网络中引入稀疏辅助网络，并迭代优化，可成功应对单目 SLAM 和深度估计中面临的多个挑战。

引入深度估计网络不仅能够提供额外的几何约束，更能从根本上提升系统的鲁棒性和精度。首先，深度信息可以作为特征点选取的重要参考依据，通过对场景深度的理解，系统能够更智能地在具有显著几何结构的区域分布特征点，同时避免在深度不连续或不可靠的区域提取特征，从而构建出更加稳定的 VO。特别值得注意的是，现代深度估

计网络通过自监督学习等先进训练范式，已经在复杂场景中展现出出色的性能，能够在光照变化、存在动态物体及反光表面等充满挑战的环境中提供可靠的深度预测，为增强 SLAM 系统在实际应用场景中的适应性提供了有力支持。

深度信息在 SLAM 系统的多个核心模块中都具有巨大的应用潜力。在回环检测方面，结合深度的场景描述子能够捕获环境的三维结构特征，在外观发生显著变化（如日夜更替、季节变换）的场景中仍能保持较高的识别准确率。在初始化阶段，深度估计网络提供的先验信息可以大幅提高系统的启动速度和稳定性，解决了传统单目 SLAM 系统在低纹理或静止场景下的初始化困难问题。此外，深度网络对远距离物体的感知能力，使得系统能够更好地利用环境中的远景信息。

3 NeRF 和 3DGS 与 SLAM 相结合的方法

3.1 NeRF 在 SLAM 中的应用方法总结

最初将两者结合的项目是 NICE-SLAM[59]，如图 4 所示，该框架直接利用图像损失优化网络参数和相机位姿，完全没有利用经典的特征点指导优化。绿色部分左侧展示了传感器收集的 RGB-D 真实数据，包括深度信息和 RGB 图像；右侧则显示了模型推断出的预测深度值和 RGB 图像。为优化这两部分之间的差异，采用了 NICE-SLAM 系统。经过训练，系统生成了右侧黄色区域内的分层特征网格和相机位姿，并将这两者输入到蓝色的可微渲染器中，从而完成对 RGB 图像和深度图像的预测。如果要使用 NICE-SLAM，必须获得深度图。这是因为在初始阶段，仅依靠 RGB 图像训练网络会导致收敛效果不佳。深度图中蕴含了光线传播方向上物体的先验概率分布，这种先验信息能够帮助网络更快地收敛。在此背景下，一个创新点在于：针对仅有 RGB 图像的情况，利用前文提到的深度估计网络可实现快速收敛。具体来说，可以利用深度估计网络，对相对深度进行估计，并将其融入到框架中。需要注意的是，通过单目图像进行深度估计时，得到的多张图像之间的深度一致性往往较差，需要添加正则项来一致化深度，可以通过式（4）来转化。

$$D_i^* = \alpha_i D_i + \beta_i$$

$$\boldsymbol{L}_{\text{depth}} = \sum_i^N \| D_i^* - \hat{D}_i \| \tag{4}$$

其中，$\hat{D}_i$ 是 NeRF 渲染得到的深度图；D_i 是单目图像深度估计得到的深度图；α_i 和 β_i 分别为比例修正项和偏移修正项，通过这种设计确保深度的一致性。还可以通过一些非深度学习的方法对深度估计的不确定性进行评估，并据此调整深度损失。另一个可行的创新点在于结合运动信息（IMU），以充分利用不同时刻的相机位姿。

此外，NeRF 与 SLAM 结合可以得到由 NeRF 构建的隐式地图，这种地图可以成为视觉导航的基础地图，有利于实现后续多样化的定位功能。

如图5 所示，RNR-Map[60] 是一种用于视觉导

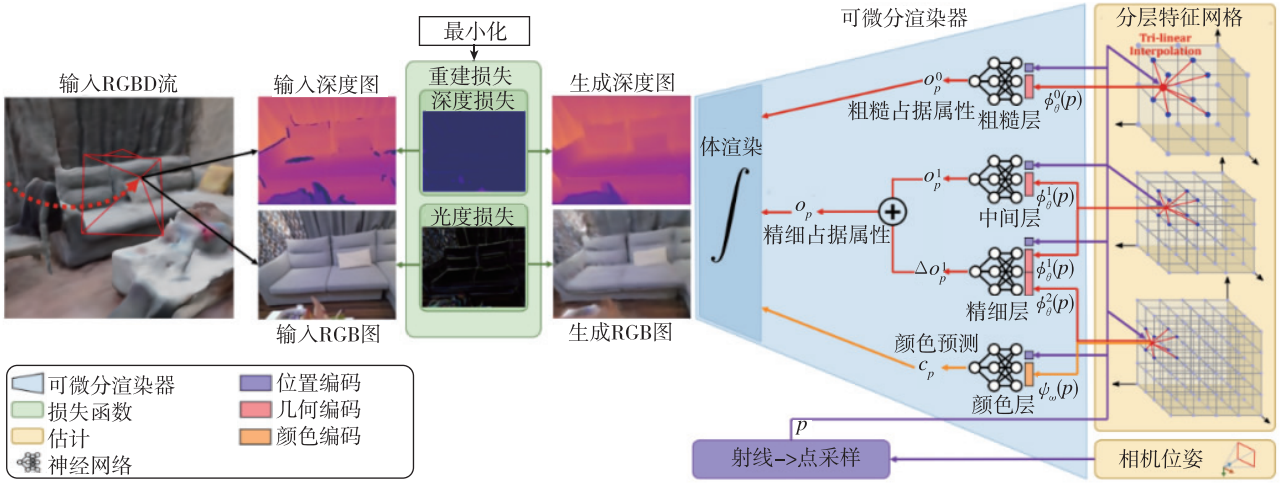
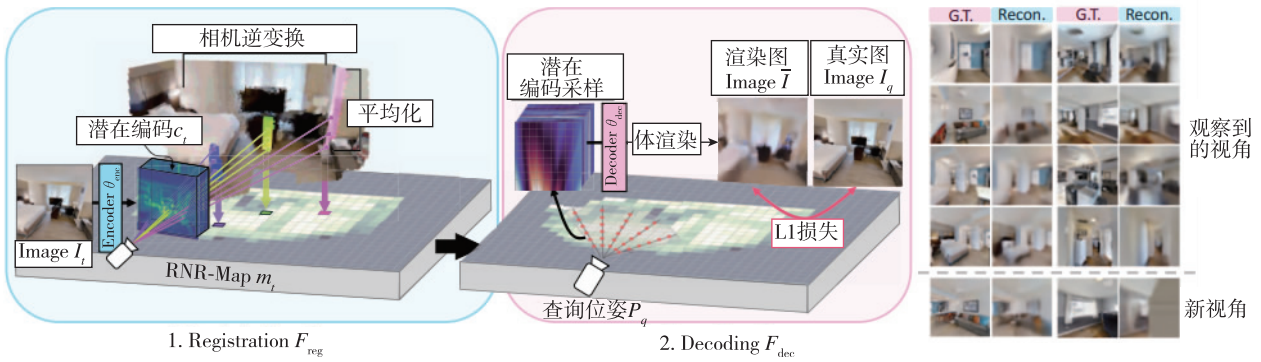


图 4 NICE-SLAM 系统概述

Fig. 4 NICE-SLAM system overview

航的基于 NeRF 的新型地图，构建时需要准确的里程计信息，确保其能包含 3D 环境的整体视觉信息。RNR-Map 具有网格形式，由每个像素处的潜在代码组成。这些潜在代码是从图像观察中嵌入的，可以转换为 NeRF，从而在给定相机姿势的情况下实现图像渲染。记录的潜在代码隐式地包含有关环境的视觉信息，这使 RNR-Map 具有视觉描述性。RNR-Map 中的这种视觉信息可以成为视觉定位和导航的有用指南。因此，利用隐式的 NeRF 进行渲染的一个好处就是，这部分网络的内存大小在一定范围内不会随着地图的扩大而增加，而且与深度学习提取的视觉描述子之间的衔接显得更加自然。



（a）重建框架　　（b）渲染图像示例

图 5　RNR-Map 系统概述

Fig. 5　RNR-Map system overview

考虑到经典算法维护的地图通常较为稀疏且轻量，因此存在一种工程解决方案 NeRFBridge[61]，如图 6所示，可有效地将传统里程计模块与 NeRF 建图模块相结合，以实现更优的性能表现。该系统通过搭载摄像头的机器人（如无人机）进行信号采集，采集到的图像数据流被传输到 NeRFBridge 节点，利用 ORB-SLAM3 实时估计相机位姿，随后通过 NeRFstudio 进行场景重建。无人机运行在树莓派 4B 平台上，负责操控摄像头并与 NeRFBridge 节点进行网络交互。同时，NeRFBridge 节点在一台配备 RTX 3090 的后台服务器上并行运行 ORB-SLAM3，NeRFstudio 及系统控制程序。

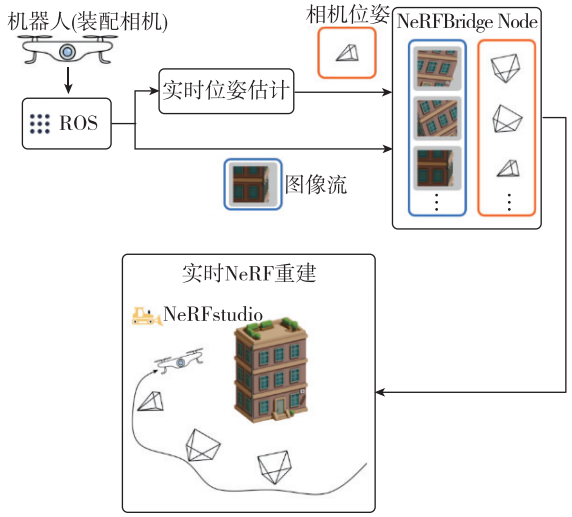


图 6　NeRFBridge 系统概述

Fig. 6　NeRFBridge system overview

除此之外，NeRF 可以存储占据信息。如图 7 所示，IR-MCL 框架[62]将 NeRF 这样的隐式环境表示扩展到移动机器人二维室内定位任务中，提出了一种神经占据场，使用神经网络隐式地表示，用于替代机器人定位任务中的二维地图。通过预训练的神经网络可以渲染合成当前场景下任意机器人姿态所对应的 2D 激光扫描。基于该隐式地图，提出了一个观测模型以计算渲染和真实扫描之间的相似性，并将其集成到蒙特卡罗定位系统中进行准确定位。

因此，可以通过构建一个由 NeRF 网络维护的地图，将其转化为传感器所需的数据格式以实现定位。例如，可以利用单目相机捕获的图像进行定位，同时也可以通过 NeRF 网络生成激光点云格式的数据。由于在给定视角下已知该视角的占据信息，进而可以利用点云进行匹配和定位。因此，从这一角度来看，NeRF 网络能够有效地应对不同模态间的定位挑战。

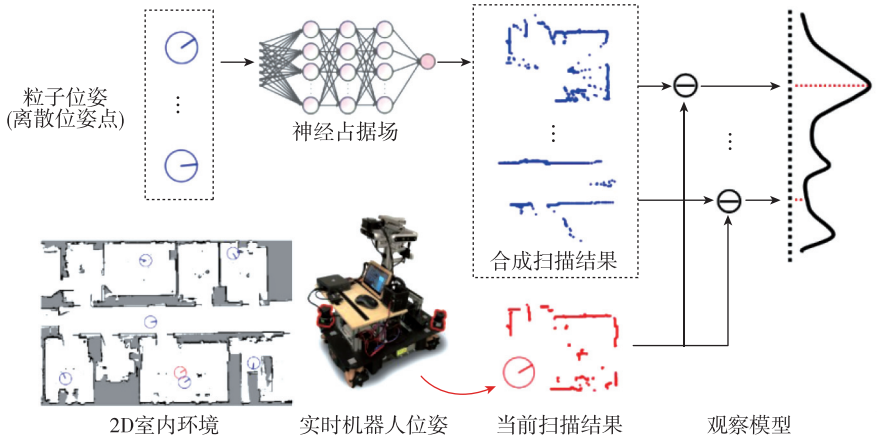


图 7 IR-MCL 系统概述

Fig. 7 IR-MCL system overview

3.2 3DGS 在 SLAM 中的应用方法总结

在 3DGS 方法中，也有一些针对同时定位和重建 3D 场景的 SLAM 方法。SuGaR[63] 根据当前高斯点和新帧捕获的深度优化几何位置。为了避免重复密化，SplaTAM[64] 利用视点独立的颜色判断新的帧是否需要密化。GS-SLAM[65] 提出了一种自适应的 3D 高斯扩展策略，通过捕捉的深度值添加新的 3D 高斯点，并删除旧的不准确高斯点。为了稳定定位和建图，Gaussian splatting SLAM[66]（见图 8）和 Gaussian-SLAM[67] 在高斯点的尺度上增加了额外的空间正则化损失，以鼓励各向同性高斯点。SemGauss-SLAM[68]、NEDS-SLAM[69] 和 SGS-SLAM[70] 进一步考虑高斯点在同时定位和建图过程中语义信息的重要性，通过蒸馏数据集提供的 2D 分割信息获取语义信息。CG-SLAM[71] 在训练过程中引入不确定性图，并基于渲染深度优化 3D 重建精度。Deng 等[72] 通过改进基于滑动窗口掩模和向量量化的方法，避免了冗余高斯分裂，从而进一步提高了 3DGS 的精度，系统概述如图 9 所示。该系统通过精确的场景重建和深度图生成，增强了对下游任务的支持，尤其是为复杂环境中的物体识别、语义标注及障碍物检测等任务提供了可靠的基础。此外，最近有研究者提出了 DepthSplat[73]，该方法利用稀疏视角的前馈 3DGS 与稳健的单目深度估计之间的互补性，提高了 3DGS 在稀疏视角下的表现，并增强了网络深度估计的一致性。

表 4 显示了不同 SLAM 方法在 TUM-RGBD 数据集和 Replica 数据集上的相机追踪精度的定量比较。

目前，基于 3DGS 和 SLAM 的技术仍在发展中，理论上 3DGS 具备更好地筛选物体的能力与增加物体的能力。例如通过判断动态物体的范围，可以筛选出属于动态物体的 3D 高斯球，从而在有动态物体干扰的场景下，依然能实现动态场景下的跟

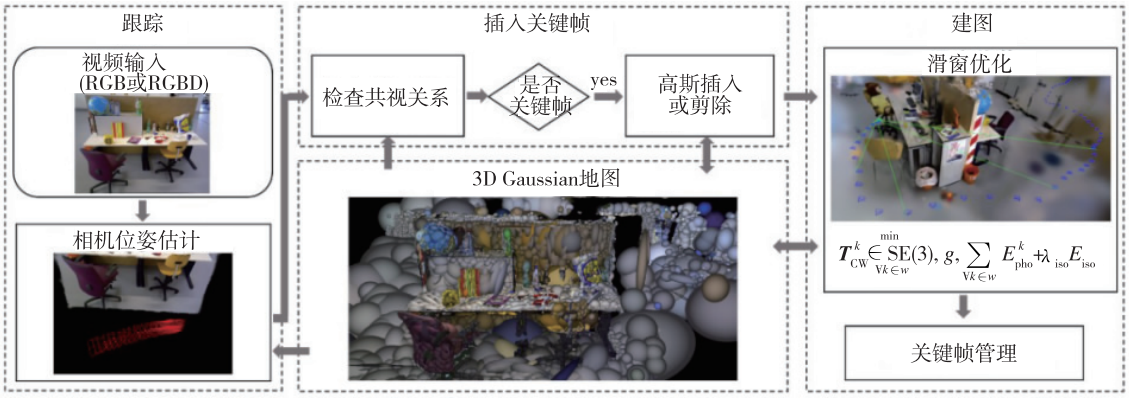


图 8 Gaussian splatting SLAM 系统概述

Fig. 8 Gaussian splatting SLAM system overview

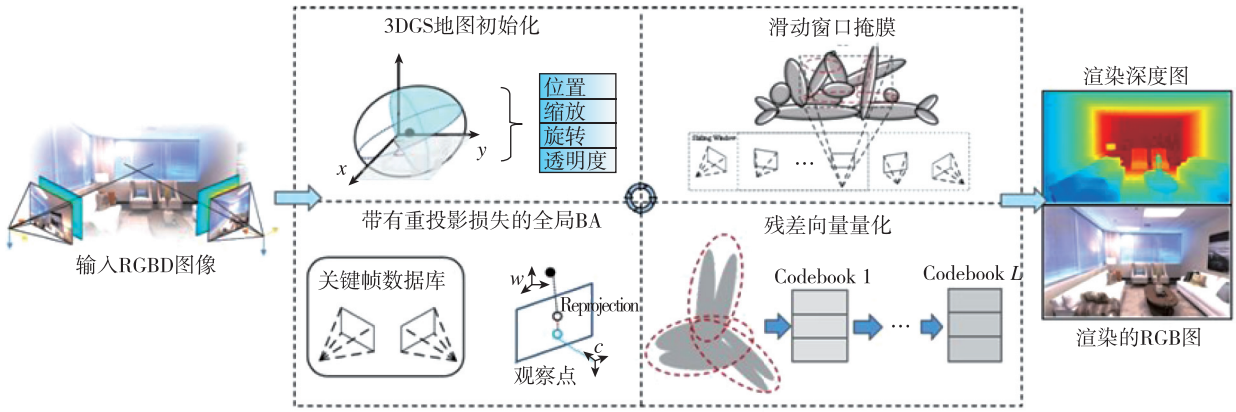


图 9　基于 GS 的 SLAM 系统概述[72]

Fig. 9　Compact 3D Gaussian splatting dense visual SLAM system overview[72]

表 4　3DGS SLAM 方法相机追踪精度

Tab. 4　Tracking accuracy of 3DGS SLAM method camera

3DGS SLAM 方法	TUM-RGBD (ATE RMSE/cm)	Replica (ATE RMSE/cm)
SplaTAM[64]	3.25	0.36
GS-SLAM[65]	3.7	0.5
Gaussian Splatting SLAM[66]	1.58	0.79
Gaussian-SLAM[67]	2.9	0.31

踪以及静态地图的构建[74]。此外，与 CNN 中不同卷积核大小对应不同特征提取能力类似，3DGS 球也具备不同大小的场景提取能力，即可以通过少量的大体积的高斯球构建粗略的场景，或者通过大量的小体积的高斯球构建精细的场景，这尤其有利于在保留大场景粗略结构的同时，保存精细物体的结构信息，具体地可以实现场景识别以及视觉导航。

4　结论

1）本研究系统性地探讨了基于深度学习的深度感知技术与 SLAM 的融合进展。在 SLAM 系统的感知和定位两个关键环节，深度感知方法都带来了重要改进：在感知端，深度网络提供的深度先验信息帮助系统更好地理解场景结构，特别是在低纹理区域的特征匹配；在定位端，NeRF 和 3DGS 的引入允许构造新的地图形式，增强了系统处理场景和匹配场景的能力。这些改进使得基于深度学习深度感知的 SLAM 系统能够更好地应对传统方法的局限性，如光照变化等复杂环境。

2）单目深度估计技术的引入为 SLAM 系统带来了新的可能性，特别是在缺乏立体视觉或深度传感器的场景中。无论是作为独立模块的直接深度估计网络，还是集成到 SLAM 流程中的间接深度估计网络，都能为系统提供有价值的深度信息。直接深度估计网络提供了灵活的集成方案，可与现有 SLAM 框架协同工作；而间接深度估计网络通过端到端优化，展现出更强的场景适应性。然而，这些方法在计算效率和实时性方面仍面临挑战，特别是在资源受限的设备上的应用仍需深入研究。

3）NeRF 和 3DGS 为 SLAM 系统提供了新的思路：首先，其场景表示方式比传统点云更加紧凑和结构化，有助于提高系统的运行效率；其次，这种新型的场景表示方式为后续的场景理解和语义分析提供了更丰富的信息基础。

4）尽管基于深度学习的深度感知技术在 SLAM 应用中取得了显著进展，但仍存在一些亟待解决的技术难题：①实时性与精度的权衡：如何在保证高精度的同时提高系统的实时性，特别是在处理大规模环境时；②长期运行的稳定性：如何有效解决累积误差和闭环检测问题，确保系统在长时间运行中的稳定性；③多传感器融合：如何高效整合 IMU，GPS 等多源数据，以提高系统的鲁棒性和适应性；④动态环境处理：如何准确识别和处理动态物体，最小化其对 SLAM 系统的干扰。

5）未来研究方向可能包括：①开发计算效率更高的深度估计算法；②优化 3DGS 技术，如引入图神经网络等先进结构，以提高其表达能力和效率；③拓展 NeRF 隐式地图的应用场景，如研究在大规模室外环境、极端光照条件下的 SLAM 解决方案；④深化 SLAM 与高级计算机视觉任务的结合，如语义 SLAM、实例级 SLAM 等，实现更深层次的场景理解；⑤探索在线学习和增量式学习方法，提高

SLAM 系统的长期自主性和环境适应能力。

深度感知技术的引入为 SLAM 领域注入了新的活力，不仅显著提升了系统性能，还拓展了其应用边界。随着这些技术的持续演进，预见 SLAM 系统将在精度、鲁棒性和泛化能力等方面取得突破性进展。未来，SLAM 技术有望在自主导航、空间感知及环境理解等领域发挥更为关键的作用，推动智能系统的革新。然而，要充分释放这些技术的潜力，研究者们仍需在算法优化、硬件适配及大规模数据处理等方面进行深入探索，以克服现有的技术瓶颈，推动 SLAM 技术向更高水平迈进。

参考文献

[1] MUR-ARTAL R, MONTIEL J M M, TARDÓS J D. ORB-SLAM: a versatile and accurate monocular SLAM system[J]. IEEE Transactions on Robotics, 2015, 31(5): 1147-1163.

[2] HESS W, KOHLER D, RAPP H, et al. Real-time loop closure in 2D LIDAR SLAM[C]// Proceedings of 2016 IEEE International Conference on Robotics and Automation (ICRA). Stockholm: IEEE, 2016: 1271-1278.

[3] MILDENHALL B, SRINIVASAN P P, TANCIK M, et al. NeRF: representing scenes as neural radiance fields for view synthesis[C]// Proceedings of European Conference on Computer Vision (ECCV). Glasgow: Springer, 2020, 12346: 405-421.

[4] KERBL B, KOPANAS G, LEIMKÜHLER T, et al. 3D Gaussian splatting for real-time radiance field rendering[J]. ACM Transactions on Graphics, 2023, 42(4): 139.

[5] ZHANG T, HU X, XIAO J, et al. A survey of visual navigation: from geometry to embodied AI[J]. Engineering Applications of Artificial Intelligence, 2022, 114: 105036.

[6] BHAT S F, BIRKL, R, WOFK D, et al. ZoeDepth: zero-shot transfer by combining relative and metric depth[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver: IEEE, 2023.

[7] SAYED M, ALEOTTI F, WATSON J, et al. DoubleTake: geometry guided depth estimation[J]. arXiv preprint arXiv: 2406.18387, 2024.

[8] YU Z, PENG S, NIEMEYER M, et al. MonoSDF: exploring monocular geometric cues for neural implicit surface reconstruction[C]// Proceedings of 36th International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc., 2022: 25018-25032.

[9] WARBURG F, WEBER E, TANCIK M, et al. NeRFBusters: removing ghostly artifacts from casually captured NeRFs[C]// Proceedings of IEEE/CVF International Conference on Computer Vision. Paris: IEEE, 2023: 18120-18130.

[10] GOLI L, READING C, SELLÁN S, et al. Bayes' rays: uncertainty quantification for neural radiance fields[C]// Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2024: 20061-20070.

[11] CAO J, WANG H, CHEMERYS P, et al. Real-time neural light field on mobile devices[C]// Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023: 8328-8337.

[12] ZHANG C, ZHOU Y, ZHANG L. Vosh: voxel-mesh hybrid representation for real-time view synthesis[J]. arXiv preprint arXiv: 2403.06505, 2024.

[13] KATO Y, TARASHIMA S. Plug-and-play acceleration of occupancy grid-based NeRF rendering using VDB grid and hierarchical ray traversal[J]. arXiv preprint arXiv: 2404.10272, 2024.

[14] CHEN Z, FUNKHOUSER T A, HEDMAN P, et al. MobileNeRF: exploiting the polygon rasterization pipeline for efficient neural field rendering on mobile architectures[C]// Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023: 16569-16578.

[15] 黄军，王聪，刘越，等. 单目深度估计技术进展综述[J]. 中国图象图形学报，2019，24(12)：2081-2097.
HUANG Jun, WANG Cong, LIU Yue, et al. The progress of monocular depth estimation technology [J]. Journal of Image and Graphics, 2019, 24(12): 2081-2097(in Chinese).

[16] GUI M, FISCHER J S, PRESTEL U, et al. DepthFM: fast monocular depth estimation with flow matching[J]. arXiv preprint arXiv: 2403.13788, 2024.

[17] SCHÖPS T, SCHÖNBERGER J L, GALLIANI S, et al. A multi-view stereo benchmark with high-resolution images and multi-camera videos[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017: 3260-3269.

[18] CHEN W, FU Z, YANG D, et al. Single-image depth perception in the wild[C]// Proceedings of 30th International Conference on Neural Information Processing Systems. Barcelona: Curran Associates Inc., 2016: 730-738.

[19] STURM J, ENGELHARD N, ENDRES F, et al. A benchmark for the evaluation of RGB-D SLAM systems[C]// Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems. Vilamoura: IEEE, 2012: 573-580.

[20] BUTLER D J, WULFF J, STANLEY G B, et al. A naturalistic open source movie for optical flow evaluation [C]// Proceedings of 12th European Conference on Computer Vision. Florence: Springer, 2012: 611-625.

[21] GEIGER A, LENZ P, URTASUN R. Are we ready for autonomous driving? The KITTI vision benchmark suite [C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Madison: IEEE, 2012: 3354-3361.

[22] SILBERMAN N, HOIEM D, KOHLI P, et al. Indoor segmentation and support inference from RGBD images[C]// Proceedings of European Conference on Computer Vision. Florence: Springer, 2012: 746-760.

[23] VASILJEVIC I, KOLKIN N, ZHANG S, et al. DIODE: a dense indoor and outdoor depth dataset[J]. arXiv preprint arXiv: 1908.00463, 2019.

[24] SAXENA A, SUN M, NG A Y. Make3D: learning 3D scene structure from a single still image[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2009, 31(5): 824-840.

[25] SAXENA A, CHUNG S H, NG A Y, et al. Learning depth from single monocular images[C]// Proceedings of 18th International Conference on Neural Information Processing Systems. Cambridge: MIT Press, 2005: 1-8.

[26] SAXENA A, CHUNG S H, NG A Y. 3-D depth reconstruction from a single still image[J]. International Journal of Computer Vision, 2008, 76(1): 53-69.

[27] CORDTS M, OMRAN M, RAMOS S, et al. The Cityscapes dataset for semantic urban scene understanding [C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016: 3213-3223.

[28] XIAN K, ZHANG J, WANG O, et al. Structure-guided ranking loss for single image depth prediction[C]// Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020: 611-620.

[29] YAO Y, LUO Z, LI S, et al. BlendedMVS: a large-scale dataset for generalized multi-view stereo networks [C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020: 1787-1796.

[30] LI Z, SNAVELY N. MegaDepth: learning single-view depth prediction from internet photos[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 2041-2050.

[31] GUIZILINI V, AMBRUS R, PILLAI S, et al. 3D packing for self-supervised monocular depth estimation [C]// Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020: 2485-2494.

[32] ADAMS W J, ELDER J H, GRAF E W, et al. The Southampton-York natural scenes (SYNS) dataset: statistics of surface attitude[J]. Scientific Reports, 2016, 6(1): 35805.

[33] SPENCER J, RUSSELL C, HADFIELD S, et al. Deconstructing self-supervised monocular reconstruction: the design decisions that matter[J]. arXiv preprint arXiv: 2208.01489, 2022.

[34] EIGEN D, PUHRSCH C, FERGUS R. Depth map prediction from a single image using a multi-scale deep network[C]// Proceedings of International Conference on Neural Information Processing Systems. Montreal: Springer, 2014: 2366-2374.

[35] RANFTL R, LASINGER K, HAFNER D, et al. Towards robust monocular depth estimation: mixing datasets for zero-shot cross-dataset transfer[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(3): 1623-1637.

[36] RANFTL R, BOCHKOVSKIY A, KOLTUN V. Vision transformers for dense prediction[C]// Proceedings of IEEE/CVF International Conference on Computer and Vision. Montreal: IEEE, 2021: 12179-12188.

[37] RANFTL R, BOCHKOVSKIY A, KOLTUN V. Vision transformers for dense prediction[C]// Proceedings of IEEE/CVF International Conference on Computer Vision. Montreal: IEEE, 2021: 12159-12168.

[38] BHAT S F, BIRKL R, WOFK D, et al. ZoeDepth: zero-shot transfer by combining relative and metric depth[J]. arXiv preprint arXiv: 2302.12288, 2023.

[39] FELDMANN C, SIEGENHEIM N, HARS N, et al. NeRFmentation: NeRF-based augmentation for monocular depth estimation[J]. arXiv preprint arXiv: 2401.03771, 2024.

[40] KE B, OBUKHOV A, HUANG S, et al. Repurposing diffusion-based image generators for monocular depth estimation[C]// Proceedings of IEEE/CVF International Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2024.

[41] FU X, YIN W, HU M, et al. GeoWizard: unleashing the diffusion priors for 3D geometry estimation from a single image[J]. arXiv preprint arXiv: 2403.12013, 2024.

[42] GARG R, BG V K, CARNEIRO G, et al. Unsupervised CNN for single view depth estimation: geometry to the rescue[C]// Proceedings of European Conference on Computer Vision. Amsterdam: Springer, 2016: 740-756.

[43] ALEOTTI F, TOSI F, POGGI M, et al. Generative adversarial networks for unsupervised monocular depth prediction[C]// Proceedings of European Conference on Computer Vision. Munich: Springer, 2018: 337-354.

[44] WANG Y, WANG P, YANG Z, et al. UnOS: unified unsupervised optical-flow and stereo-depth estimation by watching videos[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019: 8071-8081.

[45] GODARD C, AODHA O M, BROSTOW G J. Unsupervised monocular depth estimation with left-right consistency[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017: 270-279.

[46] GODARD C, AODHA O M, FIRMAN M, et al. Digging into self-supervised monocular depth estimation [C]// Proceedings of IEEE/CVF International Conference on Computer Vision. Seoul: IEEE, 2019: 3828-3838.

[47] KUMAR A C S, BHANDARKAR S M, PRASAD M. Monocular depth prediction using generative adversarial networks[C]// Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Lake City: IEEE, 2018: 300-308.

[48] ZHAO S, FU H, GONG M, et al. Geometry-aware symmetric domain adaptation for monocular depth estimation[C]// Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019: 9788-9798.

[49] ZHU J Y, PARK T, ISOLA P, et al. Unpaired image-to-image translation using cycle-consistent adversarial networks[C]// Proceedings of IEEE International Conference on Computer Vision. Venice: IEEE, 2017: 2223-2232.

[50] YANG L, KANG B, HUANG Z, et al. Depth anything: unleashing the power of large-scale unlabeled data[C]// Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2024.

[51] GUI M, FISCHER J S, PRESTEL U, et al. DepthFM: fast monocular depth estimation with flow matching[J]. arXiv preprint arXiv: 2403.13788, 2024.

[52] EFTEKHAR A, SAX A, MALIK J, et al. Omnidata: a scalable pipeline for making multi-task mid-level vision datasets from 3D scans[C]// Proceedings of IEEE/CVF International Conference on Computer Vision. Montreal: IEEE, 2021: 10786-10796.

[53] ZHANG C, YIN W, WANG B, et al. Hierarchical normalization for robust monocular depth estimation [C]// Proceedings of International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc., 2022.

[54] LIU Z, MALIS E, MARTINET P. A new dense hybrid stereo visual odometry approach[C]// Proceedings of 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Kyoto: IEEE, 2022: 6998-7003.

[55] YANG N, VON STUMBERG L, WANG R, et al. D3VO: deep depth, deep pose and deep uncertainty for monocular visual odometry[C]// Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle: IEEE, 2020: 1278-1289.

[56] ALMALIOGLU Y, TURAN M, SAPUTRA M R U, et al. SelfVIO: self-supervised deep monocular visual-inertial odometry and depth estimation[J]. Neural Networks, 2022, 150: 119-136.

[57] SUN L, YIN W, XIE E, et al. Improving monocular visual odometry using learned depth[J]. IEEE Transactions on Robotics, 2022, 38(5): 3173-3186.

[58] XING X, CAI Y, LU T, et al. Joint self-supervised monocular depth estimation and SLAM[C]// Proceedings of 2022 26th International Conference on Pattern Recognition (ICPR). Montreal: Springer, 2022: 4030-4036.

[59] ZHU Z, PENG S, LARSSON V, et al. NICE-SLAM: neural implicit scalable encoding for SLAM[C]// Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans: IEEE, 2022.

[60] KWON O, PARK J, OH S. Renderable neural radiance map for visual navigation[C]// Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023: 9099-9108.

[61] YU J, LOW J E, NAGAMI K, et al. NeRFBridge: bringing real-time, online neural radiance field training to

robotics[J]. arXiv preprint arXiv: 2305.09761, 2023.

[62] KUANG H, CHEN X, GUADAGNINO T, et al. IR-MCL: implicit representation-based online global localization[J]. IEEE Robotics and Automation Letters, 2023, 8(3): 1627-1634.

[63] GUÉDON A, LEPETIT V. SuGaR: surface-aligned Gaussian splatting for efficient 3D mesh reconstruction and high quality mesh rendering[C]// Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2024.

[64] KEETHA N V, KARHADE J, JATAVALLABHULA K M, et al. SplaTAM: splat, track & map 3D Gaussians for dense RGB-D SLAM[C]// Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2024.

[65] YAN C, QU D, WANG D, et al. GS-SLAM: dense visual SLAM with 3D Gaussian splatting[C]// Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2024.

[66] MATSUKI H, MURAI R, KELLY P H J, et al. Gaussian splatting SLAM[C]// Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2024.

[67] YUGAY V, LI Y, GEVERS T, et al. Gaussian-SLAM: photo-realistic dense SLAM with Gaussian splatting[J]. arXiv preprint arXiv: 2312.10070, 2023.

[68] ZHU S, QIN R, WANG G, et al. SemGauss-SLAM: dense semantic Gaussian splatting SLAM[J]. arXiv preprint arXiv: 2403.07494, 2024.

[69] JI Y, LIU Y, XIE G, et al. NEDS-SLAM: a novel neural explicit dense semantic SLAM framework using 3D Gaussian splatting[J]. IEEE Robotics and Automation Letters, 2024, 9(10): 8778-8785.

[70] LI M, LIU S, ZHOU H. SGS-SLAM: semantic Gaussian splatting for neural dense SLAM[J]. arXiv preprint arXiv: 2402.03246, 2024.

[71] HU J, CHEN X, FENG B, et al. CG-SLAM: efficient dense RGB-D SLAM in a consistent uncertainty-aware 3D Gaussian field[C]// Proceedings of European Conference on Computer Vision. Milan: Springer, 2024: 93-112.

[72] DENG T, CHEN Y, ZHANG L, et al. Compact 3D Gaussian splatting for dense visual SLAM[J]. arXiv preprint arXiv: 2403.11247, 2024.

[73] XU H, PENG S, WANG F, et al. DepthSplat: connecting Gaussian splatting and depth[J]. arXiv preprint arXiv: 2410.13862, 2024.

[74] LUITEN J, KOPANAS G, LEIBE B, et al. Dynamic 3D Gaussians: tracking by persistent dynamic view synthesis[J]. arXiv preprint arXiv: 2308.09713, 2023.

（编辑：孟彬）