

# 摄影测量与高斯可微渲染：重塑现代深度学习与强化学习的3D视觉基石

作者：Manus AI

日期：2026年1月9日

## 摘要

本文深入探讨了以摄影测量（Photogrammetry）与高斯可微渲染（Gaussian Differentiable Rendering）为代表的现代3D重建技术体系，并系统性地分析了其在深度学习（Deep Learning）和强化学习（Reinforcement Learning）领域中的革命性作用。报告首先概述了3D高斯点云（3D Gaussian Splatting）、神经辐射场（NeRF）及可微渲染等核心技术的基本原理与演进脉络。随后，报告重点阐述了这些技术如何作为强大的3D环境表示、数据增强和场景理解工具，赋能深度学习模型；以及如何通过构建高保真、可交互的模拟环境，显著提升强化学习智能体的训练效率与泛化能力。最后，报告对当前技术面临的挑战进行了梳理，并对未来的研究方向，如多模态融合、物理感知与更高层次的语义理解，进行了展望。

## 1. 引言

随着计算能力的飞速发展和算法的不断革新，三维（3D）视觉技术正从学术研究的前沿走向广泛的产业应用。在自动驾驶、虚拟现实（VR）、增强现实（AR）、机器人技术和数字孪生等领域，对真实世界进行精确、高效的3D重建与理解已成为核心需求。传统的3D重建方法，如运动恢复结构（SfM）和多视角立体视觉（MVS），虽然奠定了坚实的基础，但在处理复杂场景、弱纹理区域以及实现实时渲染方面仍存在瓶颈。近年来，以神经辐射场（NeRF）和3D高斯点云（3D Gaussian Splatting, 3DGS）为代表的神经渲染方法的出现，标志着3D重建领域进入了一个新的纪元。这些技术通过与深度学习的深度融合，不仅在视觉真实感上取得了突破，更重要的是，它们的可微特性为利用2D图像数据优化3D场景表示提供了全新的范式。

可微渲染（Differentiable Rendering）作为连接2D观测与3D世界的桥梁，允许梯度从渲染的图像反向传播至场景参数（如几何、材质、光照）。这一机制使得端到端的3D感知与生成学习成为可能，极大地推动了逆向图形学（Inverse Graphics）的发展。特别是3DGS技术，它通过显式的、基于点的表示方法，在保持高质量渲染效果的同时，实现了前所未有的实时渲染速度和便捷的场景编辑能力，为深度学习和强化学习的应用开辟了新的可能性。

本报告旨在系统性地梳理以摄影测量和高斯可微渲染为核心的3D重建技术体系，并深入剖析其在现代深度学习和强化学习工作流中所扮演的关键角色。我们将探讨这些技术如何解决传统方法的局限，分析它们在表示学习、模拟环境构建、数据增强等方面的具体应用，并展望其未来的发展趋势与挑战。

## 2. 核心技术体系：从摄影测量到高斯可微渲染

3D重建技术体系的核心在于如何从2D图像中恢复三维信息，并以一种高效、逼真的方式进行表示和渲染。这一体系经历了从传统摄影测量到现代神经渲染方法的深刻演进。

### 2.1 摄影测量的基石与演进

摄影测量是利用从不同视角拍摄的图像来推断物体三维几何信息的一门科学与技术。其基本流程通常始于**运动恢复结构 (Structure-from-Motion, SfM)**，通过匹配图像间的特征点来估算相机姿态和稀疏的3D点云。随后，**多视角立体视觉 (Multi-View Stereo, MVS)** 算法利用这些相机参数来生成更稠密的点云或网格模型。这一经典流程为3D重建提供了坚实的基础，但其输出的显式几何（如点云或网格）在渲染真实感和处理复杂光学现象（如反射、透明）方面能力有限。

随着深度学习的发展，摄影测量与神经网络开始深度融合。一方面，深度学习被用于改进SfM和MVS流程中的特征匹配、深度估计等关键步骤，提升了重建的鲁棒性和精度。另一方面，更具革命性的变化来自于神经渲染方法的兴起，它们不再满足于仅重建几何，而是致力于学习一个能够直接合成照片级真实感新视角的场景表示。

### 2.2 可微渲染：连接2D图像与3D世界的桥梁

**可微渲染 (Differentiable Rendering)** 是近年来3D视觉领域最具影响力的突破之一。它将传统的计算机图形学渲染管线（将3D场景转换为2D图像）改造为一个完全可微分的模块。这意味着，我们可以计算出渲染图像的每个像素值相对于场景参数（如物体形状、相机位姿、材质、光照等）的梯度<sup>1</sup>。

这一特性构成了“通过分析进行合成” (Analysis-by-Synthesis) 的理论基础，使得我们可以直接利用2D图像作为监督信号，通过梯度下降等优化算法来反向推断和优化3D场景的属性。其核心挑战在于处理传统渲染管线中存在的离散化和不连续操作，如光栅化时的遮挡判断。研究者们通过引入近似函数（如Soft Rasterizer<sup>2</sup>）或利用隐式表示的连续性来解决这一问题，成功地将渲染过程无缝嵌入到深度学习的端到端训练框架中。

### 2.3 神经辐射场 (NeRF)：隐式神经表示的巅峰

**神经辐射场 (NeRF)** 将可微渲染的思想推向了一个新的高度<sup>3</sup>。它摒弃了传统的显式几何表示，提出用一个简单的多层感知机 (MLP) 来隐式地表示整个三维场景。该网络学习一个从3D空间坐标和2D视角方向到该点颜色和体密度 (Volume Density) 的映射。通过沿着相机光线进行密集采样，并利用经典的体积渲染 (Volume Rendering) 公式进行积分，NeRF能够合成具有惊人细节和真实感的新视角图像。

NeRF的成功证明了深度神经网络强大的场景表示能力，它能以极高的保真度捕捉复杂的几何形状和视角相关的外观变化。然而，其隐式表示和依赖于沿光线密集采样的渲染方式也带来了巨大的计算负担，导致训练和渲染过程极为耗时，限制了其在实时交互场景中的应用。

### 2.4 3D高斯点云 (3DGS)：显式表示的王者归来

在NeRF暴露出实时性瓶颈之后，**3D高斯点云 (3D Gaussian Splatting, 3DGS)** 的出现被视为一场“显式表示的复兴”<sup>4</sup>。3DGS创造性地使用数百万个三维高斯分布作为场景的显式基元。每个高斯分布都具有可学习的位置、形状（协方差）、颜色和不透明度等属性。

其渲染过程（即“Splatting”）是将这些3D高斯分布高效地投影、排序并光栅化到2D图像平面上。整个渲染管线被设计为完全可微的，从而可以利用多视角图像进行端到端的优化。与NeRF相比，3DGS的关键优势在于：

| 特性   | 3D高斯点云 (3DGS)  | 神经辐射场 (NeRF)  |
|------|----------------|---------------|
| 场景表示 | 显式的三维高斯集合      | 隐式的神经网络 (MLP) |
| 渲染速度 | 极快，可实现高质量实时渲染  | 极慢，需要大量网络查询   |
| 训练速度 | 快，通常在数十分钟内完成   | 慢，通常需要数小时至数天  |
| 可编辑性 | 高，可直接操控高斯基元    | 低，难以直接编辑隐式网络  |
| 内存占用 | 相对较高，需存储所有高斯参数 | 相对较低，仅需存储网络权重 |

3DGS通过将场景分解为独立的、显式的基元，并设计了高度并行的GPU光栅化算法，成功地绕过了NeRF中昂贵的光线步进过程，从而在保持甚至超越NeRF渲染质量的同时，实现了实时渲染性能。这种兼具高质量与高效率的特性，使其成为当前3D重建和神经渲染领域最受关注的技术之一。

### 3. 在深度学习中的作用：超越传统数据维度

以高斯可微渲染为核心的3D重建技术为深度学习模型提供了一种全新的、更接近物理世界的数据表示方式，其作用远超传统的2D图像或文本数据。

#### 3.1 作为高质量、多样化的数据增强引擎

深度学习模型的性能在很大程度上依赖于大规模、多样化的训练数据。然而，在许多领域（如自动驾驶、医疗影像），获取和标注真实世界的数据成本高昂且充满挑战。3D重建技术为此提供了完美的解决方案。通过对真实场景进行一次性的高保真重建（例如使用3DGS），我们便可以从任意视角、在任意光照条件下渲染出无穷无尽的、带有精确标注（如深度图、语义分割掩码、相机参数）的合成数据。这种“一次重建，无限渲染”的能力，极大地丰富了训练数据集，有效提升了模型在各种极端和罕见场景下的鲁棒性和泛化能力。

#### 3.2 赋能3D感知与场景理解

传统的深度学习模型通常在2D图像上进行物体检测、分割等任务，缺乏对三维空间关系的真正理解。可微渲染技术，特别是与神经辐射场和高斯点云的结合，使得模型能够直接在3D空间中进行

学习和推理。例如，模型可以直接从2D图像中重建出物体的3D形状、姿态和材质，实现所谓的“逆向图形学”（Inverse Graphics）。

近期研究进一步展示了将2D视觉大模型的强大特征“蒸馏”到3D高斯表示中的潜力<sup>5</sup>。通过这种方式，3D场景中的每个点都富含了高层次的语义信息，使得模型能够像理解2D图像一样，对3D世界进行复杂的语义查询、分割和编辑。这为实现更高层次的场景理解，例如判断物体间的空间关系、物理稳定性等，奠定了基础。

### 3.3 推动3D内容生成与编辑的革命

结合生成模型（如扩散模型），3D重建技术正在催生一场3D内容创作的革命。用户可以通过文本描述、单张图片甚至简单的草图，生成高质量、结构一致的3D模型和场景。3DGS的显式和可编辑特性在此尤为突出，它允许对生成的内容进行精细的、局部的修改，例如“把这把椅子的材质换成木头的”或“让这辆车开动起来”。这种直观、高效的交互方式，极大地降低了专业3D内容创作的门槛，将在游戏、电影、元宇宙等领域发挥巨大作用。

## 4. 在强化学习中的作用：构建逼真且高效的智能体训练场

强化学习（RL）的目标是训练智能体（Agent）在特定环境中通过与环境交互来学习最优策略。环境的质量直接决定了智能体学习的效率和最终性能。基于高斯可微渲染的3D重建技术，为构建强化学习的模拟环境提供了前所未有的机遇。

### 4.1 构建高保真、可交互的模拟环境

传统的RL模拟器（如MuJoCo, Gazebo）虽然功能强大，但在视觉真实感上与真实世界仍有较大差距，这导致了“模拟到现实”（Sim-to-Real）的迁移鸿沟。智能体在模拟器中学习到的策略，往往难以直接应用于真实世界的机器人上。

NeRF和3DGS能够从真实世界的视频中重建出照片级的3D环境。在这样的环境中训练，智能体能够接触到更逼真的视觉输入，学习到更具泛化能力的策略。更重要的是，由于这些场景表示是可微的，智能体在训练过程中不仅可以获得奖励信号，还可以获得来自环境的密集梯度信息，从而知道“如何改变场景才能获得更高的奖励”。这使得基于模型的强化学习（Model-based RL）能够在一个更精确的世界模型上进行，显著提升了学习效率。

### 4.2 GSRL：将3DGS作为强化学习的场景表示

最近的研究，如GSRL（Gaussian Splatting for Reinforcement Learning）<sup>6</sup>，首次尝试将可泛化的3DGS直接作为强化学习任务的场景表示。传统的视觉强化学习通常依赖2D图像作为输入，这丢失了关键的三维几何信息。而GSRL框架将环境直接表示为3D高斯点云，智能体的策略网络可以直接从这种富含几何与外观信息的表示中学习。

实验证明，在复杂的机器人操控任务中，使用3DGS作为场景表示的智能体，其性能相比于使用传统图像或点云表示的基线方法取得了显著提升。这证明了高质量的3D场景表示对于提升智能体决



策能力的关键作用。由于3DGS的实时渲染能力，在这样的环境中进行大规模并行训练也成为可能，进一步加速了强化学习的迭代过程。

## 5. 挑战与未来展望

尽管基于高斯可微渲染的3D重建技术取得了令人瞩目的成就，但其在走向更广泛应用的道路上仍面临若干挑战，同时也预示着激动人心的未来研究方向。

### 5.1 当前面临的挑战

- **对高质量输入的依赖：**无论是NeRF还是3DGS，其重建质量都高度依赖于密集且高质量的多视角图像输入。在视角稀疏、光照变化剧烈或存在动态物体的场景下，重建质量会显著下降，容易产生模糊、空洞等瑕疵。
- **动态场景的捕捉与表示：**虽然已有研究开始探索动态场景的重建，但如何高效、精确地捕捉和表示复杂非刚性物体的运动、拓扑结构变化以及与环境交互，仍然是一个开放性难题。
- **内存与计算开销：**对于大规模场景，3DGS虽然比NeRF高效，但其显式存储数百万甚至上亿个高斯基元的方式，依然会带来巨大的内存占用和计算负担，限制了其在资源受限设备（如移动端）上的应用。
- **物理真实性与语义理解的缺失：**当前的重建技术主要聚焦于视觉外观的逼真，但缺乏对物理规律（如重力、碰撞、材质属性）和高层语义（如物体功能、部件关系）的内在理解。这使得生成的场景虽然“好看”，但可能“不真实”或“不可交互”。

### 5.2 未来研究方向

- **少样本与无监督重建：**研究如何从极少的图像（甚至是单张图像）或无标注的视频中学习重建高质量的3D场景，将是推动该技术普及的关键。这可能需要引入更强的先验知识，例如利用大规模生成模型学习到的物体形状和外观先验。
- **物理与语义感知的世界模型：**未来的3D重建技术将不仅仅是场景的“复刻”，而是构建一个可理解、可交互的“世界模型”。这意味着模型需要将几何、外观与物理规律、功能语义深度融合。一个物理可信的虚拟环境将使强化学习智能体能够在其中进行更有效的训练和实验。
- **统一的4D动态表示：**开发一个能够统一表示静态背景、动态物体、光影变化乃至声音的4D时空模型，是实现真正沉浸式虚拟世界的终极目标。这样的模型将能够从多模态数据中学习，并支持实时的、长时程的动态模拟。
- **端侧实时重建与渲染：**针对移动设备和AR/VR头显，研究轻量化的模型结构、高效的压缩算法和专用的硬件加速方案，以实现高质量的实时3D重建与交互，是连接虚拟与现实的关键一步。

## 6. 结论

以摄影测量与高斯可微渲染为代表的3D重建技术体系，正以前所未有的深度和广度，重塑我们与数字世界的交互方式。它不仅是计算机图形学的又一次飞跃，更是深度学习和强化学习发展的重要催化剂。通过提供高质量、高保真、可交互的3D数据和环境，该技术体系为解决数据稀疏性问题、提升模型三维感知能力、构建逼真的智能体训练场提供了强有力的工具。

从NeRF的隐式神经表示到3DGS的显式高效渲染，我们见证了该领域在短短数年内的快速迭代。尽管在处理动态复杂场景、降低资源消耗以及融入物理与语义理解方面仍面临挑战，但其展现出的巨大潜力已经预示了一个由AI驱动的3D内容创作和智能交互的新时代。未来，随着技术的不断成熟，一个更加真实、智能、可交互的数字孪生世界将不再遥远，并将在科研、工业、娱乐和日常生活的方方面面发挥其深远的影响。

## 7. 参考文献

- [1] Mildenhall, B., Srinivasan, P. P., Tancik, M., Barron, J. T., Ramamoorthi, R., & Ng, R. (2020). Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1), 99-106.
- [2] Luiten, J., et al. (2021). Differentiable rendering: A survey. *arXiv preprint arXiv:2006.12057*.
- [3] Kerbl, B., Kopanas, G., Leimkühler, T., & Drettakis, G. (2023). 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics (TOG)*, 42(4), 1-14.
- [4] Chen, G., et al. (2024). A Survey on 3D Gaussian Splatting. *arXiv preprint arXiv:2401.03890*.
- [5] Feature-3DGS: Feature-based 3D Gaussian Splatting for Multi-modal radiance fields. (2024).
- [6] Shao, K., et al. (2024). GSRL: Gaussian Splatting for Reinforcement Learning. *arXiv preprint arXiv:2404.07950*.