中图法分类号:TP391.4;TP18　　**文献标识码**:A　　**文章编号**:1006-8961(2026)03-0657-29

论文引用格式:Yuan Z L, Li Z H, Chen K H, Mao T L, Jiang H and Wang Z Q. 2026. Multiview stereo 3D reconstruction research: a survey. Journal of Image and Graphics, 31(3):0657-0685(袁祯泷, 李泽昊, 陈科桦, 毛天露, 蒋浩, 王兆其. 2026. 多视角立体匹配三维重建研究综述. 中国图象图形学报, 31(3):0657-0685)[DOI:10.11834/jig.250348]

多视角立体匹配三维重建研究综述

袁祯泷,李泽昊,陈科桦,毛天露*,蒋浩,王兆其

1. 中国科学院计算技术研究所,北京　100190; 2. 中国科学院大学,北京　100049

摘　要:多视角三维重建是计算机视觉与图形学中的关键问题之一,广泛应用于虚拟现实、增强现实、自动驾驶和文物修复等领域。其核心目标是从多个视角的图像或视频中恢复出三维场景的几何结构信息,实现物体和场景的高精度三维建模。本文创新性地从图像投影—几何推理与全局—局部两个维度将现有多视角三维重建方法分成4个类别,然后简要介绍了各类方法的典型模型、最新研究进展和它们的适用性及局限性。此外,本文还探讨了多视角三维重建中常用的数据集和评价指标,并从场景、方法优缺点等多个角度对各类方法进行了详细评估。最后,本文深入分析了在多模态大模型、元宇宙等背景下三维重建面临的机遇和挑战,提出了未来的研究和发展方向。

关键词:多视角立体匹配(MVS);三维重建;三维视觉;神经辐射场(NeRF);三维高斯泼溅(3DGS)

Multiview stereo 3D reconstruction research: a survey

Yuan Zhenlong, Li Zehao, Chen Kehua, Mao Tianlu*, Jiang Hao, Wang Zhaoqi

1. Institute of Computing Technology, Chinese Academy of Sciences, Beijing　100190, China;

2. University of Chinese Academy of Sciences, Beijing　100049, China

Abstract: Multiview stereo (MVS) 3D reconstruction aims to recover the three-dimensional geometric structure of a scene from multiple two-dimensional images or videos, enabling precise modeling of objects and environments. This task is fundamental to applications such as virtual reality, autonomous driving, smart city construction, and cultural heritage preservation. Traditional MVS methods rely on geometric constraints such as homography mappings and epipolar geometry to align multiview images and infer depth. However, these approaches are limited by their dependence on hand-crafted features and struggle with low-texture regions or dynamic scenes. The advent of deep learning has driven significant advancements, enabling end-to-end networks to automatically learn image-depth relationships and achieve higher accuracy and robustness. Despite these breakthroughs, challenges such as computational efficiency, generalization to complex scenes, and real-time performance remain unresolved. To systematically analyze the current state of MVS 3D reconstruction, this work categorizes existing methods into image projection-driven and geometry reasoning-driven paradigms, further subdividing them on the basis of local modeling and global modeling strategies. The paper also explores widely used datasets, evaluation metrics, and technical innovations within each category while identifying key challenges and proposing future research directions. First, this study investigates commonly used datasets and evaluation indicators in MVS 3D reconstruction. Datasets serve as the foundation for training and benchmarking algorithms and can be classified into three categories: synthetic data-

收稿日期:2025-07-25;**修回日期**:2025-09-02;**预印本日期**:2025-09-09

*通信作者:毛天露　ltm@ict.ac.cn

基金项目:中国科学院战略性先导科技专项资助(XDB 1290302);国家自然科学基金项目(62172392)

Supported by: Strategic Priority Research Program of the Chinese Academy of Sciences (XDB 1290302); National Natural Science Foundation of China (62172392)

sets, real-scene datasets, and hybrid datasets. Synthetic datasets, such as DTU and Tanks and Temples, provide high-resolution, precisely annotated data for algorithm training and validation. Real-scene datasets, including ETH3D and LLFF, capture realistic environments with varying textures and lighting conditions, enabling evaluation under practical constraints. Hybrid datasets, such as nerf-synthesis and Synthetic Soccer NeRF, combine synthetic and real-world elements to address domain adaptation challenges. Evaluation metrics are critical for quantifying algorithm performance and are primarily divided into geometric accuracy metrics and rendering quality metrics. Geometric accuracy metrics, such as chamfer distance (CD) and F-score, measure the spatial fidelity between reconstructed and ground-truth models, with lower CD values and higher F-score values indicating better performance. Rendering quality metrics, including peak signal-to-noise ratio (PSNR), structural similarity index measure (SSIM), and learned perceptual image patch similarity (LPIPS), are used to assess the visual consistency of synthesized views, with higher PSNR and SSIM scores and lower LPIPS scores reflecting superior perceptual quality. These metrics collectively provide a comprehensive framework for evaluating MVS methods across diverse scenarios. Second, the paper classifies MVS 3D reconstruction methods on the basis of their technical frameworks and innovations. Methods can be broadly categorized into image projection-driven approaches and geometry reasoning-driven approaches. Image projection-driven methods, such as MVSNet and its variants (e. g., CasMVSNet, R-MVSNet), focus on explicit cost volume construction and global optimization to infer depth maps. These methods leverage feature pyramids and attention mechanisms to enhance robustness in low-texture regions and complex geometries. Geometry reasoning-driven methods, including NeRF, Plenoxels, and 3D Gaussian splatting, adopt implicit or explicit representations to model scenes. NeRF and its derivatives (e. g., mip-NeRF, Ref-NeRF) use neural radiance fields to achieve photorealistic view synthesis, while 3DGS introduces Gaussian point clouds for efficient rendering and dynamic scene modeling. Within these categories, methods can further be divided into local modeling and global modeling subcategories. Local modeling approaches, such as PatchMatch and Gipuma, emphasize pixel-level matching and plane-induced homography, while global modeling techniques prioritize holistic scene consistency through volumetric or hierarchical representations. Innovations in recent works include sparse-to-dense depth estimation, adaptive sampling strategies, and hybrid architectures that combine explicit structures (e. g., octrees) with implicit neural networks to balance efficiency and detail preservation. Third, the paper identifies unresolved challenges and outlines future research directions. Data-related challenges include the difficulty of collecting high-quality multi-view datasets for dynamic scenes and the need for scalable annotation tools to reduce human labor. From a methodological perspective, existing techniques face limitations in computational efficiency (e. g., cubic memory complexity in NeRF) and generalization to unseen environments (e. g., natural vegetation or translucent objects). Training paradigms also need to be improved; supervised methods depend on expensive 3D annotations, while unsupervised and self-supervised approaches often sacrifice reconstruction quality for reduced data dependency. To address these issues, future research should focus on the following: 1) lightweight and real-time optimization, which involves developing hardware-aware architectures (e. g., edge computing frameworks) and dynamic resolution adjustment to reduce GPU memory consumption and inference latency; 2) dynamic scene modeling, integrating temporal consistency constraints and physics-based priors to handle motion, occlusions, and deformable objects in real-world applications; 3) cross-modal fusion, leveraging multimodal inputs (e. g., LiDAR, inertial sensors) to enhance robustness in low-texture or adverse lighting conditions, and 4) semantic-aware reconstruction, incorporating semantic segmentation and instance-level reasoning to enable editable 3D models for virtual production and interactive design. The convergence of MVS with emerging technologies such as multimodal large models and metaverse platforms will further expand its applications. For example, real-time 3D reconstruction in autonomous vehicles requires not only geometric precision but also semantic understanding for obstacle detection. Similarly, metaverse environments demand high-fidelity, scalable reconstructions of large-scale urban scenes. To bridge the gap between academic advancements and industrial deployment, future efforts must prioritize domain adaptation, computational efficiency, and cross-modal learning. By addressing these challenges, MVS 3D reconstruction will play a pivotal role in shaping next-generation immersive technologies, robotics, and digital twins. Additionally, the integration of multitask learning and meta-learning could further enhance model generalization by jointly optimizing geometric and semantic tasks. For instance, meta-learning frameworks could adapt to novel object categories with minimal samples, while multitask architectures might simultaneously

reconstruct geometry and infer surface properties (e. g. , material reflectance). Furthermore, the development of 3D foundation models that are pretrained on vast heterogeneous datasets could unlock universal reconstruction capabilities across diverse domains. Ethical considerations, such as privacy in public space reconstruction and bias in dataset curation, must be addressed to ensure responsible deployment. Addressing these multidimensional challenges will enable MVS to evolve from a research-oriented technique to an indispensable tool for real-world 3D perception.

Key words: multiview stereo (MVS); 3D reconstruction; 3D vision; neural radiance field (NeRF); 3D Gaussian splatting (3DGS)

0 引 言

多视角立体匹配(multiview stereo,MVS)三维重建作为计算机视觉与图形学领域的核心问题,其目标是从多视角图像或视频中精确恢复场景几何结构,为虚拟现实、自动驾驶等关键领域提供高精度三维建模支撑。传统的三维重建方法(刘草 等,2025)依赖单应性映射与极线约束实现逐像素匹配,但受限于人工设计的特征的局限性,难以应对低纹理区域的建模需求。近年来,深度学习与神经渲染技术的突破使技术路线呈现多样化趋势:一方面,基于特征金字塔网络的端到端方法,如 MVSNet(Yao 等,2018),通过代价体(cost lolume)构建显著提升了遮挡区域的重建精度;另一方面,神经辐射场(neural radiance field,NeRF)与三维高斯点云泼溅(3D Gaussian splatting,3DGS)等新型方法通过隐式函数建模实现了更灵活的视角合成。上述技术范式的演进不仅体现在精度提升上,更反映出从显式特征匹配向隐式几何推理的范式迁移(于柳和吴晓群,2024)。

当前方法分类体系尚未形成统一框架(杨航等,2023)。早期研究多采用局部匹配与全局优化的二元划分,但随着 NeRF(Mildenhall 等,2022)、3DGS(Kerbl 等,2023)等新兴技术的涌现,传统分类已难以覆盖方法演进的多样性。具体而言,图像投影驱动方法(如 MVSNet)通过多视角立体匹配与特征金字塔网络实现逐像素视差优化,其代价体构建虽提升了低纹理区域的鲁棒性,但显式体积计算对 GPU(graphics processing unit)内存的高依赖性限制了其在边缘设备上的部署。相比之下,几何推理驱动方法(如 NeRF、3DGS)通过隐式密度场或显式高斯泼溅(也称高斯点云)表示,实现了更高效的视角合成:NeRF 利用体积渲染技术生成连续视角插值,3DGS 则通过可微分渲染的高斯点云在保持几何精度的同时优化动态场景效率。值得注意的是,现有综述文献多聚焦图像投影驱动方法的演进脉络,对几何推理方法的系统性梳理仍显不足,目前主要存在两个关键问题:1)多数工作仅关注单一维度的演进(如单纯比较 NeRF 与 3DGS 的差异),难以揭示方法论的内在演化逻辑;2)针对动态场景建模、跨模态融合等新兴需求,现有分类体系未能体现不同技术路线在场景适应性与计算效率间的权衡机制。

多视角三维重建技术的发展始终围绕“如何高效且精确地从二维图像中恢复三维结构”这一核心问题展开。早期基于传统几何的方法通过单应性映射与全局能量优化实现重建,但受限于手工特征提取的局限性,难以应对低纹理区域的建模需求。深度学习的引入使端到端神经网络架构能够自动学习图像特征与深度关系,显著提升了重建精度,但代价体积构建与正则化过程对计算资源的高依赖性成为其落地瓶颈。研究者并未局限于单一维度的性能优化:从显式代价体(MVSNet)到隐式神经场(NeRF)的建模范式转变,体现了从图像投影驱动向几何推理驱动的技术迁移;局部细节增强(edge-Gaussians)与全局一致性优化(Plenoxels(Fridovich-Keil 等,2022))的并行发展,则揭示了建模尺度从全局向局部的演进趋势。这种技术路径的交织演进,本质上源于对场景复杂度、计算效率与重建精度三者权衡机制的持续探索。多视角三维重建技术的发展始终沿着“图像投影—几何推理”与“全局建模—局部建模”两条技术路径交织演进。

因此,本文创新性地提出基于双维度的 4 类划分体系:1)图像投影驱动的局部建模方法;2)图像投影驱动的全局建模方法;3)几何推理驱动的局部建模方法;4)几何推理驱动的全局建模方法。该体系不仅系统梳理了方法演进的时空坐标,更通过量化评估框架揭示了各类方法在动态场景建模、多模态融合等前沿领域的适用边界,为研究者提供了更清

晰的前景视角。

本文的创新性贡献体现在3个方面：1）建立基于双维度的4类方法划分体系，系统梳理多视角三维重建的技术演进；2）从场景复杂度、光照变化、遮挡处理等维度构建量化评估框架，揭示不同方法的适用边界；3）结合典型数据集，如DTU（Denmark Technical University）（Jensen等，2014）、Tanks and Temples（Knapitsch等，2017）的实验结果，对比分析各方法在精度、效率与泛化能力上的表现差异。通过上述工作，本文旨在为研究者提供清晰的技术发展脉络与未来研究方向，推动多视角三维重建技术在实际应用场景中的落地。

1 数据集与评价指标

1.1 数据集

在传统PatchMatch（Barnes等，2009）和基于深度学习的MVSNet领域，其性能高度依赖于高质量数据集的支持。本文选取了4类具有代表性的公开数据集进行分析，涵盖不同场景复杂度、分辨率及评估标准，旨在全面呈现当前MVS研究的基准框架与挑战。

Strecha数据集（Strecha等，2008）以高分辨率（3 072×2 048像素）著称，包含6个室外建筑场景（如fountain-P11、Herz-Jesu-P8等），并提供激光扫描点云与部分场景的真实深度图。其优势在于清晰纹理与理想光照条件，为算法效率优化提供了基础，但受限于场景同质性，对复杂环境的适应性验证不足。相比之下，DTU数据集通过124个实验室场景（1 600×1 200像素）覆盖了从弱纹理表面到复杂几何结构的多样化测试，其固定相机轨迹与精确参数标注更利于鲁棒性评估，但小尺度物体与有限视点变化限制了其泛化能力。

针对高精度重建需求，ETH3D（Eidgenössische technische Hochschule Zürich 3D）（Schöps等，2017）高分辨率分支（6 048×4 032像素）提供了26个场景的完整训练/测试集，结合真实深度图与点云模型，成为高分辨率算法验证的黄金标准。而Tanks and Temples数据集则聚焦重建流程的整体性能，包含室内外混合场景（1 920×1 080像素），通过Intermediate与Advanced两阶段划分，分别针对中等复杂度与大规模室内重建任务进行挑战设计。值得注意的是，该数据集未提供相机参数，更贴近实际应用中的未知条件。

总体而言，上述数据集在分辨率、场景多样性及评估维度上形成互补：Strecha与ETH3D侧重高精度与高分辨率算法开发，DTU强化鲁棒性测试，Tanks and Temples则模拟真实重建流程。这种差异化设计为研究者提供了从基础理论到工程落地的完整实验框架，推动了MVS技术的持续演进。

在神经辐射场（NeRF）和神经三维高斯泼溅（3DGS）领域，数据集的设计同样体现了对算法性能与应用场景的针对性适配。NeRF的代表性数据集如Diffuse Synthetic 360（Sitzmann等，2019）和Realistic Synthetic 360（Mildenhall等，2022），分别包含4类简单几何物体（512×512像素）和8类复杂材质物体（800×800像素）的合成场景，为算法在光照一致性与非朗伯反射特性下的鲁棒性测试提供了基础。而真实场景数据则以8组手机拍摄的复杂场景（1 008×756像素）为主，强调实际应用中的视角连续性与动态遮挡挑战。

针对动态场景建模，Synthetic Soccer NeRF数据集（Lewin等，2023）通过Blender生成3类运动场景，每类场景配备20～30台同步校准相机以25帧/s采集4 s长视频，提供图像与相机位姿的完整标注，为动态NeRF训练提供了标准化基准。此外，nerf-synthesis数据集（Mildenhall等，2022）包含8个典型物体（如ship、mic等），每个场景划分100幅训练集、100幅验证集及200幅测试集，支持从静态物体到复杂结构的多尺度重建任务。

BlenderNeRF工具进一步降低了NeRF与Gaussian splatting数据集的生成门槛，允许用户通过单键操作在Blender中获取渲染图像与相机参数，兼顾灵活性与高效性。对于真实场景处理，LLFF（local light field fusion）数据集（Mildenhall等，2019）通过手持相机采集的Fern植物等场景（1 008×756像素），结合COLMAP（co-localization and mapping）预处理与NeRF渲染流程，成为自我监督训练的典型基准。

值得注意的是，3DGS相关研究多沿用NeRF框架下的合成或真实数据，其专用数据集仍存在信息缺失。尽管NeRF数据集已覆盖从低分辨率静态物体到高动态范围大型建筑的全链条需求，但针对3DGS的评估标准与任务目标仍有待明确。总体而言，NeRF与3DGS数据集在合成场景、动态建模及真

实数据适配方面形成互补，为算法从理论验证到实际部署提供了多层次支撑。

1.2 评价指标

在进行点云模型评估的过程中，通常会采用准确、完整性以及它们的综合指标这几种方式来进行量化分析。准确性主要是对重建出来的点云与真实值点云之间的近似程度进行评估，而完整性则是对真实值点云被重建点云所覆盖的范围进行度量。设定真实值点云为$\boldsymbol{G}$，重建点云为$\boldsymbol{R}$，经过配准处理后，可以计算出重建点云中任一点$r \in \boldsymbol{R}$到真实值点云$\boldsymbol{G}$的最短距离，具体为

$$e_{r\to G} = \min_{g\in 2} \left| r - g \right| \tag{1}$$

同时，真实值点云中任一点$g \in \boldsymbol{G}$到重建点云$\boldsymbol{R}$的最短距离为

$$e_{g\to R} = \min_{r\in R} \left| g - r \right| \tag{2}$$

当前，评估点云质量常用的度量单位主要有两种：距离度量和百分比度量。接下来，分别对这两种度量单位进行详细阐述。

在距离度量中，准确性的定义为重建点云中所有点到真实值点云的平均距离。具体为

$$Acc = \frac{1}{|\boldsymbol{R}|} \sum_{r\in R} \left| e_{r\to G} \right| \tag{3}$$

而完整性的定义则为真实值点云中所有点到重建点云的平均距离。即

$$Comp = \frac{1}{|\boldsymbol{G}|} \sum_{g\in G} \left| e_{g\to R} \right| \tag{4}$$

基于这两者，可以定义一个综合得分，用于综合反映准确性和完整性。因此当距离度量得分越低时，重建的点云质量越高。而在百分比度量中，准确性定义为在给定距离阈值τ下，重建点云中满足条件的点的数量占总点数的百分比。具体为

$$P(\tau) = \frac{100}{|\boldsymbol{R}|} \sum_{r\in R} \left[e_{r\to G} < \tau \right] \tag{5}$$

而完整性则定义为在给定距离阈值τ下，真实值点云中满足条件的点的数量占总点数的百分比。即

$$R(\tau) = \frac{100}{|\boldsymbol{G}|} \sum_{g\in G} \left[e_{g\to R} < \tau \right] \tag{6}$$

基于这两者，可以定义F_1得分，即这两者的调和平均数。具体为

$$F_1(\tau) = \frac{2P(\tau)R(\tau)}{P(\tau) + R(\tau)} \tag{7}$$

从这个定义可以看出，百分比度量得分越高说明重建的点云质量越好。将准确性和完整性结合起来考虑，可以更全面地反映模型的性能。单纯关注准确性可能会导致多视图立体视觉算法生成高确定性但稀疏的点云，如仅重建纹理丰富区域的准确三维模型。而如果只关注完整性，则可能导致算法倾向于生成覆盖整个三维空间的大量噪声点云，而忽视了点云质量的问题。因此，综合运用上述评价指标才能更加全面、准确地评估点云模型的性能。

此外，对于渲染质量，NeRF 的评估依赖视觉感知指标（如，峰值信噪比（peak signal-to-noise ratio，PSNR）、结构相似性指数（structural similarity index，SSIM）、学习感知图像块相似度（learned perceptual image patch similarity，LPIPS））和几何一致性指标（如倒角距离（chamfer distance，CD）、F-score）。视觉感知指标用于量化 NeRF 渲染结果与真实图像之间的差异，其定义及作用如下：

峰值信噪比（PSNR）基于均方误差（mean squared error，MSE）进行计算，衡量像素级差异。具体为

$$PSNR = 10 \times \lg\left(\frac{MAX^2}{MSE}\right) \tag{8}$$

式中，MAX为像素最大值（如 8 位图像为 255），MSE是均方误差，表示原图像与压缩图像之间的平均像素差异，其计算式为

$$MSE = \sum \left| C^{(r)} - x_c \right|^2 \tag{9}$$

式中，$C^{(r)}$为沿光线r生成的像素颜色（NeRF 模型预测的颜色），x_c为真实颜色（真实点云中的颜色值），$\left| C^{(r)} - x_c \right|^2$是计算颜色之间的平方差，这代表每个像素的误差，$MSE$是所有像素误差的总和，即均方误差，表示了渲染图象和真实点云颜色之间的总体差异。此误差反映了渲染结果与真实场景在颜色分布上的偏离程度。

PSNR 值越高（单位为 dB），表示图像质量越接近真实值。例如，在 DTU 数据集中，优化后的 NeRF 模型 PSNR 可达 34.18 dB，显著高于传统方法（如 28.11 dB）。

结构相似性指数（SSIM）通过对比亮度、对比度和结构信息，评估两幅图像的相似性。其值范围为[0，1]，1 表示完全一致，0 表示完全不一致。相比于 PSNR，SSIM 更符合人眼对结构失真的敏感性，能更好地反映纹理模糊或边缘错位的影响。

学习感知图像块相似度(LPIPS)基于预训练神经网络(如VGG(Visual Geometry Group)),提取图像特征并计算特征空间的差异。其核心思想是:人类感知的图像差异与特征空间的距离更相关。LPIPS值越小,表示感知质量越优。该指标在稀疏视角场景中表现稳定,常用于评估NeRF的细节保留能力。

2 图像投影的多视角立体重建方法

多视角立体重建通过单应性映射将多视角图像进行深度投影,实现从二维图像到三维结构的转换。该方法利用单应性矩阵建立不同视角间的数学关系,使图像平面的二维点可对应到三维空间坐标。根据算法设计范式,现有方法可分为局部投影与全局投影两类:1)局部方法侧重单视角局部特征匹配,具有较高精度但对图像质量敏感;2)全局方法通过多视角信息融合提升鲁棒性,更适合复杂场景重建。下文将分别展开论述。

2.1 基于局部的图像投影方法

局部的图像投影方法聚焦于通过局部区域间的像素匹配实现三维结构恢复,其核心在于利用图像块间的相似性进行视差估计。该方法通过对多视角图像进行逐像素的局部匹配,基于单应性映射建立像素间的对应关系,并通过分析局部投影变化推导出物体的三维几何特性。其代表性工作为PatchMatch算法,其通过随机采样与传播机制显著提升了匹配效率。如图1所示,其核心流程主要包含以下3个关键技术步骤:

在随机初始化阶段,算法首先为每个像素点随机分配一个偏移量,来表示其在参考图像与源图像间的潜在匹配位置。这种随机性打破了局部最优陷阱,确保初始匹配覆盖全局可能性。

随后在传播阶段,基于图像的空间连续性,将已匹配的"良好"偏移量向邻域像素传播。例如,若点A的匹配偏移为$f(A)$,则其邻域点B的候选偏移可设为$f(A)+f(A-B)$。此过程通过局部一致性快速扩展匹配质量,形成连贯的匹配区域。最后,算法在传播基础上引入多尺度随机扰动,逐步修正匹配结果。每轮迭代包含两步:1)随机搜索:以当前偏移为中心,随机生成扰动并评估匹配效果;2)平面细化:对匹配区域的视差平面参数进行局部优化,提升精度。

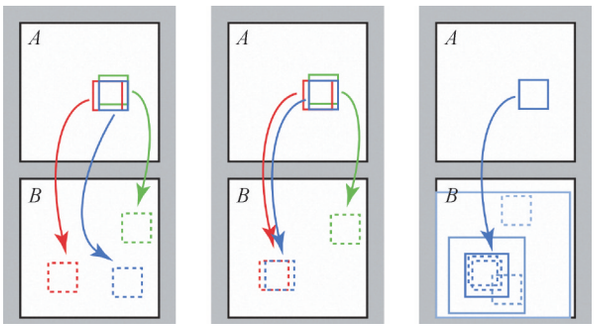

图1 PatchMatch算法主要流程

Fig. 1 The main process of PatchMatch algorithm

该方法通过图1中分阶段的流程设计,实现了从随机采样到全局优化的渐进式匹配,从而显著提升了计算效率与鲁棒性。继PatchMatch算法之后,PMS(PatchMatchStereo)(Bleyer等,2011)首次将PatchMatch算法的思路迁移到了立体视觉中,核心思想是通过动态调整支持窗口的倾斜方向,增强对复杂几何结构的匹配能力。之后Gipuma(Galliani等,2016)首次将PMS的核心思想通过红黑棋盘格的方式部署在了GPU中,大大提升了其运行效率,为局部的图像投影后续的方向研究提供了基石。Gipuma方法具体来说,对每个像素点$\boldsymbol{p}=(x,y)$,随机生成初始视差平面参数$\boldsymbol{P}_0=\{d_0,n_0\}$,其中,$d_0$为初始视差值,$\boldsymbol{n}_0$为平面法向量。平面方程可表示为

$$\boldsymbol{P}: a(x-x_0)+b(y-y_0)+c(d-d_0)=0 \tag{10}$$

式中,(a,b,c)由法向量$\boldsymbol{n}_0$定义,(x_0,y_0,d_0)为平面基准点。这一参数化方式将传统视差空间中的平面扩展到三维场景空间,通过欧几里得几何约束直接建立多视角间的几何关联。

在场景空间中,深度值$\boldsymbol{Z}$的计算依赖于相机内参矩阵$\boldsymbol{K}$与平面参数的联合关系。具体而言,假设参考相机位于坐标原点,目标点$\boldsymbol{X}=[\boldsymbol{X},\boldsymbol{Y},\boldsymbol{Z}]^{\mathrm{T}}$的深度$\boldsymbol{Z}$可推导为

$$\boldsymbol{Z}=\frac{-d\boldsymbol{c}}{\left[x-u,\alpha(y-v),c\right]\cdot\boldsymbol{n}} \tag{11}$$

式中,d为平面到相机光心的距离参数,c代表深度,$\boldsymbol{n}=\left[\boldsymbol{n}_x,\boldsymbol{n}_y,\boldsymbol{n}_z\right]^{\mathrm{T}}$为法向量,径向畸变系数$\alpha$及主点坐标$(u,v)$。该式将二维图像坐标$(x,y)$映射到三维深度$\boldsymbol{Z}$,同时隐式约束了平面法向量与相机参数的关系。通过这一映射,算法能够在不显式进行极线校正的情况下,直接利用多视角信息优化深度估计。

为了进一步实现多视角间的图像点映射，Gipuma 方法引入了平面诱导的单应性变换。具体为

$$\boldsymbol{H}_{\pi} = \boldsymbol{K}'\left(\boldsymbol{R} - \frac{1}{d}\boldsymbol{t}\boldsymbol{n}^{\mathrm{T}}\right)\boldsymbol{K}^{-1}, \boldsymbol{x}' = \boldsymbol{H}_{\pi x} \tag{12}$$

式中，$\boldsymbol{H}_{\pi}$是由平面参数 n 和深度 d 生成的单应性矩阵，R 和 t 分别表示相机外参的旋转和平移，$\boldsymbol{x}$ 为参考相机中的图像点坐标，$\boldsymbol{x}'$ 为目标相机中对应的图像点坐标。该式通过几何约束建立了多视角间的直接映射关系，避免了传统方法中复杂的极线搜索，显著提升了多视图匹配的效率。随后，Gpuma 采用细化策略生来进一步优化法向量。具体为

$$\boldsymbol{n} = \left[1 - 2S, 2q\sqrt{1 - S_1}, 2q\sqrt{1 - S_2}\right]^{\mathrm{T}} \tag{13}$$

式中，q_1, q_2 从区间（−1，1）的均匀分布中采样，直到满足 $S = q_1^2 + q_2^2 < 1$。该方法通过球面参数化技术确保法向量在可见半球上均匀分布，有效避免了初始化偏差。

针对局部图像投影方法的如何有效重建弱纹理区域，现有研究主要围绕 3 类核心策略展开：多尺度粗到细的分层匹配方法、基于平面化约束的后处理方法，以及通过补丁变形补偿形变的策略。以下将系统梳理各类方法的技术路径，并分析其在解决局部投影共性问题中的有效性与局限性。

2.1.1　基于多尺度粗到细的方法

粗到细策略通过多尺度金字塔架构逐步细化深度估计，旨在通过低分辨率到高分辨率的迭代优化提升效率。然而，受限于固定网格划分和几何简化，该类方法在处理大面积纹理缺失区域时存在深度不连续和细节丢失的问题。ACMM（Xu 和 Tao，2019）提出了自适应棋盘方案（adaptive checker board scheme），在低分辨率图像中生成稀疏深度图，再通过插值和细化逐步提升分辨率。其优势在于计算效率较高，但因固定网格划分易导致纹理缺失区域的深度不连续。ACMMP（Xu 等，2023）在 ACMM（Xu 和 Tao，2019）的基础上引入三角化平面化策略，通过连接可靠像素生成连通区域，辅助粗到细阶段的深度传播。然而，平面化过程可能过度简化几何细节，影响复杂场景的重建精度。MG-MVS（Wang 等，2020）采用金字塔架构提取多维特征，结合几何一致性约束优化深度估计。尽管提升了多尺度特征融合能力，但对纹理缺失区域的深度恢复仍显不足。HPM-MVS（Ren 等，2023）利用 KD（K-dimensional tree）树加速平面化过程，通过分层聚类可靠像素生成局部平面假设。其优势在于高效处理大规模场景，但平面化策略对非平面区域（如曲面物体）适应性较差。

2.1.2　基于平面化后处理的方法

平面化策略通过连接可靠像素生成平面区域，假设纹理缺失区域可由平面近似。该类方法依赖平面几何约束，但易导致曲面区域出现重建偏差。PCF-MVS（Kuhn 等，2019）提出平面补全与过滤，通过超像素（super pixel）分割生成平面假设，并利用深度一致性剔除异常值。其核心创新在于结合平面几何约束，但超像素边界可能与真实深度边缘不匹配，导致误差传播。TAPA-MVS（Romanoni 和 Matteucci，2019）针对无纹理区域引入纹理感知补丁匹配，通过局部平面假设减少匹配歧义。然而，平面化策略在复杂曲面场景中易产生几何偏差。TSAR-MVS（Yuan 等，2024b）结合实例分割与关联优化，动态调整平面化区域。其优势在于利用语义信息增强平面假设的鲁棒性，但依赖高质量分割结果，对遮挡敏感。CLD-MVS（Li 等，2020）提出置信度驱动的估计器，通过置信度图筛选可靠像素并生成平面区域。尽管提升了深度估计的稳定性，但平面化过程对非刚性形变区域适应性有限。

2.1.3　基于补丁变形的方法

补丁变形策略通过动态调整固定大小补丁的采样范围，扩大感受野以应对纹理缺失区域。该类方法需解决变形稳定性（如边缘跳跃）和深度边缘对齐问题。PHI-MVS（Sun 等，2021）引入空洞卷积动态调整补丁尺寸，通过扩大采样间隔提升特征提取能力。然而，固定角度的补丁变形可能导致深度边缘对齐不足。API-MVS（Sun 等，2022）基于熵值计算动态调整补丁大小，优先保留高信息量区域。其优势在于减少冗余采样，但熵值计算对噪声敏感，影响可靠性。SD-MVS（Yuan 等，2024a）利用实例分割提取深度边缘，通过垂直/水平方向距离比例缩放补丁尺寸。尽管提升了边缘对齐能力，但缺乏多路径扩散机制，导致复杂场景中补丁变形不稳定。APD-MVS（Wang 等，2023）通过分解不可靠像素的补丁为多个子补丁，结合高可靠性像素计算匹配代价。其核心创新在于局部化优化，但忽略深度连续性前提，易导致跨深度边缘的错误匹配。

2.1.4　基于局部的图像投影方法小结

表1总结了3类基于局部的图像投影方法的设计思路与性能特征。该类方法以统一的代价构建和全视角正则化为核心，在低纹理、遮挡和复杂几何条件下显著提升了深度估计的精度与稳定性。多尺度粗到细方法通过金字塔结构降低计算资源消耗，适合大规模场景处理；平面化后处理方法基于几何平面假设对可靠区域进行连接与补全，适用于规则表面场景；补丁变形类方法通过调整补丁大小扩大感受野，提高了弱纹理区域的匹配能力和鲁棒性。

表2和图2分别展示了上述方法在ETH3D与Tanks and Temples等数据集上的定量和定性对比结果，ACMMP、MG-MVS和HPM-MVS等方法在F_1-score指标上较早期ACMM取得显著优化，而TAPA-MVS与CLD-MVS通过引入局部几何建模和置信度筛选策略，进一步提升了无纹理区域的重建稳定性。这些实验结果验证了不同技术路径的有效性，也反映出方法设计与任务场景之间的耦合关系。

表1　基于局部的图像投影方法总结

Table 1　Summary of local based image projection methods

方法类别	设计思路	优点	缺点
多尺度粗到细	多尺度金字塔架构逐步细化深度，低到高分辨率迭代优化	减少计算资源需求，适合大规模场景处理	复杂纹理缺失区域的深度不连续及细节丢失
平面化后处理	连接可靠像素生成平面区域，假设纹理缺失区域可由平面近似	利用平面几何约束增强稳定性，适合规则表面重建	对曲面和非刚性形变区域适应性差
补丁变形	通过动态调整固定大小补丁的采样范围扩大感受野	利于弱纹理区域特征提取，增强深度估计鲁棒性	复杂场景中易出现跨深度边缘的错误匹配

表2　基于局部的图像投影方法定量对比

Table 2　Quantitative comparison of local based image projection methods

方法	方法类别	数据集	评价指标	对比方法	对比结果/%	实验结果/%
ACMMP	多尺度粗到细	ETH3D/ Test	F_1-score ↑	ACMM	80.78	85.89
HPM-MVS	多尺度粗到细	TNT: Inter/Advanced	F_1-score ↑	ACMM	57.27/34.02	61.39/40.80
TAPA-MVS	平面化后处理	ETH3D/ Test	F_1-score ↑	PCF-MVS	80.38	79.15
TSAR-MVS	平面化后处理	TNT: Inter/Advanced	F_1-score ↑	PCF-MVS	53.39/34.59	62.10/38.63
CLD-MVS	平面化后处理	ETH3D/ Test	F_1-score ↑	PCF-MVS	80.38	82.31
PHI-MVS	补丁变形	ETH3D/ Test	F_1-score ↑	APD-MVS	87.44	86.43
API-MVS	补丁变形	ETH3D/ Test	F_1-score ↑	APD-MVS	87.44	81.99
SD-MVS	补丁变形	TNT: Inter/Advanced	F_1-score ↑	APD-MVS	63.64/39.91	63.31/40.18

注：“↑”表示值越大越好，“↓”表示值越小越好。

总结来看，基于补丁变形的全局图像投影方法通过动态调整补丁尺寸，有效解决了局部方法在纹理缺失区域的深度不连续问题。这类方法通过扩大感受野增强了弱纹理区域的匹配能力，但其在复杂几何场景中的跨边缘匹配稳定性仍需改进。未来的研究可重点关注自适应补丁变形策略与深度连续性约束的联合优化，同时探索多尺度变形机制与几何先验融合的协同效应。此外，如何平衡计算效率与重建精度的权衡关系，仍是提升此类方法实用性的关键方向。

2.2　基于全局的图像投影方法

全局图像投影方法作为多视图立体重建领域的核心演进方向，其发展脉络可划分为3个关键阶段：早期以手工特征为基础的全局优化框架、中期结合深度学习的端到端建模体系，以及当前融合神经渲染与几何推理的混合范式。相较于局部方法依赖的显式像素匹配，全局方法通过构建跨视角的代价表达与空间一致性建模，利用神经网络的特征表达能力与端到端优化策略，实现了从图像输入到深度图输出的整体推理过程。这种范式革新不仅突破了传

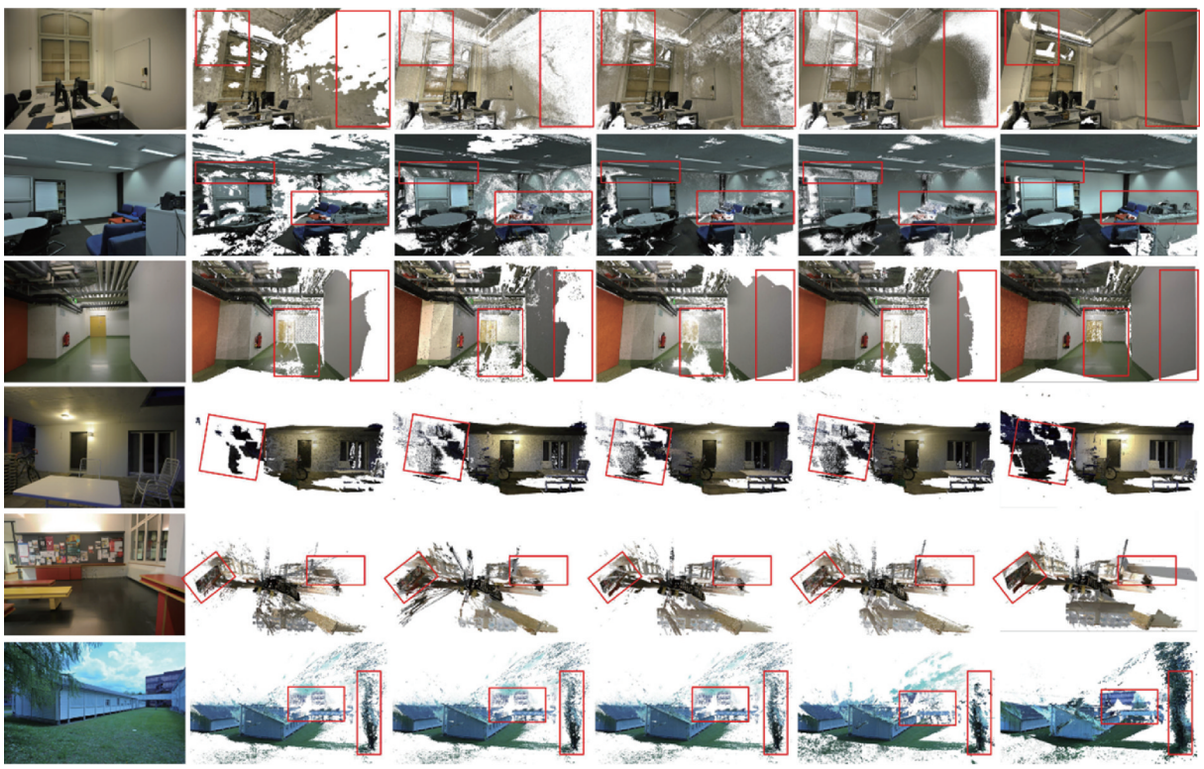

图 2　基于全局的图像投影方法定性对比

Fig. 2　Qualitative comparison of global based image projection methods ((a) original images; (b) IterMVS-LS(Wang et al. , 2022); (c) ACMMP; (d) HPM-MVS; (e) APD-MVS; (f) DVP-MVS(Yuan et al. , 2025))

统方法对纹理特征的依赖，更成为当前工业检测、自动驾驶等高精度三维感知任务的首选方案。

基于全局的图像投影方法的基础性工作是MVSNet，该方法奠定了此类方法基本原理和结构框架，后续方法大多是基于MVSNet的变体。MVSNet结构如图3所示，其基本流程如下：

1）特征提取。给定图像$\{\boldsymbol{I}_i\}_{i=1}^{N}$，每幅图像依次作为参考图像，其余图像作为源图像。通过共享卷积网络从多幅输入图像中提取对应的特征图，用于后续在不同视角进行匹配。

2）代价体构建。通过使用单应性矩阵，在三维空间中均匀采样若干深度假设，并计算光度一致性构建代价体作为深度可能性度量。具体来说，给定一对已知的相机内参$\boldsymbol{K}_r$, $\boldsymbol{K}_s$和外参$\boldsymbol{T}_r$, $\boldsymbol{T}_s$($\boldsymbol{T}_i=[\boldsymbol{R}_i, \boldsymbol{t}_i]$包含旋转矩阵和平移矩阵)，即可计算得到深度假设$d$对应的单应性矩阵$\boldsymbol{H}(d)$，以表示源图像和参考图像之间的坐标映射关系，具体过程为

$$\boldsymbol{H}(d)=\boldsymbol{K}_s\boldsymbol{R}_s\left(\boldsymbol{I}-\frac{(\boldsymbol{t}_r-\boldsymbol{t}_s)\boldsymbol{n}^{\mathrm{T}}}{d}\right)\boldsymbol{R}_r^{\mathrm{T}}\boldsymbol{K}_r^{-1} \tag{14}$$

式中，$\boldsymbol{n}$表示法线。

接着，将每个源图像映射得到的特征体与参考图像的特征体进行合并，构建集合$\{\boldsymbol{V}_i\}_{i=1}^{N}$，并通过计算特征方差得到代价体。具体为

$$C=\frac{1}{N}\sum_{i=1}^{N}\left(\boldsymbol{V}_i-\overline{\boldsymbol{V}}_i\right)^2 \tag{15}$$

3）代价体正则化。从图像特征计算得到的原始代价体可能受到噪声污染（例如非朗伯表面或物体遮挡），因此结合平滑约束做进一步处理。采用多尺度3D CNN（convolutional neural network）进行代价体正则化，并最终沿深度方向应用softmax操作进行概率归一化得到概率体，以最终衡量深度假设的置信度。

4）深度图推断。为了提供亚像素级别的深度，采用soft-argmin操作，通过计算数学期望来回归深度值。具体为

$$d=\sum_{d'=d_{\min}}^{d_{\max}}d'\times p(d') \tag{16}$$

式中，$[d_{\min}, d_{\max}]$定义了均匀采样深度假设的范围，$p(d')$是深度假设d'的概率。为避免正则化过程可能出现的过度平滑和模糊现象，应用深度残差学习网络，以参考图像作为指导，精细化得到最终深度图。

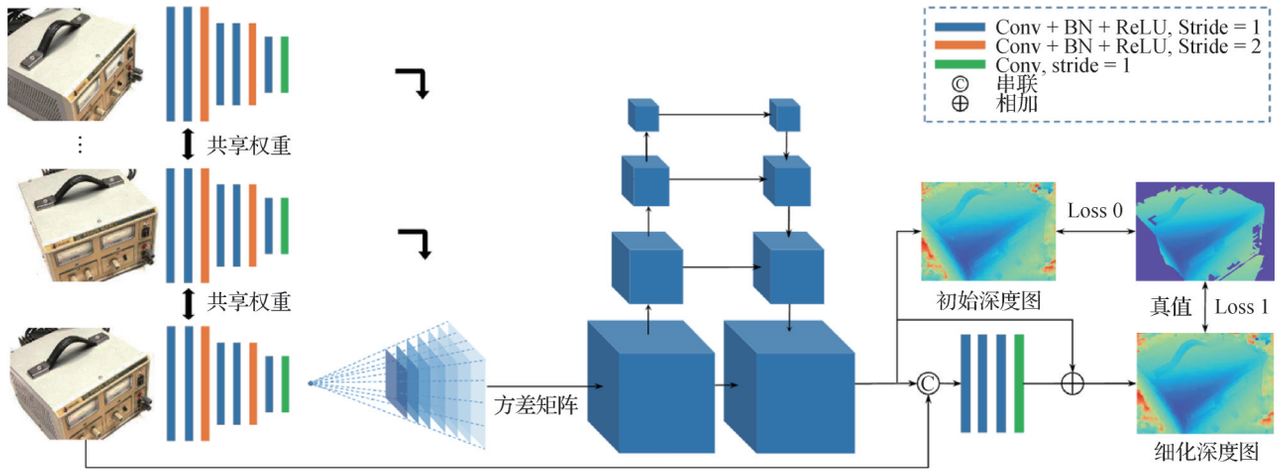

图3　MVSNet结构(Yao等,2018)

Fig. 3　Structure of MVSNet(Yao et al. ,2018)

为继续提升全局的图像投影方法的重建效果,后续工作提出了多种优化策略和技术。这些方法根据优化方向和技术特点,主要可分为基于计算成本优化的方法、基于深度推断改进的方法、基于注意力机制的方法、基于几何先验融合的方法以及基于无监督/自监督的方法。

2.2.1　基于计算成本优化的方法

基于全局图像投影的一个关键优势是代价体正则化,能够有效结合多视角图像信息,增强深度估计的准确性和一致性。然而,这一过程消耗大量内存,因为它处理的是三维体积,且随着图像分辨率的增加,内存需求呈立方增长。这种内存消耗成为了该类方法扩展到高分辨率场景的主要瓶颈之一。

为了解决这一问题,早期提出的粗到细逐步优化策略发挥了至关重要的作用。CasMVSNet(Gu等,2020)便是这一思路的经典代表。如图4所示,该方法通过级联三维代价体构建,从特征金字塔中提取多尺度特征。在初期阶段,采用较大尺度的语义二维特征和稀疏采样来构建代价体,从而实现较低的体积分辨率。随后,后期阶段利用前一阶段估计的深度图来自适应地调整深度假设的采样范围,并构建新的代价体,以应用更加精细的语义特征。通过这种自适应深度采样和特征分辨率调整,方法在显著降低计算时间和GPU内存消耗的同时,利用多层级的语义信息,提高了不同纹理目标的匹配精度,进而显著提升了最终深度估计的质量。在同一时期,CVP-MVSNet(Yang等,2020)和UCS-Net(Cheng等,2020)也提出了类似的思路。自此,粗到细逐步优化策略被广泛应用,成为后续研究的重要基础。

另一种常见的计算成本优化思路是采用循环神经网络。这种思路最初由R-MVSNet(Yao等,2019)实现,使用GRU沿深度方向顺序地对二维代价图进行正则化,通过减少同一时间内处理的维度来降低内存消耗。AA-RMVSNet(Wei等,2021)则采用RNN-CNN(recurrent neural networks—convolutional neural network)混合的方式,利用ConvLSTMCell模块,通过2D CNN的Encoder-Decoder结构聚合空间上下文信息并进行正则化。在沿深度方向的处理中,通过5个并行的RNN结构将前一个ConvLSTMCell的中间输出传递给下一个,从而在保证较低内存消耗的同时,进一步提升重建效果。

深度假设的采样数量也是影响计算成本的关键因素。GBiNet(Mi等,2022)旨在探索确保高精度所需的最小假设采样数量。原始的MVSNet被视为一种密集搜索方法,通过并行方式检查所有深度假设。相比之下,GBiNet将MVS问题表述为二分查找问题,将深度范围划分为两个区间,并通过比较逐步消除一半的搜索空间。为了增强模型的鲁棒性,GBiNet在预测范围外填充了两个误差容忍区间,以减少区间外预测错误;若某个搜索阶段对像素的预测发生了区间外错误,便停止该像素的前向传播。除此之外,GBiNet在每个搜索阶段即时更新网络参数,避免了以往方法中的多阶段梯度积累。通过这些策略,GBiNet在每个阶段仅需4个深度假设,从而显著降低了训练和推理过程中的内存消耗。实验结果表明,这种设计足以实现准确的深度预测。

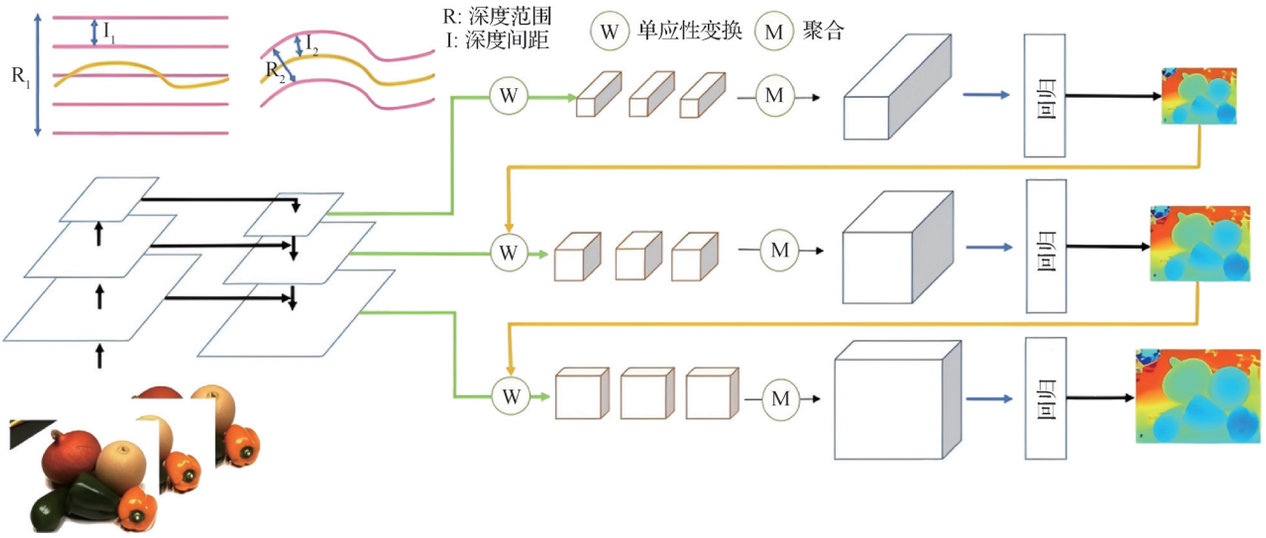

图 4　CasMVSNet 结构（Gu 等，2020）

Fig. 4　Structure of CasMVSNet (Gu et al. , 2020)

2.2.2　基于深度推断改进的方法

深度推断在学习型 MVS 中通常被视为回归或分类问题，尽管这两种表示方法均展现了出色的性能，但也存在明显的局限性。回归方法通过 soft-argmin 回归深度，理论上可实现亚像素级的精度，然而模型需要在复杂约束下学习权重组合，容易导致过拟合，并且权重组合的模糊性增加了收敛的难度。分类方法则通过选择概率最大深度假设，并利用交叉熵对概率体进行约束，但其离散预测限制了精确深度的推断。为克服这些问题，UniMVSNet（Peng 等，2022）设计了一种创新的统一表示方法，该方法通过在概率体上执行损失，并估计一个包含最佳深度假设位置及其与真实深度之间偏移的标签，从而有效融合回归和分类方法的优势。此外，传统 MVS 模型通常会密集采样深度假设，但仅有一个正样本，会导致样本不平衡问题。为了解决这一问题，UniMVSNet 扩展了焦点损失为统一焦点损失，使得模型能够处理连续深度标签，并在优化过程中加强对困难样本的关注，从而显著提高了模型的鲁棒性。

深度推断还依赖于高质量的深度假设。主流方法在进行深度假设采样时，仍沿用原始 MVSNet 的思路，采用固定的全像素深度范围和均匀的深度区间划分。然而，物体表面通常具有异构性，这种方式往往导致深度平面利用不充分，从而影响深度估计的精度。许多研究者探索了更有效的深度假设采样策略，例如，PatchMatchNet（Wang 等，2021a）将多尺度 PatchMatch 思想引入到基于端到端可学习的 MVS 框架中。在初始阶段，PatchMatchNet 将预设的深度范围划分为多个区间，并通过随机采样为每个像素生成初始深度假设。在后续阶段，它在前一个阶段得到的估计值为中心的小范围内进行采样。

2.2.3　基于注意力机制的方法

为了克服传统卷积神经网络在处理复杂数据时的局限性，尤其是在捕捉长距离依赖关系和全局上下文信息方面的不足，一些方法引入了注意力机制，以增强模型在细节恢复、特征聚合和长距离上下文建模中的能力。例如，TransMVSNet（Ding 等，2022a）设计了一个特征匹配模块，处理通过特征金字塔网络提取的图像特征，利用自注意力和跨视图注意力机制聚合图像内外的长距离上下文信息，从而降低特征匹配过程中的不确定性。MVSFormer（Cao 等，2022）利用预训练的视觉 Transformer（ViT）增强了特征学习的能力，能够通过 ViT 提供的有用先验信息学习更可靠的特征表示。经过微调的 MVSFormer 在层次化的注意力机制下，在 FPN（feature pyramid network）的基础上实现了显著的性能提升。

MVSFormer++（Cao 等，2024）在 MVSFormer 的基础上进行了进一步优化，借鉴了以往基于注意力机制的 MVS 方法的经验，旨在充分挖掘注意力机制的固有优势，以提升学习型 MVS 中各个关键模块的性能。与 MVSFormer 不同，MVSFormer++在多个方面进行改进，特别是在跨视图信息的处理和不同注意力机制的应用上。首先，MVSFormer++通过将跨视图信息注入到预训练的 DINOv2（dual-stage implicit

object-oriented network)模型中,进一步增强了MVS学习的效率和准确性。通过这种方式,模型能够更好地从多个视角融合信息,从而提高深度估计的鲁棒性。此外,MVSFormer++采用了不同的注意力机制来优化特征编码和代价体积正则化,分别聚焦于特征聚合和空间聚合,从而提升了模型在复杂场景中的性能。值得注意的是,MVSFormer++还揭示了一些设计细节对变换器模块在MVS中的性能影响,诸如归一化的3D位置编码、适应性注意力缩放、以及层归一化的位置等,都显著提升了模型的表达能力和稳定性。

2.2.4 基于几何先验融合的方法

基于几何先验融合的方法通过充分利用MVS任务中特有的一些几何特性,将其融入到MVS网络中,从而提升网络的性能。通过引入几何约束,能够有效减少不必要的计算复杂度,增强匹配精度,尤其是在处理复杂的视角变化和遮挡问题时,能够显著提高重建结果的质量。

极几何约束是传统立体匹配任务中常用于简化匹配搜索的复杂度。ET-MVSNet(Liu等,2023)基于极线几何原理,将非局部特征增强操作限制在极线对内,从而有效减少了计算开销并提升了匹配精度。该方法通过优化的搜索算法对二维特征图进行划分,并结合极线变换器在极线对之间执行特征增强,从而提高了特征匹配的效率与准确性。

模型中生成的中间结果也蕴含了丰富的几何结构信息。PFR-MVSNet(Zhao等,2023)通过动态结构感知模块,深入挖掘多视图特征中的几何信息,并在局部区域内建立空间结构关系,从而优化了特征学习和深度估计的效果。该模块有效利用中间结果中的几何信息,增强了网络对复杂几何结构的感知和建模能力,显著提高了重建精度。GeoMVSNet(Zhang等,2023b)通过双分支几何融合网络,将粗略阶段提取的几何信息与精细阶段的特征提取相结合,进一步提升了深度估计的几何感知能力。通过这种方式,网络能够更好地处理场景的几何结构,避免单纯依赖像素级匹配。

部分方法则通过引入外部获取的几何先验来辅助MVS模型,例如利用先进的单目法线模型预先获取图像的法线信息。GoMVS(Wu等,2024)采用几何一致传播模块,通过表面法线信息传播相邻代价,从而优化代价聚合过程,显著提升匹配精度和重建完整性,但其计算成本较高。GAP-MVSNet(Chen等,2025)将法线信息融入MVS的各个关键模块,设计了结构感知特征金字塔模块与空间几何增强正则化,隐式地学习目标的几何结构,以较低的计算成本实现高质量的重建效果。

2.2.5 基于无监督/自监督的方法

上述提到的方法均为有监督方法,并已取得了令人鼓舞的成果。然而,多视角深度数据集的收集非常困难,且监督数据并非始终可用,这限制了模型在实际应用中的泛化能力。因此,一些研究探索了无监督和自监督的MVS方法。

基于无监督的MVS方法通常是端到端的,并且主要依赖光度一致性假设。例如,Unsup-MVSNet(Khot等,2019)提出了第一个端到端无监督MVS框架,该方法通过预测的深度图将源图像反向映射得到重建图像,并强制执行重建图像与参考图像的光度一致性。此外,该方法还引入了结构相似性损失和深度平滑损失,以进一步提升性能。

然而,光度一致性在许多具有挑战性的区域效果不佳。M3VSNet(Huang等,2021)在此基础上引入了多尺度特征损失,通过从预训练的VGG-16网络获得的多尺度特征图作为线索,并采用法线-深度一致性来强化几何约束。CL-MVSNet(Xiong等,2023)则提出了一种新的双层对比学习方法,结合图像级和场景级对比分支来增强模型的上下文感知和表示能力,从而提高在低纹理和视角相关效应区域的鲁棒性。

基于自监督的MVS方法通常需要通过过滤和处理推断的深度图来获得可靠的伪标签,用于监督后续阶段的模型训练。这类方法通常需要复杂的预处理和微调,因此往往需要多个阶段进行训练。KD-MVS(Ding等,2022b)在这一类方法中具有代表性。该方法通过引入知识蒸馏机制,优化了自监督MVS的训练过程。具体而言,首先,教师模型基于光度一致性和特征一致性进行训练,生成初步的伪标签。随后,教师模型的知识通过概率编码转移到学生模型中,从而避免了传统方法中伪标签生成和处理的复杂性。

2.2.6 基于全局的图像投影方法小结

基于全局的图像投影方法通过构建统一的代价表达和空间一致性建模,显著提升了在复杂场景中的深度重建质量。这类方法以MVSNet及其变体为

代表，通过神经网络的特征表达和端到端优化，使深度图推断更加精准和稳定。其核心优势体现在：1）通过多视角特征融合和代价体正则化，有效解决低纹理区域、遮挡和强光变化等传统方法难以应对的挑战；2）借助深度学习框架实现从图像输入到深度输出的整体推理过程，显著提升重建稳定性；3）通过自适应深度采样和多尺度特征融合，在保证精度的同时优化计算效率。然而，该类方法仍面临三大技术瓶颈：1）计算资源消耗呈立方增长导致高分辨率场景处理困难；2）深度假设采样策略在物体边界和细节区域存在精度限制；3）现有基准数据集场景同质化问题制约方法泛化能力。

表 3 和表 4 系统展示了全局图像投影方法的设计演进与性能对比。

表 3 显示，基于计算成本优化的方法（如 CasMVSNet、GBiNet）通过逐步优化策略和轻量化设计，将内存消耗降低至传统方法的 1/8～1/4，但可能牺牲局部细节精度；基于深度推断改进的方法（如

表 3　基于全局的图像投影方法总结

Table 3　Summary of global based image projection methods

方法类别	设计思路	优点	缺点
基于计算成本优化	采用逐步优化策略，设计轻量化的网络结构	减少高分辨率场景重建所需计算资源	轻量化设计对重建质量提升有限
基于深度推断改进	改进深度估计的表示方式，优化深度假设采样策略	提高重建精度，改善细节区域重建质量	不规则的深度假设采样使得正则化过程更加复杂
基于注意力机制	强化长距离依赖关系和全局上下文信息的建模	增强复杂场景下的表现和鲁棒性	计算成本相对较高
基于几何先验融合	引入几何约束或外部几何信息	提高重建的几何一致效果和完整性	依赖高质量几何先验
基于无监督/自监督	生成高质量伪标签，学习数据的潜在结构或约束	减少数据依赖，提升泛化能力	重建效果一般且训练流程复杂

表 4　基于全局的图像投影方法定量对比

Table 4　Quantitative comparison of global based image projection methods

方法	方法类别	数据集	评价指标	对比方法	对比结果	实验结果
R-MVSNet	基于计算成本优化	DTU	Overall ↓	MVSNet	0.511	0.417
CasMVSNet	基于计算成本优化	DTU	Overall ↓	MVSNet	0.511	0.355
CVP-MVSNet	基于计算成本优化	DTU	Overall ↓	MVSNet	0.511	0.351
PatchMatchNet	基于深度推断改进	DTU	Overall ↓	CasMVSNet	0.355	0.352
AA-RMVSNet	基于计算成本优化	Intermediate	F_1-score ↑	CasMVSNet	56.42	61.51
TransMVSNet	基于注意力机制	DTU	Overall ↓	CasMVSNet	0.355	0.305
GBiNet	基于计算成本优化	Intermediate/Advanced	F_1-score ↑	CasMVSNet	56.42/31.12	61.42/37.32
UniMVSNet	基于深度推断改进	Intermediate/Advanced	F_1-score ↑	CasMVSNet	56.42/31.12	64.36/38.96
ET-MVSNet	基于几何先验融合	DTU	Overall ↓	GBiNet	0.298	0.291
ARAI-MVSNet	基于深度推断改进	Intermediate/Advanced	F_1-score ↑	GBiNet	61.42/37.32	65.89/41.52
GoMVS	基于几何先验融合	DTU	Overall ↓	GBiNet	0.298	0.287
MVSFormer++	基于注意力机制	Intermediate/Advanced	F_1-score ↑	GBiNet	61.42/37.32	67.18/41.70
GAP-MVSNet	基于几何先验融合	Intermediate/Advanced	F_1-score ↑	GBiNet	61.42/37.32	67.08/43.12
RRT-MVS	基于注意力机制	Intermediate/Advanced	F_1-score ↑	GBiNet	61.42/37.32	68.16/43.29

注：“↑”表示值越大越好，“↓”表示值越小越好。

UniMVSNet、ARAI-MVSNet)通过优化采样策略和损失函数设计,在DTU数据集上将Overall指标提升15%以上,但需应对非均匀深度分布带来的正则化挑战;注意力机制(如MVSFormer++、RRT-MVS(Jiang等,2025))和几何先验融合(如GeoMVSNet)的引入,则显著提升了复杂场景下的鲁棒性,使F_1-score在TNT Advanced子集达到43.29%。

表4的定量对比进一步验证了技术迭代的有效性:从MVSNet(0.511)到最新RRT-MVS(0.287),DTU数据集的Overall指标下降34%,表明重建精度持续提升;而TNT Advanced子集的F_1-score从31.12%提升至43.29%,增幅达40%,印证了方法在大规模场景中的进步。值得注意的是,当前方法在DTU和TNT数据集上的优异表现,主要受益于室内外建筑和小型人造物场景的密集标注,但对植被、动态场景等复杂环境的适应性仍待验证。

3 几何推理的多视角立体重建方法

在多视角三维重建任务中,几何推理方法通过建模场景中不同视角图像之间的几何关系,从而在三维空间中还原物体的结构信息。不同方法在几何信息建模的范式上存在显著差异,通常可以划分为全局几何推理方法与局部几何推理方法。前者通过统一的空间表示对整个场景进行优化与建模,强调多视角信息的全局一致性与整体结构的还原,例如基于体渲染的隐式表示方法。后者则侧重于对局部区域或离散点进行精细建模,依赖每个视角下的局部观测对物体形态进行补全,如显式的三维高斯投影表示方法。

这两类方法各有优劣:全局方法通常具有更强的视角一致性和拓扑完整性,但在细节还原和表达效率方面存在瓶颈;而局部方法具备更高的灵活性和细节表达能力,却在视角变化下易出现遮挡冲突或鬼影等一致性问题。因此,如何在全局一致性与局部细节表达之间取得平衡,成为当前三维重建方法设计的核心挑战之一。

3.1 基于全局的几何推理方法

基于全局的几何推理方法强调通过统一的空间体积或隐式函数对整个场景进行建模,以实现连续且高质量的三维重建。该类方法的核心思想是在多视角图像之间引入一个统一的全局几何表示,通常借助体素网格、辐射场(radiance field)或符号距离函数(signed distance function,SDF)等形式,在三维空间中建立像素与场景之间的映射关系。典型代表为NeRF(Mildenhall等,2022)提出的神经辐射场方法,通过训练一个坐标到颜色与密度的神经网络,实现从连续空间到图像观测的可微分渲染。其核心思想是:仅需多视角2D图像和相机位姿信息,就能重建出逼真的3D场景表示,并生成任意新视角的图像。NeRF通过将神经网络与经典计算机图形学的体积渲染技术结合,开创了3D场景表示的新范式。

如图5所示,NeRF的核心在于将3D场景表示为一个5D函数$F(\boldsymbol{x},\boldsymbol{d})\rightarrow(\boldsymbol{c},\boldsymbol{\sigma})$,式中$\boldsymbol{x}=(x,y,z)$是3D空间中的点,$\boldsymbol{d}=(d_x,d_y,d_z)$是视角方向,$\boldsymbol{c}=(r,g,b)$是颜色,$\boldsymbol{\sigma}$是体积密度。通过多层感知机(multilayer perceptron,MLP),NeRF将这一函数参数化,并结合可微分体积渲染技术,将神经网络与经典渲染流程无缝衔接。

这种隐式建模方式无需显式几何表示(如网格或点云),而是通过神经网络直接学习场景的几何和光照特性。NeRF的关键技术之一是体积渲染公式,它描述了光线穿过场景时的颜色累积过程。对于一

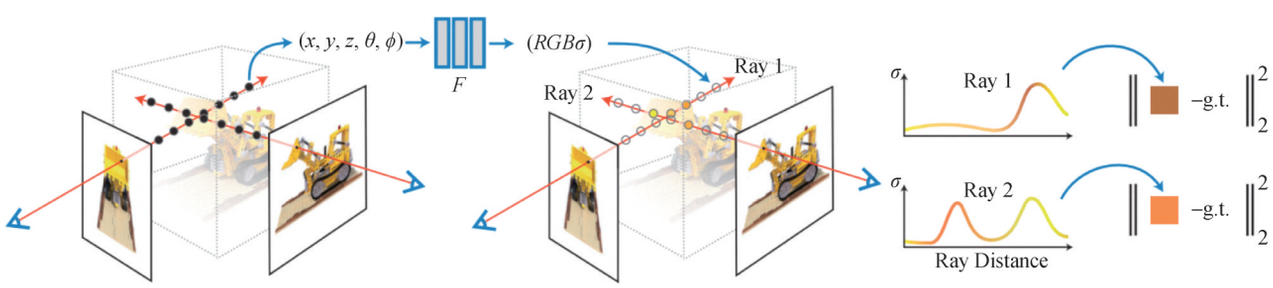

图5 NeRF结构(Mildenhall等,2022)

Fig. 5 Structure of neural radiance fields(Mildenhall et al.,2022)

条相机射线$r(t)=o+td$，其最终颜色$C(r)$由以下积分计算，具体为

$$C(r)=\int_{t_{\text{near}}}^{t_{\text{far}}}T(t)\cdot\sigma\big(r(t)\big)\cdot c\big(r(t),d\big)\,\mathrm{d}t \tag{17}$$

式中，$T(t)=\exp\left(-\int_{t_{\text{near}}}^{t}\sigma\big(r(u)\big)\,\mathrm{d}u\right)$是累积透射率。在实际计算中，这一积分通过分层采样策略离散化处理，分为粗采样和细采样两阶段，后者基于粗采样的概率分布重新采样，聚焦于高密度区域（如物体表面）。

NeRF 的 MLP 分为两个阶段：几何阶段输入 3D 坐标x，输出体积密度σ和 256 维特征向量$\boldsymbol{f}$；外观阶段将特征向量$\boldsymbol{f}$与视角方向d拼接后，输出颜色c。此外，为了增强模型对高频细节的建模能力，NeRF 引入了位置编码，将输入坐标和方向映射到高维正弦波空间。具体为

$$\begin{aligned}\gamma(v)=\big(&\sin\big(2^0\pi v\big),\cos\big(2^0\pi v\big),\cdots,\\&\sin\big(2^{L-1}\pi v\big),\cos\big(2^{L-1}\pi v\big)\big)\end{aligned} \tag{18}$$

式中，$L=10$（位置）或$L=4$（方向）。

NeRF 的训练目标是通过最小化光度损失$L=\sum_{r\in\mathbf{R}}\left|\hat{C}(r)-C_{\text{gt}}(r)\right|_2^2$来优化 MLP 参数。整个流程包括：生成相机射线、分层采样、MLP 预测密度与颜色、体积渲染生成像素值，并通过反向传播更新参数。尽管 NeRF 训练速度较慢且显存消耗大，但其首次实现了复杂场景的光栅化级真实感渲染，成为新视角合成领域的里程碑式工作。

3.1.1 隐式表示方法

全局的几何推理方法中，隐式表示方法通过神经网络直接学习 3D 场景的几何和外观信息，无需显式的 3D 结构（如网格或点云），其核心思想是将 5D 输入（3D 位置 + 2D 视角方向）映射为颜色和体积密度。这一类方法在 NeRF 的早期发展中占据主导地位，但受限于训练效率和显存消耗，逐渐被显式或混合表示方法取代。Baseline NeRF（Mildenhall 等，2022）是隐式表示的开山之作，其通过多层感知机（MLP）直接建模 5D 函数$F(x,d)\rightarrow(c,\sigma)$，结合可微分体积渲染技术，首次实现了复杂场景的高质量新视角合成。然而，Baseline NeRF 的训练速度较慢且难以处理抗锯齿问题。为此，mip-NeRF（Barron 等，2021），引入锥形采样（cone sampling）和集成位置编码（integrated positional encoding，IPE），将光线从传统线性采样扩展为锥形区域采样，并通过多变量高斯分布近似积分，显著提升了低分辨率下的图像质量。此外，mip-NeRF 还设计了新的正则化策略以防止几何漂浮和背景崩溃，成为隐式表示方法的重要里程碑。

尽管隐式表示在几何建模上具有灵活性，但其对反射表面的处理能力有限。为解决这一问题，Ref-NeRF（Verbin 等，2022）通过参数化反射方向与局部法向量之间的关系，将辐射场拆分为漫反射和镜面反射分量。具体而言，Ref-NeRF 将密度 MLP 改为方向无关的网络，输出漫反射颜色、镜面反射颜色、粗糙度和表面法向量，并通过反射视角方向优化镜面反射路径。该方法在复杂材质（如玻璃、金属）的场景中表现出色，显著提升了反射区域的渲染精度。

隐式表示方法的优势在于其对复杂拓扑的适应性，但其高计算开销限制了大规模应用。然而，隐式表示方法的存储和计算瓶颈仍难以彻底解决。随着 Gaussian splatting（Kerbl 等，2023）的提出，显式表示方法逐渐成为主流，隐式表示转向特定任务（如动态建模、符号化几何）。尽管如此，隐式表示方法在小规模场景和高质量渲染中仍具不可替代性，其核心思想（如体积渲染、符号距离函数约束）持续影响后续研究。

3.1.2 显式表示方法

显式表示方法通过直接存储 3D 点云、网格或体素等显式数据结构，结合神经网络优化，显著提升了渲染效率和几何可控性。这类方法通常利用显式结构的查询速度优势，同时借助神经网络处理复杂细节，是当前基于全局的几何推理方向中，NeRF 研究的主流方向之一。然而，显式表示的灵活性受限于数据结构的设计，难以建模复杂拓扑（如多孔物体），其应用更多集中在大规模场景和实时渲染任务中。

Plenoxels（Fridovich-Keil 等，2022）是早期显式表示的典型工作，其核心创新在于将场景体素化（voxelization）并结合球谐函数（spherical harmonics，SH）压缩颜色信息。Plenoxels 通过预训练的 NeRF 模型提取体素密度和球谐系数，并存储为稀疏体素网格。在推理阶段，Plenoxels 仅需查询体素参数，避免了 MLP 的高计算开销，使得推理速度比 Baseline NeRF 快数倍。此外，Plenoxels 通过自适应体素化策略（如动态调整体素分辨率）进一步减少存储需求。

尽管Plenoxels在效率上取得显著进展，但其固定分辨率的体素网格限制了细节建模能力，且难以适应动态场景变化。

Plenoctree（Yu等，2021）进一步优化了显式体素表示，其核心贡献在于构建基于八叉树（Octree）的稀疏体素结构。Plenoctree通过球谐函数编码颜色，并结合深度学习优化体素参数。具体而言，Plenoctree利用八叉树结构动态调整体素分辨率，在高密度区域保留精细体素，而在空旷区域使用稀疏表示。这一设计显著降低了存储开销（相比Plenoxels减少约80%），同时保持了较高的渲染质量。此外，Plenoctree引入深度传播策略，通过多尺度体素共享参数进一步提升效率。然而，Plenoctree仍需依赖预训练NeRF模型生成初始体素参数，限制了其在动态场景中的应用。

PointNeRF（Xu等，2022）将显式点云与神经网络结合，通过预训练的深度估计网络生成点云，并利用PointNet（Qi等，2017）回归密度和颜色。具体而言，PointNeRF首先通过代价体生成密集点云，再通过PointNet学习点云的局部密度和颜色特征。这一方法跳过了传统NeRF的MLP渲染过程，通过点云查询加速渲染，推理速度比Baseline NeRF快3倍以上。然而，PointNeRF的点云生成依赖于预训练深度模型，导致其在复杂场景中的几何精度受限。此外，点云的显式存储需求较高，限制了其在大规模场景中的扩展性。

显式表示方法的核心优势在于其高效性和几何可控性，但其灵活性受限于数据结构的设计。例如，Gaussian splatting通过点云实现高速渲染，但难以建模复杂拓扑；Plenoctree通过八叉树优化存储，但固定分辨率限制了细节表达。为解决这些问题，研究者探索了全局的几何推理方法下动态显式结构和混合优化策略。例如，Neuralangelo（Li等，2023）结合网格优化和神经网络，提升了几何编辑能力；Plenoxels通过球谐函数压缩颜色，降低了存储需求。尽管显式表示方法在效率上取得显著进展，其在动态场景和复杂拓扑建模中的局限性仍需进一步突破。

3.1.3 混合表示方法

混合表示方法通过结合显式数据结构（如体素、哈希表）与隐式神经网络（如MLP），旨在平衡效率与精度。这类方法通常通过显式结构预计算部分场景信息，减少神经网络的计算负担，从而加速训练和推理。然而，显式结构的设计需要权衡存储开销与灵活性，而神经网络仍需承担复杂细节的建模任务。

SNeRG（Hedman等，2021）是早期混合表示的代表性工作，其核心思想是将NeRF预计算到稀疏体素网格（sparse voxel grid）中。具体而言，SNeRG通过预训练的NeRF模型提取场景的漫反射颜色、密度和特征向量，并将其存储在体素网格中。在推理阶段，MLP仅需预测镜面反射颜色，并通过alpha混合生成最终像素值。这种设计将大部分计算量转移到预计算阶段，使得推理速度比Baseline NeRF快约3 000倍，同时保持了高质量渲染效果。然而，SNeRG的显式网格存储需求较高，且难以适应动态场景的变化。

Instant-NGP（Müller等，2022）的核心贡献在于引入多分辨率哈希编码（multi-resolution Hash encoding）。Instant-NGP将场景坐标映射到哈希表中，通过动态调整哈希表的分辨率和特征维度，实现灵活的细节控制。这一设计不仅保留了MLP的高精度建模能力，还大幅提升了训练和推理速度（在合成NeRF数据集上，训练时间从数小时缩短至数分钟）。此外，Instant-NGP通过指数步进（exponential stepping）和空域跳过（empty space skipping）优化采样过程，进一步加速了体积渲染。该方法成为混合表示领域的里程碑，为后续显式表示（如Gaussian splatting）的提出奠定了基础。

混合表示方法的核心优势在于其对显式结构和隐式网络的协同优化。例如，PlenOctree通过球谐函数压缩颜色信息，Instant-NGP通过哈希编码提升效率，而KiloNeRF（Reiser等，2021）通过分块策略降低计算复杂度。然而，这些方法仍面临显式结构存储开销大、灵活性受限的挑战。随着Gaussian splatting的提出，显式点云表示逐渐成为主流，混合表示方法的应用场景逐步转向特定任务（如大规模场景建模）。尽管如此，混合表示方法在NeRF早期发展中起到了承上启下的作用，其核心思想（如分层优化、哈希编码）仍对当前研究具有重要参考价值。

3.1.4 符号化表示方法

符号化表示方法通过数学符号（如符号距离函数、占用场）定义场景几何，并结合神经网络进行参数化优化。这类方法的核心优势在于其几何语义的明确性，能够直接定义表面或空间点的归属性，但需依赖特定的符号先验设计（如SDF形式）。符号化表

示通常与隐式表示方法交叉，但在动态场景建模和物理模拟中展现出独特优势。

NeuS（Wang等，2021b）是符号化表示的代表性工作，其核心创新在于结合符号距离函数（SDF）与体积渲染框架。NeuS通过Eikonal约束强制SDF梯度的单位长度，确保表面定义的准确性。具体而言，NeuS将SDF定义为神经网络的输出，并通过可微分渲染联合优化SDF和体积密度。在渲染阶段，NeuS利用SDF定义的表面进行光栅化，而非传统体积积分，显著减少了计算量。此外，NeuS引入双目约束策略，通过多视角深度估计减少几何歧义。这一方法在表面重建精度上优于隐式方法，但其显式网格的生成依赖于SDF优化，计算复杂度较高，且难以处理动态场景。

Occupancy Networks（Mescheder等，2019）通过占用场（occupancy field）定义空间点是否属于物体。该方法将3D坐标输入到神经网络，输出一个二分类概率值（0表示空旷，1表示占用），从而构建隐式表面。Occupancy Networks的优势在于其简单性和灵活性，能够建模复杂拓扑结构（如多孔物体）。然而，占用场的离散预测限制了精确几何的推断，且训练过程中需密集采样空间点，导致效率较低。为解决这一问题，后续研究如Neural Implicit Surfaces（Yariv等，2021）引入梯度估计策略，通过SDF的梯度计算表面法向量，从而提升几何重建的鲁棒性。MVSNeRF（Chen等，2021）提出了一种结合多视图立体与神经辐射场的通用型NeRF重建框架，能够仅依赖3幅输入图像进行快速几何感知渲染。MVSNeRF引入平面扫描代价体构建几何感知编码体积，并通过3D CNN与MLP解码器共同回归体素密度与方向相关的颜色，从而实现端到端的物理一致体积渲染。该方法显著减少了对逐场景优化的依赖，在DTU、LLFF和Synthetic等数据集上展现了良好的泛化能力。

符号化表示方法的核心优势在于其几何语义的明确性，能够直接定义表面或空间点的归属性，便于与物理模拟或同步定位与地图构建（simultaneous localization and mapping，SLAM）任务结合。例如，NeuS在SLAM中通过SDF定义表面，提升了位姿估计的鲁棒性；NeRFReN（Guo等，2022）通过符号化路径约束优化反射场景，解决了传统NeRF在复杂材质建模中的不足。然而，符号化表示方法需设计复杂的符号函数（如SDF、占用场），且依赖先验知识，限制了其通用性。此外，符号函数的优化通常需密集采样空间点，导致计算复杂度较高。

3.1.5 基于全局的几何推理方法小结

基于全局的几何推理方法通过统一的空间体积或隐式函数在多视角图像间建立全局几何表示，实现了高质量的三维重建，尤其在处理复杂场景时表现出显著优势。NeRF及其衍生方法成功将体积渲染与神经网络相结合，提供了逼真的新视角图像合成。然而，这类方法仍然面临训练速度慢和显存消耗大的问题，限制了其在高分辨率、大规模场景中的应用。虽然显式和混合表示方法的提出有效提升了渲染效率，但它们仍受到计算资源和存储需求的制约。未来研究可以聚焦于优化计算资源的使用，如引入高效的采样策略和加速训练过程，同时解决复杂场景中对细节建模的挑战。

表5和表6系统对比了不同几何建模方法在图像质量与训练效率方面的表现及其设计要点。从定量对比可见，NeRF及其改进版（如mip-NeRF、ref-NeRF）在PSNR和SSIM上表现优越，但训练时间普遍超过2 h，显存消耗大，难以满足实时需求；显式方法如Plenoxels和PlenOctree通过体素化和八叉树结构兼顾渲染效率与几何可控性，适用于实时场景，但对复杂拓扑适应性较弱；混合方法（如Instant-NGP和TensoRF（Chen等，2022））通过多分辨率哈希编码或张量分解策略，在保持较高渲染质量的同时，将训练时间压缩至分钟级，展现出良好的实用性；符号化方法如MVSNeRF虽然训练速度快（约15 min），但合成质量略低，更适合几何感知任务。4类方法在精度、效率与适用场景间形成互补：隐式表示擅长建模动态场景但不可编辑；显式表示渲染高效但拓扑适应性差；混合方法平衡效率与精度；符号化表示强调几何语义但依赖复杂先验。

3.2 基于局部的几何推理方法

随着显式建模方法的发展，基于局部的几何推理方法逐渐受到关注。其中3DGS作为代表性方法，后续方法大多是基于3DGS的变体。具体地，3DGS通过在三维空间中构建一组具有空间位置、朝向、尺度和颜色等属性的高斯点，显式表示场景的局部结构。该方法不再依赖全局一致的体素网格或隐式场，而是利用每个高斯点独立建模其所在区域的局部几何与外观信息，实现高效渲染与三维重建。3DGS的结构如图6所示，其基本流程如下：

表5　基于几何推理图像投影方法总结

Table 5　Summary of geometric reasoning based image projection methods

方法类别	设计思路	优点	缺点
基于隐式表示	通过神经网络直接学习5D场景函数，结合可微分体积渲染	高精度建模复杂拓扑，适应动态场景	训练速度慢，显存消耗大，难以提取显式几何
基于显式表示	使用3D点云显式数据结构存储场景信息，结合神经网络优化	高效渲染，几何可控性强，兼容SLAM和点云	灵活性受限，难以建模复杂拓扑
基于混合表示	结合显式结构（体素/哈希表）与隐式网络，预计算场景信息	平衡效率与精度，保留MLP的细节建模能	存储需求较高，需权衡分辨率与灵活性
基于符号化表示	通过数学符号（如SDF）定义几何，神经网络负责参数化	几何语义明确，支持复杂材质建模（如反射）	依赖符号函数设计，优化复杂度较高

表6　基于几何推理的图像投影方法定量对比

Table 6　Quantitative comparison of geometric reasoning based image projection methods

方法	类型	PSNR/dB	SSIM	LPIPS	训练时间
NeRF	隐式	31.01	0.947	0.081	> 12 h
mip-NeRF	隐式	33.09	0.961	0.043	~ 3 h
ref-NeRF	隐式	35.96	0.967	0.058	–
Plenoxels	显式	31.71	0.958	0.049	–
PlenOctree	显式	31.71	0.958	0.053	> 12 h
FastNeRF（Stephan等，2021）	显式	29.97	0.941	0.053	> 12 h
SNeRG	混合	30.38	0.950	0.05	> 12 h
Instant-NGP	混合	33.18	–	–	–
TensoRF	混合	31.56～33.14	0.949～0.963	–	–
MVSNeRF	符号化	27.07	0.931	0.163	约15 min

注：“–”表示对应缺少相关信息。

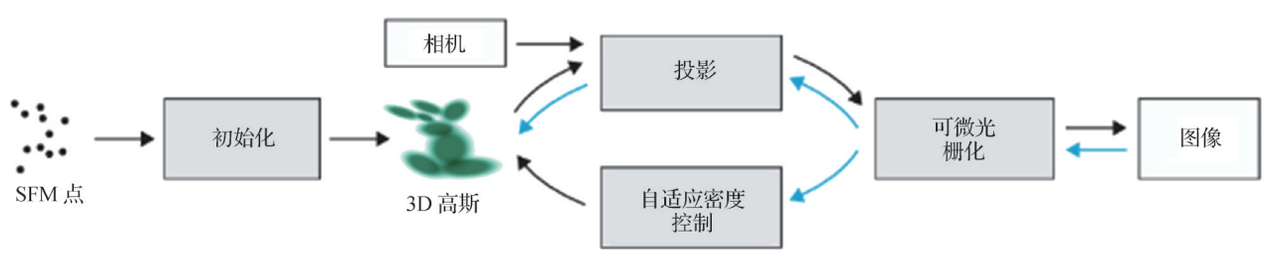

图6　3DGS结构

Fig. 6　Structure of 3D Gaussian splatting

1）高斯点初始化。给定带相机位姿的多视角RGB图像 $\boldsymbol{I}_i(i=1,\cdots,N)$，初始化生成一组可学习的具有三维位置、朝向、尺度和颜色等属性的初始高斯点，每个点记做 $\boldsymbol{G}=\{\boldsymbol{x},\boldsymbol{r},\boldsymbol{s},c,\alpha\}$，其中，$\boldsymbol{x}$ 表示中心位置，$\boldsymbol{r}$ 表示旋转，$\boldsymbol{s}$ 表示尺度，c 为颜色，α 为不透明度。

2）高斯空间分布建模。每个高斯点在三维空间中被建模为一个具有中心位置 x 和协方差矩阵 $\boldsymbol{\Sigma}$ 的高斯函数，其密度分布定义为

$$G(\boldsymbol{X})=\exp\left(-\frac{1}{2}(\boldsymbol{X}-\boldsymbol{x})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{X}-\boldsymbol{x})\right) \tag{19}$$

式中，$\boldsymbol{\Sigma}$ 控制高斯在空间中的尺度与方向。为了简化优化过程，提升建模灵活性与优化效率，$\boldsymbol{\Sigma}$ 通常被参数化为旋转矩阵 $\boldsymbol{R}$ 和缩放矩阵 $\boldsymbol{S}$ 的组合。具体为

$$\boldsymbol{\Sigma} = \boldsymbol{RSS}^{\mathrm{T}}\boldsymbol{R}^{\mathrm{T}} \tag{20}$$

3)可微分投影渲染。在可微渲染阶段，每个高斯点根据相机投影变换被映射到二维图像平面，计算其在像素上的影响。其二维协方差矩阵$\boldsymbol{\Sigma}$可由视图矩阵$\boldsymbol{V}$(从世界坐标到相机坐标)与透视投影的雅可比矩阵$\boldsymbol{J}$计算得到。具体为

$$\boldsymbol{\Sigma}' = \boldsymbol{JV\Sigma V}^{\mathrm{T}}\boldsymbol{J}^{\mathrm{T}} \tag{21}$$

2D像素的最终颜色通过对所有投影到该像素上的高斯点集合进行深度排序，并采用α—混合方式计算。具体形式为

$$C = \sum_{i \in N} c_i \alpha_i \prod_{j=1}^{i-1}\left(1 - \alpha_j\right) \tag{22}$$

式中，N表示投影到该像素的高斯集合，c_i为颜色，α_i为不透明度权重，依据o_i，$\boldsymbol{\Sigma}_i'$及像素距离计算。

4)高斯参数优化。所有高斯点的参数通过最小化与输入图像之间的重建误差进行端到端优化。常见的优化目标包括像素级L2损失、SSIM或LPIPS等感知损失函数，训练过程中不断调整高斯点的空间位置、方向、尺度和外观属性，以逐步提升渲染图像与真实图像之间的一致性。

由于3DGS采用局部建模范式，其在细节表示方面具备更高的灵活性与效率，但也因此更加依赖于视角下的局部观测。在多视角融合过程中，若缺乏有效的约束机制，高斯点在不同视角下可能出现冗余、漂移或对齐误差，进而导致鬼影等伪影现象。特别是在存在遮挡或图像质量不一致的场景中，这类方法在几何一致性建模方面面临不小挑战。此外，局部的几何推理方法下的原始方法在点云结构、语义表达、动态建模与多模态感知等方面仍存在一定局限。因此，如何在保留局部建模优势的同时，全面提升重建效率、表达能力与适应性，已成为该类方法持续演进的核心方向。

3.2.1 基于结构稀疏化优化的方法

为了在保证渲染质量的前提下，降低冗余计算成本，基于结构稀疏化的优化策略逐渐成3DGS研究中的核心方向。原始3DGS方法在训练过程中不断生长新的高斯点以拟合观测图像，虽然这种策略有助于细节表达，但也导致了点数量膨胀、内存开销大、推理速度慢等问题，尤其在跨视角融合时容易产生重影。因此，如何在局部的几何推理中通过稀疏化机制实现高效、紧凑的场景表示，是该方向改进的关键目标。

目前的稀疏化策略主要分为两类：一类侧重于后处理剪枝，通过对训练完成后的高斯集合进行压缩、融合或采样；另一类则强调稀疏感知建模机制，在建模过程中引导高斯点分布向几何或语义显著区域聚集，从源头控制点的生成密度。

在后处理剪枝方面，Speedy-Splat(Hanson等，2025a)提出了一种高效的模型压缩与渲染加速方法，结合光栅化剪枝与训练后点融合策略，实现大量高斯点削减，在保持渲染质量的同时将推理速度提升数倍。类似地，PUP-3DGS(Hanson等，2025b)设计了基于高斯点敏感度的剪枝机制，并结合微调过程保持稳定性，在大幅压缩的前提下支持在边缘设备上高效运行。此外，LP-3DGS(Zhang等，2024)引入可微掩码机制，在训练过程中学习每个高斯点的重要性分布，实现自动化压缩，不依赖手动设定阈值。

相比之下，另一类方法更关注稀疏感知能力的建模设计。Edge-GS(Chelani等，2025)通过在二维图像中提取边缘信息，并利用这些边缘对高斯点进行空间引导优先分布，使得投影至三维空间的高斯点集中于轮廓与高频结构区域。此外，Dropout-GS(Xu等，2025)引入随机丢弃机制模拟剪枝效应，并结合边缘增强模块增强点集结构稳定性。

尽管已有多项工作取得显著成果，稀疏化过程仍面临如融合策略设计、属性连续性保持以及剪枝后误差控制等挑战。例如，如何在合并高斯点时保留颜色、法线与尺度的局部一致性，是影响模型稳定性的关键因素。部分研究也引入点间图结构或注意力机制，以辅助保留重要结构信息，缓解压缩过程的性能损失。

3.2.2 基于显式几何一致性约束的方法

尽管3DGS在建模效率与细节表达上取得显著进展，其局部建模范式仍容易受到视角漂移、遮挡误差与几何漂浮等问题的影响。为提升跨视角一致性与空间结构可靠性，近年来多项研究引入显式的几何一致性约束机制，旨在通过深度、法线、多视角投影等几何监督信号，加强高斯点的空间位置与方向约束，从而提升基于局部的几何推理方法下3D重建质量与稳定性。

早期方法如GScream(Wang等，2025)在原始3DGS基础上引入单目深度图作为监督信号，并通过交叉注意力机制确保相邻视角下的特征表达保持一致。该方法有效缓解了高斯点漂浮与边界不清的问

题，在保持端到端训练框架的同时，引入了强几何先验，提升了渲染一致性。

随后，DN-Splatter（Turkulainen等，2024）进一步引入深度图与法线估计作为联合约束。在训练阶段，深度正则化项用于约束高斯点的三维位置，法线光滑项则保持点云表面结构的连续性。该方法可在无需额外网格建模的前提下，实现高质量的曲面重建，并自然支持从高斯点集生成连续网格，增强了结构表达能力。

MVG-splatting（Li等，2026）提出从多视角图像中自适应引导点密度分布，通过基于深度误差的稀疏性控制机制，引导高斯点分布向几何不一致区域集中，从而提升跨视角几何对齐性能。该方法将多视角一致性作为核心优化目标，融合光度一致性、投影深度误差与边缘梯度，显著提升了在非刚体场景中的稳健性。

综上所述，基于显式几何一致性约束的方法通过引入多种几何监督信号，从点的位置、朝向、密度和连接性多个维度优化3DGS建模过程。这类方法有效弥补了局部建模范式在空间一致性方面的缺陷，是当前提升3DGS重建质量的关键方向之一。未来工作可进一步探索弱监督或自监督条件下的几何一致性建模，以降低对外部深度数据或多视角图像的依赖。

3.2.3　基于语义先验引导的方法

随着3DGS在场景建模和渲染中的广泛应用，其可编辑性与语义理解能力成为进一步拓展应用边界的重要突破口。传统3DGS主要关注几何与颜色属性的高质量表达，缺乏对语义信息的建模与控制，限制了其在三维编辑、语义分割和人机交互等场景中的应用能力。近期，一系列研究通过引入可控机制或语义先验，在高斯点属性、结构组织和渲染过程中融入语义信息，引导模型实现更精准、更高效的结构表达与控制。

Semantic Gaussians（Guo等，2025a）是该领域的代表性工作之一。该方法首次将开放词汇语义引入3DGS，通过将预训练2D视觉模型生成的语义嵌入映射到高斯点，实现点级语义编码。并辅以3D稀疏卷积网络，在训练过程中优化语义一致性，使高斯点在三维空间内按语义类别聚集。这一机制不仅提升了结构表达力，还使得3DGS具备语义查询、对象定位、编辑等功能，显著增强了其在语义驱动任务中的实用性。

在细粒度结构理解方面，Gaussian Grouping（Ye等，2025）引入了身份编码属性，将每个高斯点映射到低维语义空间中，并通过SAM（segment anything model）提供的2D掩膜在渲染过程中进行语义对齐。这一机制使得同一物体或语义区域的高斯点在三维空间中自动聚集，形成一致的语义簇结构。该方法不仅能够实现对象级分割与遮挡区域的结构化处理，还在多视角融合过程中显著提升了语义一致性与空间聚合度，从而支持精细的三维编辑操作与结构对齐。

此外，SAGA（Cen等，2025）则从交互语义感知的角度出发，提出了一种支持任意区域、任意类别分割的高斯感知机制，将开源2D分割模的语义能力蒸馏到3D高斯特征空间中，并结合软尺度门策略，引导高斯点在不同空间尺度上学习语义区分能力。与静态语义聚类方法如Gaussian Grouping不同，SAGA强调实时交互与动态响应能力，支持点击、框选等提示操作，在用户输入后可快速完成所选区域的语义分割与高斯聚焦，极大提升了模型在编辑、增强现实与人机交互场景中的适用性。

总体来看，基于语义先验或控制机制的3DGS扩展方法正推动从几何重建向语义可控表达的发展。它不仅增强了模型的可解释性和交互能力，也为下游任务如三维语义分割、人机交互与编辑型生成等提供了坚实基础。未来研究可进一步探索更精细的语义属性建模、更高效的语言-高斯对齐策略以及多模态条件下的统一建模框架，以实现真正意义上的语义驱动三维表示。

3.2.4　基于渲染模型增强的方法

传统3DGS在几何建模与实时渲染方面已展现出强大优势，其渲染机制仍依赖于显式编码的颜色与透明度属性，难以充分建模纹理细节、复杂材质与真实光照。为弥补显式建模的表达能力不足，近期研究逐步引入神经渲染模块，将可学习的隐式特征与高斯结构融合，构建显式—隐式协同的混合渲染框架。这些方法在保持高斯点高效几何投影特性的基础上，借助神经网络增强颜色预测、光照建模与细节表达能力，显著提升了图像重建质量与视觉真实感。

例如，Normal-GS（Wei等，2024）在渲染路径中引入法向量并结合设计的集成定向照明向量，实现

了高斯点与光照方向的显式交互建模，有效增强了镜面高光区域的表现力。这种显式几何与神经光照模型的结合在提升材质还原度的同时，增强了高斯点几何分布的稳定性。Spec-Gaussian（Yang 等，2024a）则进一步提出基于各向异性高斯函数（anisotropic spherical Gaussian，ASG）的神经外观模型，用以替代传统球谐函数，以表达更加精细的方向性反射特征。

在几何与渲染的深度协同上，GSDF（Yu 等，2024）通过并行训练 3DGS 与 SDF 神经网络，将显式点云结构与隐式表面建模融合，同时设计跨通道特征桥机制，实现了神经场重建与高斯渲染路径的统一优化。该框架兼顾结构连续性与渲染精度，显著提升了几何边界表达与深度一致性，是显式—隐式融合建模的重要实践。

这些方法还启发了更广泛的结构扩展方向，例如引入神经体积模块补充纹理细节、使用注意力机制建模透明度分布等。尽管融合神经模块会带来训练与推理的额外计算开销，但其在视觉质量与结构保真上的提升，为 3DGS 在高质量图形合成、虚拟现实和交互式编辑等应用中的拓展提供了坚实基础。

3.2.5 面向动态场景的拓展方法

3DGS 等方法主要面向静态场景建模，其渲染和优化流程基于固定相机和静止物体的假设。然而，在实际应用中，诸如动态人体、流体、交通以及动作捕捉等场景普遍具有显著的时间变化特性。为扩展 3DGS 在动态环境下的建模能力，研究者提出了一系列支持时序建模与变形表达的方法，构成面向动态场景的拓展方向。

动态扩展方法的核心挑战在于：高斯点需要在时间维度上具有一致性，同时能表达形变、遮挡、运动模糊等现象。这要求高斯表示不仅能在空间中准确建模，还需在时间上进行跟踪、插值或重建。因此，大多数方法通过引入时间编码、点轨迹建模或可学习的变形场等机制，使高斯点具备动态特性。

Deformable 3D Gaussians（Yang 等，2024b）是早期代表性工作之一，采用可学习的形变场，对每个高斯点的中心位置进行时序更新。该方法在单目视频输入下实现高保真动态重建，并引入平滑训练机制来缓解相机姿态误差造成的变形抖动。延续此方向，Motion-aware 3D Gaussian splatting（Guo 等，2025b）进一步引入基于 2D 光流的运动引导机制，将像素级运动嵌入模型约束形变过程；此外还提出时序辅助模块以加强动态区域的时间一致性。这两种方法显著减少了运动模糊与帧间漂移，并在单视角视频重建任务中表现突出。

为解决传统方法中时空耦合严重、边界模糊等问题，STDR（Li 等，2025）引入了一种可插拔的解耦框架，利用时空权重与独立形变场对高斯点进行分组建模，解耦了时间与空间高斯分布。这种方式不仅缓解了不同区域运动带来的不一致性，还提升了动态物体的边界清晰度和几何保真度。

此外，SC-GS（Huang 等，2024）提出以稀疏控制点为核心的动态建模框架，针对高斯点在时间轴上的连续变形问题，设计了基于控制点轨迹的可学习形变场。具体而言，SC-GS 通过一组稀疏、可控的三维控制点捕捉全局运动趋势，并借助一个形变 MLP 将控制点的运动映射至高斯场中的所有点，从而实现高斯点的时序变换与形变表达。该方法在保持高渲染效率的同时提供了良好的可编辑性，可支持复杂对象的动态重建与形变动画等任务。

面向动态场景的高斯建模正逐步从简单的时间内插与刚性偏移，发展为更复杂的时空联合建模与可控形变优化方向。通过引入时间维度的显式建模、运动感知机制以及解耦设计，研究者有效提升了 3DGS 在真实动态环境中的表达能力与适应性，为其在动态重建、增强现实与动作捕捉等领域的应用奠定了技术基础。

3.2.6 结合多模态信息的增强方法

在实际场景中，单一图像模态往往难以全面捕捉场景中的几何、运动与语义信息，特别是在低纹理、光照复杂或动态遮挡等条件下。为此，近期研究逐步引入多模态信息进行协同建模，结合视觉、惯性和语言等异构信号以增强高斯表示的鲁棒性与任务适应性，形成多模态驱动的 3DGS 扩展路径。

在传感器协同方面，VINGS-Mono（Wu 等，2025）和 Gaussian-LIC（Lang 等，2025）分别将惯性测量单元（inertial measurement unit，IMU）与激光雷达（light detection and ranging，LiDAR）信息引入高斯建图流程，实现了对相机位姿与空间结构的联合建模。相比传统仅依赖几何一致性的优化路径，这类方法通过外部物理传感器提供稳定的先验约束，提升了系统在光照剧烈变化或动态场景下的稳定性与追踪能力，广泛适用于 SLAM 与移动建图任务。

在音频控制领域，GSTalker（Chen等，2024）通过将音频信号作为形变驱动条件，构建了语音同步的面部动画系统，在高斯框架中实现了口型驱动与面部表情生成。进一步地，SynGauss（Zhou等，2025）引入注意力机制细化局部表达，提升了口型拟合精度与渲染一致性。这类“语音驱动”的建模方式突破了图像模态的时空限制，体现出高斯点模型在语义控制与可编辑内容生成方面的潜力。

结合多模态信息的高斯方法代表了从“纯图像驱动”向“多源感知融合”的重要演进路径。它突破了传统3DGS对图像质量与几何一致性监督的强依赖，通过语音、动作和环境等辅助信号提升建模鲁棒性、交互性与任务适配能力。与基于显式几何一致性的优化路径不同，多模态增强策略强调结构表达与外部语义或感知信号之间的协同关系，展现出更强的系统扩展性与现实应用潜力。

3.2.7　基于局部的几何推理方法小结

基于局部几何推理的高斯建模范式以3DGS为代表，突破了传统全局体素与隐式场方法在效率与细节控制方面的限制，通过构建显式可优化的高斯点集，实现了高效、高质量的三维建模与图像渲染。该范式强调局部结构表达能力，并在高斯初始化、可微渲染以及结构优化等环节形成了完整建模流程。围绕这一核心框架，当前研究逐步沿多个方向展开扩展与优化，如表7和图7所示：一方面，结构稀疏化方法通过剪枝与点分布引导，实现点集的高效压缩与推理加速；另一方面，显式几何一致性约束引入深度与法线等几何监督信号，有效缓解了跨视角不一致与结构漂移问题。

在可控性与语义表达方面，诸多方法尝试将语义嵌入、高阶语义控制或编辑机制融合进高斯建模流程，显著增强了3DGS在结构识别与人机交互任务中的应用能力。此外，为应对真实世界中动态变化与多模态感知需求，研究者提出了面向动态场景的高斯形变建模方法，以及融合语音、IMU和LiDAR等异构模态信号的多模态增强策略，进一步提升了模型的环境适应性与交互性。

总体来看，基于局部几何推理的高斯方法已逐步形成从稀疏建模、高几何一致性、语义控制以及动态形变，到多模态融合的全链条发展路径。不仅推动了三维重建技术在精度、效率与表达力上的持续突破，也为其在虚拟现实、机器人操作和动态感知等实际应用中提供了更强的落地基础与拓展潜力。未来的研究可以进一步加强动态场景下的适应性，结合时空建模和多模态信息，以提高模型的鲁棒性和实时性，继续拓宽其在现实世界中的应用范围。

表7　基于局部的几何推理方法总结

Table 7　Summary of local based geometric reasoning methods

方法类别	设计思路	优点	缺点
基于结构稀疏化优化	减少高斯点冗余，以提升建模与渲染效率	降低内存与计算开销，适配实时应用	易引发结构失真，需设计保真机制
基于几何一致性约束	引入深度、法线等几何监督提升空间一致性	提高几何重建精度与建模稳定性	对监督质量与数据完整性较为敏感
基于语义先验引导	融合语义标签或语言引导实现可控建模	支持语义理解与交互式编辑	语义映射与三维表达对齐难度较高
基于渲染模型增强	引入神经模块建模复杂光照与材质信息	提升渲染质量与细节表达能力	增加训练与推理成本
面向动态场景扩展	建立时空一致性以支持运动与形变建模	支持动态重建与时序表达	易受遮挡、模糊等因素干扰
结合多模态信息	利用非视觉模态提升感知鲁棒性与交互能力	强化模型泛化性与任务适用性	存在模态对齐与融合复杂性

4　讨论与分析

为了讨论不同类的方法针对不同种类数据集的差异化的适应性，本文从室内与室外场景的MVS方法适配性、有序街景（KITTI（Karlsruhe Institute of Technology and Toyota Technological Institute at Chicago））与无序互联网数据（LLFF）的MVS差异以及

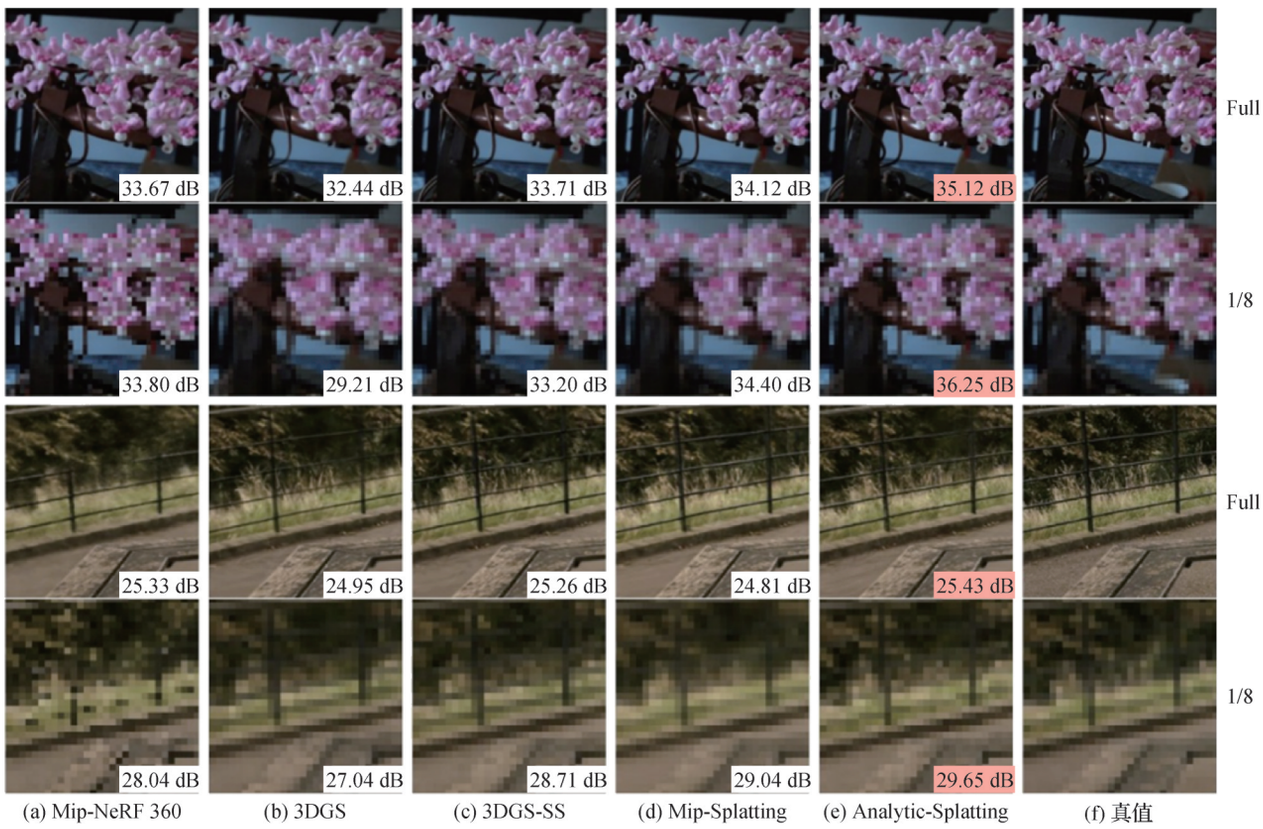

图 7　基于局部的几何推理方法定性对比

Fig. 7　Qualitative comparison of local based geometric reasoning methods

((a) Mip-NeRF 360; (b) 3DGS; (c) 3DGS-SS; (d) Mip-Splatting; (e) Analytics-Splatting; (f) ground truth)

航拍与地面视角的MVS适配策略3个角度分析了场景特性与不同方法之间适配性，具体如下：

1)室内与室外场景的MVS方法适配性。室内场景(如DTU、Tanks and Temples)通常具有可控光照、有限视角变化和相对静态的纹理分布，适合图像投影驱动方法(如MVSNet、GBiNet)。这些方法通过显式代价体构建和局部特征匹配，在低纹理区域(如KITTI道路)的深度一致性上表现更优。例如，MVS-Net在KITTI的低纹理道路区域的深度误差较NeRF降低约8%，因其代价体优化策略能有效利用纹理一致性。然而，室外场景(如LLFF、Google Birds Eye View)面临复杂光照变化、大尺度遮挡和动态物体干扰，此时几何推理驱动方法(如NeRF、3DGS)更具优势。例如，3DGS通过动态形变建模(Deformable 3DGS)在LLFF的遮挡区域将F-score提升12%，但其对计算资源的需求显著增加，需结合轻量化策略(如speedy-splat)以适应边缘设备部署。

2)有序街景(KITTI)与无序互联网数据(LLFF)的MVS差异。KITTI等有序街景数据具有固定视角分布和均匀纹理，适合依赖全局优化的图像投影方法(如R-MVSNet)，其多尺度金字塔策略能高效处理规则场景，但在无序互联网数据(如LLFF)中因视角跳跃和非刚性形变导致性能下降。相比之下，几何推理方法(如Spec-Gaussian)通过隐式场建模在LLFF的动态区域展现更强的鲁棒性，其各向异性高斯函数对反射特性建模使PSNR仅下降3.2%(静态场景基线下降11.7%)。然而，这类方法对有序场景的重建效率较低，需结合混合表示(如Instant-NGP)平衡精度与速度。

3)航拍与地面视角的MVS适配策略。航拍数据(如Google Birds Eye View)因视角跨度大且存在遮挡盲区，导致几何推理方法(如Plenoctree)的重建误差较地面数据(如Cityscapes)增加18%，主要受限于隐式场在大尺度场景中的几何歧义性。图像投影方法(如CasMVSNet)虽能处理航拍的大尺度，但其固定网格划分导致内存消耗激增，需引入动态分辨率调整策略。此外，多模态融合(如LiDAR辅助)可显著缩小航拍与地面场景的精度差距(误差缩小至

5%以内),例如通过LiDAR点云约束3DGS的高斯点分布,提升航拍场景的几何一致性。

此外,在全景图像的MVS处理中,现有方法需应对环形视角分布带来的深度歧义、大尺度几何一致性建模以及动态遮挡等核心挑战。图像投影驱动方法(如MVSNet)通过代价体构建依赖固定视角间的几何约束,在全景场景中易因视角跳跃导致深度估计不连贯,尤其在环形视角的交界区域出现显著误差。相比之下,几何推理方法(如NeRF、3DGS)通过隐式场建模可缓解视角连续性问题,但其高计算开销和对动态内容的敏感性限制了实际应用。例如,NeRF需通过体积渲染处理全景视角的连续插值,但训练时间较传统方法增加2~3倍;而3DGS通过显式高斯点表示虽能适应动态全景场景(如DeepScene),但其非刚性形变建模仍需引入额外的时间一致性约束(如Dynamic NeRF)。

5 结 语

本文基于图像投影—几何推理与全局—局部建模两个维度,将多视角立体匹配三维重建方法划分为4类典型技术路径,并通过系统梳理技术演进脉络与核心设计思路,揭示了不同方法在精度、效率与鲁棒性间的权衡机制。结合DTU、Tanks and Temples等数据集的实验结果,验证了各类方法在复杂场景下的性能表现差异。

从技术发展的纵向维度观察,三维重建技术正经历从"几何推理主导"到"数据驱动为主的演进。早期基于传统几何的方法依赖单应性映射与极线约束等原理,通过逐像素匹配与能量优化实现重建,但受限于手工特征提取的局限性,难以应对低纹理区域的建模需求。近年来,深度学习的引入显著提升了重建精度与鲁棒性,端到端的神经网络架构能够自动学习图像特征与深度关系,但其对计算资源的高消耗、对大规模标注数据的敏感性,以及在动态场景中对时间一致性建模的不足,成为当前研究的核心挑战。此外,随着元宇宙与多模态大模型的兴起,三维重建技术面临动态场景建模、跨模态融合及实时性优化等新需求,亟需在理论创新与工程实践层面寻求突破。

未来的研究方向可能集中在以下几个方面:1)轻量化与实时性优化是推动三维重建技术落地的关键。针对图像投影驱动方法的计算成本瓶颈,可探索基于知识蒸馏的模型压缩策略、动态分辨率调整机制,以及硬件加速的边缘计算框架,以提升算法在移动设备或嵌入式系统中的部署能力。2)动态场景建模与时间一致性建模将成为技术演进的重要方向。当前方法多聚焦静态场景的重建,而对人群流动、物体变形等动态场景的建模能力仍显不足,需结合时序信息建模与物理动力学约束,实现更自然的动态重建效果。3)跨模态融合与自监督学习有望缓解对大规模标注数据的依赖。通过融合RGB-D、点云、语义标签等多模态信息,可增强模型的场景理解能力;而基于自监督或弱监督的学习策略,则能在有限标注数据条件下提升模型泛化能力,尤其适用于文化遗产保护、灾害应急等数据获取受限的场景。

在应用层面,多视角三维重建技术的扩展需求正在快速多元化。在虚拟现实与元宇宙建设中,高保真三维模型是构建沉浸式体验的基础,但当前方法在大规模场景的实时渲染与交互性方面仍需突破;在自动驾驶领域,对道路环境的三维感知直接影响系统的决策能力,但复杂天气条件与动态障碍物的建模仍是技术难点;在工业检测与文化遗产保护中,非接触式的三维扫描技术虽已实现初步应用,但如何在保证几何精度的同时兼顾纹理细节的完整性,仍是亟待解决的工程问题。此外,随着数字孪生城市概念的推进,三维重建技术还需突破虚实同步的实时性壁垒,通过在线学习与增量更新机制,实现虚拟场景与物理世界的动态映射。

综上所述,三维重建技术正处于从实验室研究向实际应用加速转化的阶段。未来的研究需在模型轻量化、动态场景建模、跨模态融合与自监督学习等方向持续突破,同时需紧密结合工业、医疗和文化等领域的实际需求,推动技术从"可实现"向"可落地"转变。随着硬件计算能力的提升与多模态数据的丰富,三维重建技术有望在精度、效率与泛化能力之间实现更优平衡,为虚拟现实、文物修复和自动驾驶等前沿领域提供更坚实的技术支撑。

参考文献(References)

Barnes C, Shechtman E, Finkelstein A and Goldman D B. 2009. PatchMatch: A randomized correspondence algorithm for structural image editing. ACM Transactions on Graphics, 28(3): #24 [DOI:

10.1145/1531326.1531330］

Barron J T, Mildenhall B, Tancik M, Hedman P, Martin-Brualla R and Srinivasan P P. 2021. Mip-NeRF: a multiscale representation for anti-aliasing neural radiance fields//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 5835-5844［DOI: 10.1109/ICCV48922.2021.00580］

Bleyer M, Rhemann C and Rother C. 2011. Patchmatch stereo-stereo matching with slanted support windows//Proceedings of 2011 British Machine Vision Conference. Dundee, UK: BMVC: #14［DOI: 10.5244/C.25.14］

Cao C J, Ren X L and Fu Y W. 2022. MVSFormer: multi-view stereo by learning robust image features and temperature-based depth. Transactions on Machine Learning Research, 2022

Cao C J, Ren X L and Fu Y W. 2024. MVSFormer++: revealing the devil in transformer's details for multi-view stereo//Proceedings of the 12th International Conference on Learning Representations. Vienna, Austria: OpenReview.net

Cen J Z, Fang J M, Yang C, Xie L X, Zhang X P, Shen W, et al. 2025. Segment any 3D Gaussians//Proceedings of the 39th AAAI Conference on Artificial Intelligence. Philadelphia, USA: AAAI: 1971-1979［DOI: 10.1609/aaai.v39i2.32193］

Chelani K, Benbihi A, Sattler T and Kahl F. 2025. EdgeGaussians-3D edge mapping via Gaussian splatting//Proceedings of 2025 IEEE/CVF Winter Conference on Applications of Computer Vision. Tucson, USA: IEEE: 3268-3279［DOI: 10.1109/WACV61041.2025.00323］

Chen A P, Xu Z X, Geiger A, Yu J Y and Su H. 2022. TensoRF: tensorial radiance fields//Proceedings of the 17th European Conference on Computer Vision. Tel Aviv, Israel: Springer: 333-350［DOI: 10.1007/978-3-031-19824-3_20］

Chen A P, Xu Z X, Zhao F P, Zhang X S, Xiang F B, Yu J Y, et al. Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo//Proceedings of 2021 IEEE/CVF international conference on computer vision. 2021: 14124-14133.

Chen B, Hu S K, Chen Q, Du C P, Yi R, Qian Y M, et al. 2024. GSTalker: real-time audio-driven talking face generation via deformable Gaussian splatting［EB/OL］.［2025-07-25］. https://arxiv.org/pdf/2404.19040.pdf

Chen K H, Yuan Z L, Xiao H H, Mao T L and Wang Z Q. 2025. Learning multi-view stereo with geometry-aware prior. IEEE Transactions on Circuits and Systems for Video Technology, 2025: #3578452［DOI: 10.1109/TCSVT.2025.3578452］

Cheng S, Xu Z X, Zhu S L, Li Z W, Li L E, Ramamoorthi R, et al. 2020. Deep stereo using adaptive thin volume representation with uncertainty awareness//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 2521-2531［DOI: 10.1109/CVPR42600.2020.00260］

Ding Y K, Yuan W T, Zhu Q T, Zhang H T, Liu X Y, Wang Y J, et al. 2022a. TransMVSNet: global context-aware multi-view stereo network with transformers//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 8575-8584［DOI: 10.1109/CVPR52688.2022.00839］

Ding Y K, Zhu Q T, Liu X Y, Yuan W T, Zhang H T and Zhang C. 2022b. KD-MVS: knowledge distillation based self-supervised learning for multi-view stereo//Proceedings of the 17th European Conference on Computer Vision. Tel Aviv, Israel: Springer: 630-646［DOI: 10.1007/978-3-031-19821-2_36］

Fridovich-Keil S, Yu A, Tancik M, Chen Q H, Recht B, Kanazawa A. 2022. Plenoxels: radiance fields without neural networks//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 5491-5500［DOI: 10.1109/CVPR52688.2022.00542］

Galliani S, Lasinger K and Schindler K. 2016. Gipuma: Massively parallel multi-view stereo reconstruction. Publikationen der Deutschen Gesellschaft für Photogrammetrie, Fernerkundung und Geoinformation e. V, 25(361-369): #2

Gu X D, Fan Z W, Zhu S Y, Dai Z Z, Tan F T and Tan P. 2020. Cascade cost volume for high-resolution multi-view stereo and stereo matching//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 2492-2501［DOI: 10.1109/CVPR42600.2020.00257］

Guo J, Ma X J, Fan Y, Liu H P and Li Q. 2025a. Semantic Gaussians: open-vocabulary scene understanding with 3D Gaussian splatting［EB/OL］.［2025-07-25］. https://arxiv.org/pdf/2403.15624.pdf

Guo Y C, Kang D, Bao L C, He Y and Zhang S H. 2022. NeRFReN: IEEE: 18388- Neural radiance fields with reflections//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: #18397［DOI: 10.1109/CVPR52688.2022.01786］

Guo Z Y, Zhou W G, Li L, Wang M and Li H Q. 2025b. Motion-aware 3D Gaussian splatting for efficient dynamic scene reconstruction. IEEE Transactions on Circuits and Systems for Video Technology, 35(4): 3119-3133［DOI: 10.1109/TCSVT.2024.3502257］

Hanson A, Tu A, Lin G, Singla V, Zwicker M and Goldstein T. 2025a. Speedy-splat: fast 3D Gaussian splatting with sparse pixels and sparse primitives//Proceedings of 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 21537-21546［DOI: 10.1109/CVPR52734.2025.02006］

Hanson A, Tu A, Singla V, Jayawardhana M, Zwicker M and Goldstein T. 2025b. PUP 3D-GS: principled uncertainty pruning for 3D Gaussian splatting//Proceedings of 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 5949-5958［DOI: 10.1109/CVPR52734.2025.00558］

Hedman P, Srinivasan P P, Mildenhall B, Barron J T and Debevec P. 2021. Baking neural radiance fields for real-time view synthesis//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 5855-5864［DOI: 10.1109/ICCV48922.2021.00582］

Huang B C, Yi H W, Huang C, He Y J, Liu J B and Liu X. 2021. M3VSNet: Unsupervised multi-metric multi-view stereo network//Proceedings of 2021 IEEE International Conference on Image Processing (ICIP). Anchorage, USA: IEEE: 3163-3167 [DOI: 10.1109/ICIP42928.2021.9506469]

Huang Y H, Sun Y T, Yang Z Y, Lyu X Y, Cao Y P and Qi X J. 2024. SC-GS: Sparse-controlled Gaussian splatting for editable dynamic scenes//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 4220-4230 [DOI: 10.1109/CVPR52733.2024.00404]

Jensen R, Dahl A, Vogiatzis G, Tola E and Aanaes H. 2014. Large scale multi-view stereopsis evaluation//Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus, USA: IEEE: 406-413 [DOI: 10.1109/CVPR.2014.59]

Jiang J F, Wang L Y, Yu H C, Hu T Y, Chen J S and Ma H M. 2025. RRT-MVS: recurrent regularization transformer for multi-view stereo//Proceedings of the 39th AAAI Conference on Artificial Intelligence. Philadelphia, USA: AAAI: 3994-4002 [DOI: 10.1609/AAAI.V39I4.32418]

Kerbl B, Kopanas G, Leimkühler T and Drettakis G. 2023. 3D Gaussian splatting for real-time radiance field rendering. ACM Transactions on Graphics, 42(4): #139 [DOI: 10.1145/3592433]

Khot T, Agrawal S, Tulsiani S, Mertz C, Lucey S and Hebert M. 2019. Learning unsupervised multi-view stereopsis via robust photometric consistency [EB/OL]. [2025-07-25]. https://arxiv.org/pdf/1905.02706.pdf

Knapitsch A, Park J, Zhou Q Y and Koltun V. 2017. Tanks and temples: benchmarking large-scale scene reconstruction. ACM Transactions on Graphics, 36(4): #78 [DOI: 10.1145/3072959.3073599]

Kuhn A, Lin S and Erdler O. 2019. Plane completion and filtering for multi-view stereo reconstruction//Proceedings of the 41st DAGM German Conference on Pattern Recognition. Dortmund, Germany: Springer: 18-32 [DOI: 10.1007/978-3-030-33676-9_2]

Lang X L, Li L J, Wu C M, Zhao C, Liu L N, Liu Y, et al. 2025. Gaussian-LIC: real-time photo-realistic SLAM with Gaussian splatting and LiDAR-inertial-camera fusion//Proceedings of 2025 IEEE International Conference on Robotics and Automation. Atlanta, USA: IEEE: 8500-8507 [DOI: 10.1109/ICRA55743.2025.11128712]

Lewin S, Vandegar M, Hoyoux T, Barnich O and Louppe G. 2023. Dynamic NeRFs for soccer scenes//Proceedings of the 6th International Workshop on Multimedia Content Analysis in Sports. Ottawa, Canada: ACM: 113-121 [DOI: 10.1145/3606038.3616158]

Li Z H, Jiang H, Cai Y J, Chen J N, Bi B L, Gao S Q, et al. 2025. STDR: spatio-temporal decoupling for real-time dynamic scene rendering [EB/OL]. [2025-07-25]. https://arxiv.org/pdf/2505.22400.pdf

Li Z S, Müller T, Evans A, Taylor R H, Unberath M, Liu M Y, et al. 2023. Neuralangelo: high-fidelity neural surface reconstruction//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 8456-8465

Li Z X, Yao S L, Chu Y J, Garcia-Fernandez Á F, Yue Y, Ding W P, et al. 2026. MVG-splatting: multi-view guided Gaussian splatting with adaptive quantile-based geometric consistency densification. Information Fusion, 126: #103540 [DOI: 10.1016/j.inffus.2025.103540]

Li Z X, Zuo W M, Wang Z Q and Zhang L. 2020. Confidence-based large-scale dense multi-view stereo. IEEE Transactions on Image Processing, 29: 7176-7191 [DOI: 10.1109/TIP.2020.2999853]

Liu C, Cao T, Kang W X, Jiang Z H, Yang C H, Gui W H, et al. 2025. Single-view 3D object reconstruction based on deep learning: survey. Journal of Image and Graphics, 30(4): 0922-0952 (刘草, 曹婷, 康文雄, 蒋朝辉, 阳春华, 桂卫华, 等. 2025. 深度学习邻域单视图三维物体重建研究综述. 中国图象图形学报, 30(4): 0922-0952) [DOI: 10.11834/jig.240389]

Liu T Q, Ye X Y, Zhao W Y, Pan Z Y, Shi M and Cao Z G. 2023. When epipolar constraint meets non-local operators in multi-view stereo//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 18042-18051 [DOI: 10.1109/ICCV51070.2023.01658]

Mescheder L, Oechsle M, Niemeyer M, Nowozin S and Geiger A. 2019. Occupancy networks: learning 3D reconstruction in function space//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 4455-4465 [DOI: 10.1109/CVPR.2019.00459]

Mi Z X, Di C and Xu D. 2022. Generalized binary search network for highly-efficient multi-view stereo//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 12981-12990 [DOI: 10.1109/CVPR52688.2022.01265]

Mildenhall B, Srinivasan P P, Ortiz-Cayon R, Kalantari N K, Ramamoorthi R, Ng R, et al. 2019. Local light field fusion: practical view synthesis with prescriptive sampling guidelines. ACM Transactions on Graphics, 38(4): #29 [DOI: 10.1145/3306346.3322980]

Mildenhall B, Srinivasan P P, Tancik M, Barron J T, Ramamoorthi R and Ng R. 2022. NeRF: representing scenes as neural radiance fields for view synthesis. Communications of the ACM, 65(1): 99-106 [DOI: 10.1145/3503250]

Müller T, Evans A, Schied C and Keller A. 2022. Instant neural graphics primitives with a multiresolution hash encoding. ACM Transactions on Graphics, 41(4): #102 [DOI: 10.1145/3528223.3530127]

Peng R, Wang R J, Wang Z Y, Lai Y W and Wan R G. 2022. Rethinking depth estimation for multi-view stereo: A unified representation//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 8635-8644 [DOI: 10.1109/CVPR52688.2022.00845]

Qi C R, Su H, Mo K and Guibas L J. 2017. PointNet: deep learning on point sets for 3D classification and segmentation//Proceedings of

1017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 77-85 [DOI: 10.1109/CVPR.2017.16]

Reiser C, Peng S Y, Liao Y Y and Geiger A. 2021. KiloNeRF: speeding up neural radiance fields with thousands of tiny MLPs//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 14315-14325 [DOI: 10.1109/ICCV48922.2021.01407]

Ren C L, Xu Q S, Zhang S K and Yang J Q. 2023. Hierarchical prior mining for non-local multi-view stereo//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 3588-3597 [DOI: 10.1109/ICCV51070.2023.00334]

Romanoni A and Matteucci M. 2019. TAPA-MVS: textureless-aware patchmatch multi-view stereo//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 10412-10421 [DOI: 10.1109/ICCV.2019.01051]

Schöps T, Schönberger J L, Galliani S, Sattler T, Schindler K, Pollefeys M, et al. 2017. A multi-view stereo benchmark with high-resolution images and multi-camera videos//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 2538-2547 [DOI: 10.1109/CVPR.2017.272]

Sitzmann V, Zollhöfer M, and Wetzstein G. 2019. Scene representation networks: continuous 3D-structure-aware neural scene representations//Proceedings of the 33rd International Conference on Neural Information Processing Systems. Vancouver, Canada: ACM: #101

Stephan J G, Marek K, Matthew J, Jamie S and Julien V. 2021. FastNeRF: high-fidelity neural rendering at 200FPS//Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, Canada: IEEE: 14346-14355

Strecha C, von Hansen W, Von Gool L, Fua P, and Thoennessen U. 2008. On benchmarking camera calibration and multi-view stereo for high resolution imagery//Proceedings of 2008 IEEE Conference on Computer Vision and Pattern Recognition. Anchorage, USA: IEEE: 1-8 [DOI: 10.1109/CVPR.2008.4587706]

Sun S, Liu J J, Li Y Z, Ying H C, Zhai Z G and Mou Y R. 2022. Adaptive pixelwise inference multi-view stereo//Proceedings of the 13th International Conference on Graphics and Image Processing. Kunming, China: SPIE: #120830K [DOI: 10.1117/12.2623392]

Sun S, Zheng Y, Shi X L, Xu Z Y and Liu Y G. 2021. PHI-MVS: plane hypothesis inference multi-view stereo for large-scale scene reconstruction [EB/OL]. [2025-07-25]. https://arxiv.org/pdf/2104.06165.pdf

Turkulainen M, Ren X Q, Melekhov I, Seiskari O, Rahtu E and Kannala J. 2024. DN-splatter: depth and normal priors for Gaussian splatting and meshing//Proceedings of 2024 IEEE/CVF Winter Conference on Applications of Computer Vision. Tucson, USA: IEEE: 2421-2431 [DOI: 10.1109/WACV61041.2025.00241]

Verbin D, Hedman P, Mildenhall B, Zickler T, Barron J T and Srinivasan P P. 2022. Ref-NeRF: structured view-dependent appearance for neural radiance fields//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 5481-5490 [DOI: 10.1109/CVPR52688.2022.00541]

Wang F J H, Galliani S, Vogel C, Speciale P and Pollefeys M. 2021a. PatchmatchNet: Learned multi-view patchmatch stereo//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 14189-14198 [DOI: 10.1109/CVPR46437.2021.01397]

Wang F J, Galliani S, Vogel C and Pollefeys M. 2022. IterMVS: iterative probability estimation for efficient multi-view stereo//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. New Orleans, USA: IEEE: 8606-8615

Wang P, Liu L, Liu Y, Theobalt C, Komura T and Wang W P. 2021b. NeuS: learning neural implicit surfaces by volume rendering for multi-view reconstruction//Proceedings of the 35th Conference on Neural Information Processing Systems. [s.l.]: NeurIPS: 27171-27183

Wang Y S, Guan T, Chen Z, Luo Y W, Luo K Y and Ju L L. 2020. Mesh-guided multi-view stereo with pyramid architecture//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 2036-2045 [DOI: 10.1109/CVPR42600.2020.00211]

Wang Y S, Zeng Z J, Guan T, Yang W, Chen Z, Liu W K, et al. 2023. Adaptive patch deformation for textureless-resilient multi-view stereo//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 1621-1630 [DOI: 10.1109/CVPR52729.2023.00162]

Wang Y X, Wu Q Y, Zhang G F and Xu D. 2025. Learning 3D geometry and feature consistent Gaussian splatting for object removal//Proceedings of the 18th European Conference on Computer Vision. Milan, Italy: Springer: 1-17 [DOI: 10.1007/978-3-031-72646-0_1]

Wei M, Wu Q Y, Zheng J M, Rezatofighi H and Cai J F. 2024. Normal-GS: 3D Gaussian splatting with normal-involved rendering//Proceedings of the 38th International Conference on Neural Information Processing Systems. Vancouver, Canada: ACM: 2432

Wei Z Z, Zhu Q T, Min C, Chen Y S and Wang G P. 2021. AA-RMVSNet: adaptive aggregation recurrent multi-view stereo network//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 6167-6176 [DOI: 10.1109/ICCV48922.2021.00613]

Wu J, Li R, Xu H F, Zhao W X, Zhu Y, Sun J Q, et al. 2024. GoMVS: geometrically consistent cost aggregation for multi-view stereo//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 20207-20216 [DOI: 10.1109/CVPR52733.2024.01910]

Wu K, Zhang Z C, Tie M, Ai Z Q, Gan Z X and Ding W C. 2025. VINGS-mono: visual-inertial Gaussian splatting monocular SLAM in large scenes. IEEE Transactions on Robotics, 41: 5912-5931

[DOI: 10.1109/TRO.2025.3613536]

Xiong K Q, Peng R, Zhang Z, Feng T X, Jiao J B, Gao F, et al. 2023. CL-MVSNet: unsupervised multi-view stereo with dual-level contrastive learning//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 3749-3757 [DOI: 10.1109/ICCV51070.2023.00349]

Xu Q and Tao W. 2019. Multi-scale geometric consistency guided multi-view stereo//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 5483-5492.

Xu Q G, Xu Z X, Philip J, Bi S, Shu Z X, Sunkavalli K, et al. 2022. Point-NeRF: point-based neural radiance fields//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 5428-5438 [DOI: 10.1109/CVPR52688.2022.00536]

Xu Q S, Kong W H, Tao W B and Pollefeys M. 2023. Multi-scale geometric consistency guided and planar prior assisted multi-view stereo. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(4): 4945-4963 [DOI: 10.1109/TPAMI.2022.3200074]

Xu Y X, Wang L G, Chen M L, Ao S, Li L and Guo Y L. 2025. DropoutGS: dropping out Gaussians for better sparse-view rendering//Proceedings of 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 701-710 [DOI: 10.1109/CVPR52734.2025.00074]

Yang H, Chen R, An S P, Wei H and Zhang H. 2023. The growth of image-related three dimensional reconstruction techniquesin deep learning-driven era: a critical summary. Journal of Image and Graphics, 28(08): 2396-2409(杨航, 陈瑞, 安仕鹏, 魏豪, 张衡. 2023. 深度学习背景下的图像三维重建技术进展综述. 中国图象图形学报, 28(08): 2396-2409)[DOI: 10.11834/jig.220376]

Yang J Y, Mao W, Alvarez J M and Liu M M. 2020. Cost volume pyramid based depth inference for multi-view stereo//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 4877-4886 [DOI: 10.1109/CVPR42600.2020.00493]

Yang Z Y, Gao X Y, Sun Y T, Huang Y H, Lyu X Y, Zhou W, et al. 2024a. Spec-Gaussian: anisotropic view-dependent appearance for 3D Gaussian splatting//Proceedings of the 38th International Conference on Neural Information Processing Systems. Vancouver, Canada: ACM: #1956

Yang Z Y, Gao X Y, Zhou W, Jiao S H, Zhang Y Q and Jin X G. 2024b. Deformable 3D Gaussians for high-fidelity monocular dynamic scene reconstruction//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 20331-20341 [DOI: 10.1109/CVPR52733.2024.01922]

Yao Y, Luo Z X, Li S W, Fang T and Quan L. 2018. MVSnet: depth inference for unstructured multi-view stereo//Proceedings of the 15th European Conference on Computer Vision (ECCV). Munich, Germany: Springer: 785-801 [DOI: 10.1007/978-3-030-01237-3_47]

Yao Y, Luo Z X, Li S W, Shen T W, Fang T and Quan L. 2019. Recurrent MVSNet for high-resolution multi-view stereo depth inference//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 5520-5529 [DOI: 10.1109/CVPR.2019.00567]

Yariv L, Gu J T, Kasten Y and Lipman Y. 2021. Volume rendering of neural implicit surfaces//Proceedings of the 35th International Conference on Neural Information Processing Systems. [s.l.]: NeurIPS: #367

Ye M Q, Danelljan M, Yu F and Ke L. 2025. Gaussian grouping: Segment and edit anything in 3D scenes//Proceedings of the 18th European Conference on Computer Vision. Milan, Italy: Springer: 162-179 [DOI: 10.1007/978-3-031-73397-0_10]

Yu A, Li R L, Tancik M, Li H, Ng R and Kanazawa A. 2021. PlenOctrees for real-time rendering of neural radiance fields//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 5732-5741 [DOI: 10.1109/ICCV48922.2021.00570]

Yu L and Wu X Q. 2024. Survey of texture optimization algorithms for 3D reconstructed scenes. Journal of Image and Graphics, 29(08): 2303-2318(于柳, 吴晓群. 2024. 三维重建场景的纹理优化算法综述. 中国图象图形学报, 29(08): 2303-2318)[DOI: 10.11834/jig.230478]

Yu M L, Lu T, Xu L N, Jiang L H, Xiangli Y B and Dai B. 2024. GSDF: 3DGS meets SDF for improved neural rendering and reconstruction//Proceedings of the 38th International Conference on Neural Information Processing Systems. Vancouver, Canada: NeurIPS: #4115

Yuan Z L, Cao J K, Li Z X, Jiang H and Wang Z Q. 2024a. SD-MVS: segmentation-driven deformation multi-view stereo with spherical refinement and EM optimization//Proceedings of the 38th AAAI Conference on Artificial Intelligence. Vancouver, Canada: AAAI: 6871-6880 [DOI: 10.1609/aaai.v38i7.28512]

Yuan Z L, Cao J K, Wang Z Q and Li Z X. 2024b. TSAR-MVS: textureless-aware segmentation and correlative refinement guided multi-view stereo. Pattern Recognition, 154: #110565 [DOI: 10.1016/J.PATCOG.2024.110565]

Yuan Z L, Luo J G, Shen F, Li Z X, Liu C, Mao T L and Wang Z Q. 2025. DVP-MVS: Synergize depth-edge and visibility prior for multi-view stereo//Proceedings of the 39th AAAI Conference on Artificial Intelligence. Philadelphia, USA: AAAI: 9743-9752 [DOI: 10.1609/aaai.v39i9.33056]

Zhang S, Xu W J, Wei Z W, Zhang L L, Wang Y and Liu J Y. 2023a. ARAI-MVSNet: a multi-view stereo depth estimation network with adaptive depth range and depth interval. Pattern Recognition, 144: #109885 [DOI: 10.1016/j.patcog.2023.109885]

Zhang Z, Peng R, Hu Y X and Wang R G. 2023b. GeoMVSNet: learning multi-view stereo with geometry perception//Proceedings of

2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 21508-21518 [DOI: 10.1109/CVPR52729.2023.02060]

Zhang Z L, Song T C, Lee Y, Yang L, Peng C, Chellappa R, et al. 2024. LP-3DGS: learning to prune 3D Gaussian splatting//Proceedings of the 38th International Conference on Neural Information Processing Systems. Vancouver, Canada: NeurIPS: #3891

Zhao R, Han X, Guo X D, Kuang L Q, Yang X W and Sun F S. 2023. Exploring the point feature relation on point cloud for multi-view stereo. IEEE Transactions on Circuits and Systems for Video Technology, 2023, 33(11): 6747-6763.

Zhou Z Y, Feng Q D and Li H J. 2025. SynGauss: real-time 3D Gaussian splatting for audio-driven talking head synthesis. IEEE Access, 13: 42167-42177 [DOI: 10.1109/ACCESS.2025.3548015]

作者简介

袁祯泷，男，博士研究生，主要研究方向为计算机视觉与三维重建。E-mail: yuanzhenlong21b@ict.ac.cn

毛天露，通信作者，女，副研究员，主要研究方向为虚拟现实和智能人机交互。E-mail: ltm@ict.ac.cn

李泽昊，男，博士研究生，主要研究方向为计算机视觉、3D 高斯与三维重建。E-mail: lizehao23z@ict.ac.cn

陈科桦，男，硕士研究生，主要研究方向为神经辐射场与三维重建。E-mail: chenkehua23s@ict.ac.cn

蒋浩，男，副研究员，主要研究方向为虚拟现实和智能人机交互。E-mail: jianghao@ict.ac.cn

王兆其，男，研究员，主要研究方向为虚拟现实和智能人机交互。E-mail: zqwang@ict.ac.cn